Argumentative Writing Support by means of Natural Language Processing

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

Dissertation

zur Erlangung des akademischen Grades Dr.-Ing.

vorgelegt von
Dipl.-Inform. Christian Matthias Edwin Stab
geboren in Darmstadt

Tag der Einreichung: 13. December 2016
Tag der Disputation: 10. February 2017

Referenten: Prof. Dr. phil. Iryna Gurevych, Darmstadt
Prof. Marie-Francine Moens, Ph.D., Leuven
Prof. Dr. Manfred Stede, Potsdam

Darmstadt 2017
D17

Abstract

Persuasive essay writing is a powerful pedagogical tool for teaching argumentation skills. So far, the provision of feedback about argumentation has been considered a manual task since automated writing evaluation systems are not yet capable of analyzing written arguments. Computational argumentation, a recent research field in natural language processing, has the potential to bridge this gap and to enable novel argumentative writing support systems that automatically provide feedback about the merits and defects of written arguments.

The automatic analysis of natural language arguments is, however, subject to several challenges. First of all, creating annotated corpora is a major impediment for novel tasks in natural language processing. At the beginning of this research, it has been mostly unknown whether humans agree on the identification of argumentation structures and the assessment of arguments in persuasive essays. Second, the automatic identification of argumentation structures involves several interdependent and challenging subtasks. Therefore, considering each task independently is not sufficient for identifying consistent argumentation structures. Third, ordinary arguments are rarely based on logical inference rules and are hardly ever in a standardized form which poses additional challenges to human annotators and computational methods.

To approach these challenges, we start by investigating existing argumentation theories and compare their suitability for argumentative writing support. We derive an annotation scheme that models arguments as tree structures. For the first time, we investigate whether human annotators agree on the identification of argumentation structures in persuasive essays. We show that human annotators can reliably apply our annotation scheme to persuasive essays with substantial agreement. As a result of this annotation study, we introduce a unique corpus annotated with fine-grained argumentation structures at the discourse-level. Moreover, we present a novel end-to-end approach for parsing argumentation structures. We identify the boundaries of argument components using sequence labeling at the token level and propose a novel joint model that globally optimizes argument component types and argumentative relations for identifying consistent argumentation structures. We show that our model considerably improves the performance of local base classifiers and significantly outperforms challenging heuristic baselines.

In addition, we introduce two approaches for assessing the quality of natural language arguments. First, we introduce an approach for identifying myside biases which is a well-known tendency to ignore opposing arguments when formulating arguments. Our experimental results show that myside biases can be recognized with promising accuracy using a combination of lexical features, syntactic features and features based on adversative transitional phrases. Second, we investigate for the first time the characteristics of insufficiently supported arguments. We show that insufficiently supported arguments frequently exhibit specific lexical indicators. Moreover, our experimental results indicate that convolutional neural networks significantly outperform several challenging baselines.

Zusammenfassung

Das Schreiben von argumentativen Aufsätzen ist eine effektive Methode, Argumentationsfähigkeiten zu lehren. Bisher ist die Bewertung von argumentativen Aufsätzen eine rein manuelle Aufgabe, da automatisierte Schreibhilfen nicht in der Lage sind, Argumente automatisch zu analysieren. *Computational argumentation*, ein junges Forschungsfeld der natürlichen Sprachverarbeitung, hat das Potential diese Lücke zu schließen und neue intelligente Schreibhilfen zu ermöglichen, die automatisch konstruktive Rückmeldungen zu natürlichsprachlichen Argumenten generieren.

Die automatische Analyse von natürlichsprachlichen Argumenten unterliegt den folgenden Herausforderungen. Zum einen ist die Erstellung von annotierten Korpora ein große Hürde für neue Bereiche der natürlichen Sprachverarbeitung. Zu Beginn dieser Arbeit war es weitestgehend unbekannt, ob Argumente in argumentativen Aufsätzen mit ausreichender Übereinstimmung von menschlichen Annotatoren erkannt und bewertet werden können. Zum anderen besteht die automatische Erkennung von Argumentationsstrukturen aus mehreren komplexen und voneinander abhängigen Analyseschritten, die nicht unabhängig voneinander gelöst werden können. Zudem basieren die meisten Argumente nicht auf logischen Regeln und sind selten in einer standardisierten Form, was eine weitere Herausforderung für menschliche Annotatoren und computerbasierte Methoden darstellt.

In dieser Dissertation vergleichen wir zuerst existierende Argumentationstheorien und prüfen deren Eignung für intelligente Schreibhilfen. Wir stellen ein Argumentationsmodell vor, welches die Argumentationsstruktur eines gesamten Dokumentes als Baum modelliert. Wir zeigen erstmalig, dass menschliche Annotatoren Argumentationsstrukturen mit hoher Übereinstimmung identifizieren. Das Ergebnis dieser Annotationsstudie ist ein mit Argumentationsstrukturen annotiertes Korpus, welches der Forschungsgemeinschaft zur freien Verfügung steht. Darüber hinaus stellen wir einen neuen automatischen Ansatz zur Erkennung von Argumentationsstrukturen vor. Dieser Ansatz erkennt die Grenzen von Argumentkomponenten auf Wortebene. Zusätzlich stellen wir ein neues Modell zur Erkennung von Argumentationsstrukturen vor, welches die Funktion von Argumentkomponenten und argumentative Relationen gemeinsam modelliert. Die Evaluationergebnisse zeigen, dass dieser Ansatz nicht nur konsistente Argumentationsstrukturen erkennt, sondern auch im Vergleich zu mehreren heuristischen Ansätzen signifikant bessere Erkennungsraten erzielt.

Zusätzlich stellen wir zwei weitere Ansätze zur Bewertung der Argumentqualität vor. Der erste Ansatz erkennt Bestätigungsfehler, welche in der Kognitionspsychologie als eine Tendenz zur Vernachlässigung von Gegenargumenten bekannt sind. Die Evaluationsergebnisse zeigen, dass die Erkennung von Bestätigungsfehlern mit einer Kombination aus lexikalischen Merkmalen, syntaktischen Eigenschaften und adversativen Phrasen die besten Ergebnisse erzielt. Für den zweiten Ansatz untersuchen wir erstmals die Eigenschaften von unzureichend begründeten Argumenten. Wir zeigen, dass unzureichend begründete Argumente oft spezifische lexikalische Eigenschaften aufweisen. Zudem stellen wir einen Ansatz basierend auf neuronalen Netzen vor, welcher unzureichend begründete Argumente automatisch erkennt und im Vergleich mit mehreren Baselinesystemen signifikant bessere Erkennungsraten erzielt.

Acknowledgments

First and foremost, I thank Prof. Dr. Iryna Gurevych for giving me the opportunity to conduct this research, for her continuous support, and for her excellent supervision in recent years. Moreover, I would like to thank Prof. Dr. Marie-Francine Moens and Prof. Dr. Manfred Stede for finding the time to evaluate my thesis. This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, the German Federal Ministry of Education and Research (BMBF) as a part of the Software Campus project AWS under grant No. 01|S12054 and by IBM with a PhD Fellowship.

I'm very thankful to all my colleagues at UKP Lab and the very constructive feedback that I received during numerous talks and discussions. In particular, I would like to thank Dr. Ivan Habernal, Dr. Judith Eckle-Kohler and Christian Kirschner for the very valuable discussions during our weekly meetings of the computational argumentation working group. I'm also very grateful to Dr. Christian M. Meyer for the numerous helpful hints and his guidance. Also I would like to express my gratitude to Ilya Kuznetsov for his guidance in integer linear programming. I thank my student assistants Can Diehl, Radhika Gaonkar, Piyush Paliwal, Nikhil Priyatam Pattisapu, Krish Perumal, and Anshul Tak for their valuable contributions and their support. I also would like to thank Dr. Richard Eckart de Castilho, Dr. Johannes Daxenberger and the whole DKPro team for providing excellent software frameworks without which this thesis would have been way more laborious.

Last but not least, I would like to express my deepest gratitude to my family and my wife. Without your outstanding support and understanding this thesis would not have been possible.

Contents

Chapter 1

Introduction

Argumentation is a verbal activity that aims at increasing or decreasing the acceptability of a controversial standpoint (van Eemeren et al., 1996, p. 5). It is a routine which is omnipresent in our daily verbal communication and thinking. Well-reasoned arguments are not only important for making justified decisions but also play a crucial role in drawing widely-accepted conclusions and deriving novel knowledge in epistemic activities. Moreover, the ability to develop well-reasoned arguments is a fundamental requirement for learning itself (Davies, 2009, p. 94). It enables students to critically assess evidence and to reason rationally.

Argumentative writing and, more specifically, persuasive essay writing is a powerful pedagogical tool for teaching argumentation (Botley, 2014, p. 46). However, a great many of students are still not adequately prepared in constructing strong arguments (Butler and Britt, 2011, p. 70; Wolfe and Britt, 2009, p. 183). One reason for this shortcoming is that teachers are not able to provide sufficient writing assignments in view of growing class sizes and the enormous load for manually analyzing arguments (Burstein et al., 2004, p. 27). Automated essay scoring (Shermis and Burstein, 2013a) or computer-supported writing systems (Burstein et al., 2004) aim to reduce the load of teachers, but these systems are limited to feedback about spelling, grammar, mechanics, discourse structures, or lexical richness (Shermis and Burstein, 2013b) and do not address argumentation (Lim and Kahng, 2012). Novel developments in computational argumentation could bridge this gap and enable innovative argumentative writing support systems that provide tailored feedback about written arguments.

Computational argumentation is a recent and rapidly evolving field of research in computational linguistics that addresses the analysis of natural language arguments. Most of the existing approaches are trained on manually annotated corpora using supervised machine learning. These methods learn the identification of arguments, their components, and relations by means of known texts and different types of linguistic features. One of the first approaches in computational argumentation has been introduced by Mochales-Palau and Moens (2009), who focused on the identification of argument components and their structures in legal texts. Recently, researchers in computational argumentation made numerous proposals for automatically extracting arguments. For example, Habernal and Gurevych (2016a) identified argument components in user-generated web discourse and Peldszus and Stede (2015) addressed argumentation structures in microtexts. Other approaches

like those proposed by researchers at IBM focus on mining claims (Levy et al., 2014) and corresponding evidence (Rinott et al., 2015) across multiple Wikipedia articles. While these approaches achieve promising results, they are not sufficient for analyzing arguments in student essays because of the following challenges which have not been adequately addressed by previous work:

First of all, recognizing argumentation structures in text is a difficult task even for humans. On the one hand, ambiguity and vagueness of realistic texts often impede a definite interpretation of natural language arguments (Govier, 2010), and on the other hand, components of the same argument may be too far apart to reliably recognize argumentative relations in texts. In fact, it has been mostly unknown whether humans agree on the identification of argumentation structures in realistic texts at the beginning of this research and whether it is possible to create reliable corpora for training argumentation structure parsers.

Second, most of the existing approaches in computational argumentation focus on particular subtasks of argumentation structure parsing such as identifying argument components (Moens et al., 2007; Levy et al., 2014; Al-Khatib et al., 2016), classifying the type of argument components (Kwon et al., 2007; Rooney et al., 2012), or recognizing argumentative discourse relations (Biran and Rambow, 2011a; Cabrio and Villata, 2012a). These tasks are, however, not independent of each other. For instance, the type of argument components depends on the argumentative discourse structure and vice versa. Therefore, modeling these tasks independently does not suffice to recognize consistent argumentation structures. Despite this interdependence most approaches operate locally and do not globally optimize argumentation structures. Recently and during the writing of the current thesis, Peldszus and Stede (2015) proposed a promising approach based on *minimum spanning trees* (MST) which jointly models several subtasks in a single model. This approach achieves promising results for identifying argumentation structures by considering the function and role of already known argument components. An end-to-end approach covering all subtasks that is capable of separating several arguments is, however, still missing.

Third, assessing the quality of arguments is a highly subjective task. The persuasive power of arguments is a product of many different criteria which may depend on personal preferences and the previous knowledge of individuals who evaluate the argument (Thomas, 1973). For instance, the quality of an argument is subject to the level of trust that an individual has in the arguer (ethos), the emotions appealed by the argument (pathos), the kind and quality of reasoning (logos), and the circumstances in which the argument appears in (kairos) (Schiappa and Nordin, 2013). The high degree of subjectivity is a major impediment for developing methods that automatically assess the quality of arguments and for creating argumentative writing support systems.

These challenges give rise to the following three research questions, which we want to approach in the current thesis:

RQ1 Annotating Argumentation Structures: We want to investigate if argumentation models proposed in argumentation theory are applicable to persuasive essays. In this respect, we want to evaluate if, how and to which extent human annotators agree on argumentation structures in persuasive essays and

if it is possible to create annotated corpora of high quality.

RQ2 Parsing Argumentation Structures: The second research question addresses the automatic identification of argumentation structures. We want to investigate which linguistic features are effective for several subtasks involved in the identification of argumentation structures and if jointly modeling several subtasks improves the accuracy and consistency of the predicted argumentation structures.

RQ3 Assessing Argument Quality: The third research questions addresses the automatic evaluation of natural language arguments. We want to investigate which quality criteria can be employed for providing objective feedback to students, if human annotators agree on their application to realistic texts, and how accurate computational models can predict them.

Approaching these research questions is a first step towards a better understanding of arguments in natural language texts. On the one hand, investigating the applicability of theoretical argument models to persuasive essays will enable an assessment of the viability of argumentative writing support systems. On the other hand, we seek to get an estimate of how well current natural language processing methods recognize arguments in text. This could potentially promote the development of enabling technologies for innovative information retrieval systems and decisions support systems in the near future.

1.1 Contributions

The contributions of this thesis can be divided into (*i*) a part on parsing argumentation structures, i.e. an approach for identifying the components and argumentative relations of natural language arguments, and (*ii*) a part on assessing argument quality, i.e. approaches for analyzing the merits or defects of arguments. The following lists provide an overview of these contributions:

Parsing Argumentation Structures:

- We provide a systematic summary of existing approaches in computational argumentation and introduce related work. We propose a taxonomy and categorize existing approaches and corpora accordingly.

- We introduce an annotation scheme for modeling argumentation structures in persuasive essays derived from argumentation theory. It models the arguments of a persuasive essay as a tree structure and is not limited to isolated single arguments or specific aspects of argumentation.

- We show that our annotation scheme can be reliably applied to persuasive essays by human annotators. We introduce an unique corpus that represents (at the date of writing this thesis) the largest language resource annotated with fine-grained argumentation structures at the discourse-level.

- For the first time, we present an end-to-end argumentation structure parser that covers all steps required for identifying fine-grained argumentation structures at the discourse-level. Our approach is based on supervised machine learning and joint modeling that globally optimizes argumentation structures. More specifically, our approach combines the type of argument components and argumentative relations in order to find an optimal structure. Our parser also separates argumentative from non-argumentative text units and recognizes the boundaries of argument components at the token level.

Assessing Argument Quality:

- Argument quality is a product of various aspects and there are numerous proposals for evaluating the quality of arguments. We provide an overview of the most prominent approaches in argumentation theory, clarify their relationships and compare them with respect to their suitability for argumentative writing support.

- We create the first corpus of arguments for studying if an argument is sufficiently supported. We employ the RAS-criteria proposed by Johnson and Blair (2006) and show that human annotators can reliably differentiate between sufficiently supported and insufficiently supported arguments.

- For the first time, we investigate the characteristics of insufficiently supported arguments. We show that insufficiently supported arguments frequently exhibit specific lexical indicators. Our experimental results indicate that convolutional neural networks significantly outperform several challenging baselines and recognize insufficiently supported arguments with a promising accuracy.

- We present an approach for identifying myside biases in persuasive essays which is a tendency to ignore opposing arguments and to formulate arguments biased towards one's own prior beliefs. Our experimental results show that the absence of opposing arguments can be recognized by using a combination of lexical features, syntactic features and features based on adversative transitional phrases.

1.2 Publication Record

Several parts of this thesis have been previously published in international peer-reviewed conference and workshop proceedings from major events in natural language processing, e.g. *EMNLP* and *COLING*.[1] We list all the publications below and indicate the chapters and sections of this thesis which build upon them:

- **Christian Stab** and Iryna Gurevych. 2017. Recognizing Insufficiently Supported Arguments in Argumentative Essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '17, pp. (to appear), Valencia, Spain (Section 6.1)

[1] A major part of this thesis is currently under review in *Computational Linguistics*: **Christian Stab** and Iryna Gurevych. Parsing Argumentation Structures in Persuasive Essays, Under review in *Computational Linguistics*, First submission: 26 October 2015. Revised submission: 15 July 2016.

- **Christian Stab** and Iryna Gurevych. 2016. Recognizing the Absence of Opposing Arguments in Persuasive Essays. In *Proceedings of the 3rd Workshop on Argument Mining*, pp. 113–118, Berlin, Germany (Section 6.2)
- Beata Beigman Klebanov, **Christian Stab**, Jill Burstein, Yi Song, Binod Gyawali and Iryna Gurevych. 2016. Argumentation: Content, Structure, and Relationships with Essay Quality. In *Proceedings of the 3rd Workshop on Argument Mining*, pp. 70–75, Berlin, Germany
- **Christian Stab** and Ivan Habernal. 2016. Existing Resources for Debating Technologies. *Report of Dagstuhl Seminar on Debating Technologies (15512)*, pp. 32, Wadern, Germany (Section 3.1)
- **Christian Stab** and Ivan Habernal. 2016. Detecting Argument Components and Structures. *Report of Dagstuhl Seminar on Debating Technologies (15512)*, pp. 32-33, Wadern, Germany (Section 3.2)
- Iryna Gurevych and **Christian Stab**. 2016. Argumentative Writing Support: Structure Identification and Quality Assessment of Arguments. *Report of Dagstuhl Seminar on Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments (16161)*, pp. 87-88, Wadern, Germany
- **Christian Stab** and Iryna Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pp. 46–56, Doha, Qatar (Sections 5.4.1 and 5.4.2)
- **Christian Stab** and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, pp. 1501–1510, Dublin, Ireland (Chapter 4)
- Christian M. Meyer, Margot Mieskes, **Christian Stab**, and Iryna Gurevych. 2014. DKPro Agreement: An Open-Source Java library for Measuring Inter-Rater Agreement. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations* COLING '14, pp. 105–109, Dublin, Ireland (Sections 4.2.2, 6.2.1, and 6.1.1)
- **Christian Stab**, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pp. 40–49, Bertinoro, Italy

1.3 Thesis Organization

This thesis is structured in seven chapters. In this section, we provide an overview of the organization of this thesis and the content of each chapter:

Chapter 2: "*Argumentation: Overview and Background*":
The first approaches on studying argumentation date back to the ancient Greeks. Therefore, it is not surprising that there is a huge number of different argumentation theories which cover different aspects of argumentation. The same applies to quality criteria of arguments. In Chapter 2, we provide a non-exhaustive overview of existing argumentation models and quality criteria. We compare their suitability for argumentative writing support and introduce the terminology with respect to argumentation theory used throughout this thesis.

Chapter 3 "*Computational Argumentation*":
Computational argumentation is a recent field in natural language processing. In Chapter 3, we provide an overview of related work. We introduce a taxonomy for categorizing existing approaches and review existing corpora in order to facilitate the selection of language resources or the orientation of future annotation studies. We also discuss the similarity and differences between argumentation structures and discourse analysis.

Chapter 4 "*Annotating Argumentation Structures*":
In Chapter 4, we approach the first research question (RQ1) by deriving an annotation scheme from argumentation theory and investigating its applicability to persuasive essays. We show that human annotators can apply our annotation scheme to persuasive essays with substantial agreement and introduce a novel corpus of persuasive essays annotated with fine-grained argumentation structures at the discourse-level. Furthermore, we provide detailed statistics of the annotated argumentation structures in order to better understand the characteristics of arguments in persuasive essays and to derive the requirements for computational approaches.

Chapter 5 "*Parsing Argumentation Structures*":
In Chapter 5, we investigate the second research questions (RQ2) and present a novel method for parsing argumentation structures which consists of several consecutive analysis steps. First, we segment a persuasive essay in order to identify relevant argument components. Second, we jointly model the classification of argument component types and the identification of argumentative relations using integer linear programming for ensuring consistent argumentation structures. Third, we recognize the stance of each argument component in order to discriminate between argumentative support and attack relations. We show that our approach considerably improves the identification of argumentation structures and significantly outperforms a challenging heuristic baseline. In addition, we propose novel feature sets for all tasks and evaluate their effectiveness to better understand the characteristics of written arguments.

Chapter 6 "*Quality Assessment*":
In Chapter 6, we approach the third research question by analyzing two quality criteria in persuasive essays (RQ3). First, we investigate the applicability of theoretical quality criteria to arguments in persuasive essays. In particular, we focus on the sufficiency criterion that an argument complies with if its premises provide sufficient support for accepting its claim. We show that human annotators agree

on the sufficiency criterion and show that insufficiently supported arguments can be identified with a promising accuracy using convolutional neural networks. Second, we investigate myside biases in persuasive essays and analyze which linguistic features are informative for recognizing them. We model this task as a binary document classification and consider an essay as biased if it does not include opposing arguments.

Chapter 7 "*Summary*":
Finally, we summarize the main contributions of this thesis and present a prospect of future work.

Chapter 2

Argumentation: Overview and Background

The study of argumentation is a comprehensive and interdisciplinary research field. It involves philosophy, communication science, logic, linguistics, psychology and computer science. The first approaches to study argumentation date back to the ancient Greek sophists and evolved in the 6^{th} and 5^{th} centuries B.C. In particular, the influential work of Aristotle on traditional logic, rhetoric, and dialectic sets an important milestone in the study of argumentation and is still an essential cornerstone of modern argumentation theories. In order to capture the diversity of the field, van Eemeren et al. (1996) propose the following definition of argumentation:

> *"Argumentation is a verbal and social activity of reason aimed at increasing (or decreasing) the acceptability of a controversial standpoint for the listener or reader, by putting forward a constellation of propositions intended to justify (or refute) the standpoint before a rational judge." (van Eemeren et al., 1996, p. 5)*

It defines argumentation as a *verbal* activity since an arguer puts forward opinions, claims, justifications or reasons either in spoken or written form. The definition also states that argumentation is a *social* activity. Certainly, the social characteristic is most evident in dialogical communication when arguments are directed towards other people. However, even if someone deliberates on a decision in an internal monologue, the consideration of pros and cons is basically a social activity because it anticipates the reactions of a potential opponent.

Argumentation requires a *standpoint* on a topic. However, taking a standpoint is not sufficient to begin an argumentation. It also requires that standpoints differ or are supposed to differ. Thus, argumentation presupposes that the standpoint is *controversial* and that there is another standpoint that diverges from the standpoint of the arguer. Furthermore, the definition determines the objective of argumentation as *increasing (or decreasing) the acceptability* of a controversial standpoint. Thus, the goal of argumentation is to persuade a listener or reader of the arguer's standpoint and to reject other standpoints.

Finally, the definition refers to a *constellation of propositions* as the means of argumentation. It states that the purpose of these propositions is to either *justify* or *refute* the standpoint before drawing a *rational judge*.

Due to the diversity of the field, there are numerous proposals for modeling argumentation. Bentahar et al. (2010) propose a taxonomy including the following three types of argumentation models: (1) monological models, (2) dialogical models, and (3) rhetorical models (Figure 2.1).

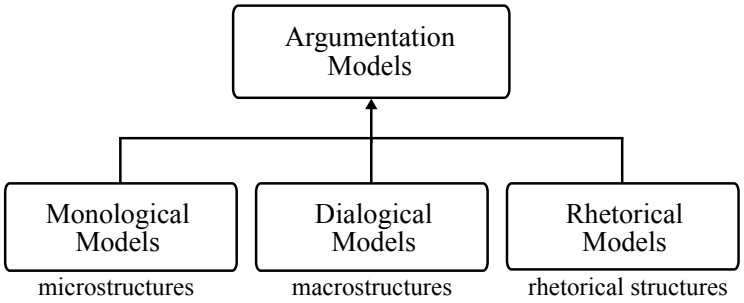


Figure 2.1: Taxonomy of argumentation models adapted from Bentahar et al. (2010).

Monological models address the internal structure of arguments. Their objective is to define the types of argument components, the links between them, and the types of reasoning. Thus, the focus of monological models is the *microstructure* of arguments. For instance, Toulmin's argument model formalizes the internal microstructure by means of six different argument components (Toulmin, 1958, p. 97), whereas argumentation schemes proposed by Hastings (1963) define different types of reasoning that can be observed in everyday discourse. Most of these models stem from the field of *informal logic* that evolved in the early 1950^{th}. Monological models focus on the analysis and evaluation of arguments as *product*, whereas dialogical models focus on the *process* of argumentation (Johnson, 2000; O'Keefe, 1977).

Dialogical models address the relations between arguments and the external *macrostructure* respectively. They focus primarily on formalizing conversations such as discussions, debates or negotiations and usually ignore the internal microstructure of single arguments. Examples of dialogical models are MacKenzie's model of formal dialectics (MacKenzie, 1981), the abstract argumentation framework proposed by Dung (1995), or Amgoud's argumentative dialog modeling framework (Amgoud et al., 2000).

Rhetorical models neither consider the microstructure nor the macrostructure but rather rhetorical patterns of arguments. In contrast to monological and dialogical models, they also consider the audience's perception and aim at studying the way of using arguments as a means of persuasion. Examples of rhetorical models are the new rhetoric theory proposed by Perelman and Olbrechts-Tyteca (1969) which defines rhetorical schemes that are successful in practice (Bentahar et al., 2010, p. 243), or the rhetorical approach for persuasive negotiation suggested by Ramchurn et al. (2007).

These three perspectives on the study of argumentation are closely related (Bentahar et al., 2010; Reed and Walton, 2003; Walton and Godden, 2007). The formulation of a single argument, for instance, is part of the process of argumentation, and dialogical situations presuppose arguments as product as well as the consideration of the audience's perception. Accordingly, more recent argumentation theories combine these perspectives. They consider argumentation as a hypothetical dialectical

exchange between a proponent and an opponent and attempt to develop a holistic theory of argumentation (van Eemeren and Grootendorst, 2004; Freeman, 2011; Peldszus and Stede, 2013a).

In the remainder of this chapter, we focus primarily on monological models. On the one hand, the types of argument components defined by monological models are an important foundation for creating annotation schemes and for modeling arguments in natural language texts respectively. On the other hand, monological models do not presuppose the presence of several interlocutors, and are thus well suited for modeling argumentation structures in persuasive essays. Furthermore, monological models address the microstructure of arguments, and are therefore appropriate for a fine-grained analysis of arguments in texts.

In Section 2.1, we introduce the most prevalent approaches for modeling the microstructure of arguments. In Section 2.2, we introduce formal and informal criteria for assessing the quality of arguments.

2.1 Argument Models

The microstructure of an *argument* consists of several *argument components*. It includes a claim and one or more premises (Govier, 2010, p. 1; Damer, 2009, p. 13; Hurley, 2012, p. 1). The *claim* is a controversial statement and the central component of an argument.[1] The *premises* constitute the reasons for believing the claim to be true or false (Damer, 2009, p. 14).[2] In addition, an argument includes a *consequence relation* that connects the premise(s) to the claim and may determines the reasoning type of the argument. Accordingly, an argument includes at least a claim, a single premise, and a consequence relation (Figure 2.2).

Figure 2.2: Minimal form of an argument including a claim, a single premise, and a consequence relation.

In comparison with an *opinion* or a personal belief, an argument is a supported claim that includes at least one premise intended to justify the claim. An opinion is merely an unsupported claim without justification (Damer, 2009, p. 15). The following example illustrates an argument with a single premise about cloning.

> Premise: "*Scientists showed that cloning can be used to raise organs.*"
>
> ———
>
> Claim: "*Humankind will benefit from modern cloning technology.*"

The claim is a controversial statement about cloning. It states that cloning will be beneficial for humankind. The premise provides a reason for supporting the claim by stating that scientists successfully demonstrated that cloning can be used to raise

[1] The claim is also called *conclusion* in related literature. In this thesis, we exclusively use the term *claim* to refer to the central component of an argument.

[2] The premise is also called *reason*, *evidence*, *justification* or *data*.

organs. In this example, the structure is already given and the claim and premise are known. However, in ordinary language arguments are hardly ever present in a structured form. Consequently, it is necessary to recognize the components of an argument and its structure before further analysis steps (Govier, 2010, p. 22). For instance, the following text includes an argument with two premises:

> "*Cloned organs will match perfectly to the blood group and tissue of patients. It also shortens the waiting time. Therefore, cloning is beneficial for medical purposes.*"

The first two sentences are the premises and the last sentence includes the claim. The indicator "*therefore*" facilitates the identification of the claim. Other indicators for claims are for instance "*consequently*", "*thus*", or "*hence*". Indicators like "*because*", "*since*", or "*given that*" may signal the presence of premises. In the following argument, the identification of argument components is more difficult.

> "*Finding an appropriate donor can take several years in particular cases.*"

This example represents an incomplete argument. It neither includes several components nor indicators that reveal the role of the component. Instead it consists of a single premise that supports an unstated and implicit claim. These arguments are called *enthymemes* (Hurley, 2012, Chapter 5) and can be reconstructed by referring to contextual information. For instance, by the knowledge that the topic of the debate is cloning, we can reconstruct the standard form of the argument as:

> Premise 1: "*Cloning organs shortens the healing process.*"
> Premise 2: "*Shorter healing is beneficial for transplantation patients.*"
> ———
> Claim: "*Therefore, cloning is beneficial for transplantation patients.*"

The *standard form* of an argument explicitly lists all components of an argument. It includes a list of all premises followed by the claim of the argument. The standard form also reconstructs all enthymemes and makes all components of the argument explicit (Damer, 2009, p. 17). However, deriving the standard form of natural language arguments is a challenging task even for humans. It is highly subjective and requires well-informed domain knowledge and contextual information. More information about the reconstruction of arguments can be found in various textbooks on argument evaluation, for instance those written by Govier (2010) or Damer (2009).

2.1.1 Toulmin's Argument Model

Most argument models include one type of premise and do not distinguish between different reasons. However, we can easily observe different types of premises in everyday discourse (Bentahar et al., 2010, p. 216). For example, a premise could provide empirical evidence, eyewitness or a justification why the reasoning of an argument is correct. Toulmin (1958) suggests an argument model with six different types of argument components that contribute to the argument's strength in different ways:

1. *Claim*: The claim is the central component of the argument in the same way as in the previously defined model. It is a controversial statement which the author or speaker wants to persuade the listener or reader of.

2. *Data*: Toulmin's counterpart of the common premise is called data (van Eemeren et al., 2014; Reed and Rowe, 2006). It either specifies facts as evidence for justifying the claim or previous beliefs that are related to the current argument.

3. *Warrant*: A justification that the reasoning from data to claim is correct. It answers the question why the data counts in favor of the claim.

4. *Backing*: This component justifies the reliability of the warrant. It is relevant if the warrant is attacked by an opponent.

5. *Qualifier*: Indicates the degree of certainty of the claim or any condition for the truth of the claim.

6. *Rebuttal*: Exceptions or situations under which the argument might not hold true, e.g. opposing arguments or circumstances in which the claim can be false.

The model defines a normative view of an optimal argument. It can also be considered as a tool for evaluating the strength of arguments (Sampson and Clark, 2006). Each component can be used as a question in order to identify the weak points of an argument. For instance, asking for a warrant answers the question why the data is relevant to the claim. Similarly, investigating and determining potential rebuttals strengthens the argument against opposing opinions. In this way, the model serves as a guideline for constructing strong arguments since the defined components anticipate critique from potential opponents.

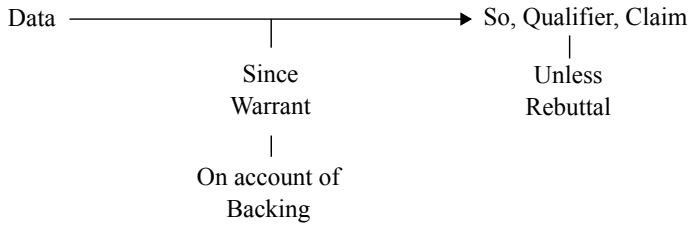


Figure 2.3: Toulmin's argument model (Toulmin, 2003, p. 97).

The structure of Toulmin's model is illustrated in Figure 2.3. The model includes a premise (data) that supports the claim in the same way as in the minimal form of an argument (cf. Figure 2.2). In addition to the minimal model, the warrant can be considered as a consequence relation expressed in words. Figure 2.4 shows an instantiation of Toulmin's model taken from his original textbook. It illustrates the roles of all six argument components.
Although Toulmin's argument model represents an excellent guideline for manually analyzing and constructing well-reasoned arguments, there are several drawbacks of applying it to natural language arguments and for computational purposes respectively. First of all, it is difficult to model everyday arguments with the original

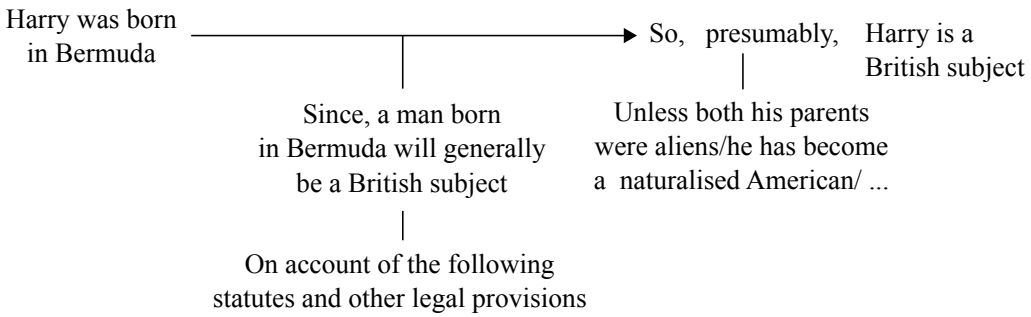


Figure 2.4: Example argument in Toulmin's model (Toulmin, 2003, p. 97).

Toulmin model (Mochales-Palau and Moens, 2009) due to the fuzzy distinction between the defined argument components, In particular, the distinction between data, warrant and backing is often vague in practice (Freeman, 2011; Hitchcock, 2003). Second, the warrant is almost never stated in everyday arguments (van Eemeren et al., 1996, p. 140; Toulmin, 2003, p. 92). The backing is also irrelevant in most practical situations since it presupposes the presence of a warrant. For the same reason, Habernal and Gurevych (2016a) omit the original warrant and redefine the role of backing as "*a single personal experience or statement that gives credibility or attributes certain expertise to the author*" (Habernal and Gurevych, 2016a, p. 18). Third, Toulmin's model includes only a single attacking argument component (rebuttal). It is not possible to model an attack of the rebuttal and thus the attack of potential opposing standpoints which is a common practice in practical argumentation.

2.1.2 Argumentation Schemes

Argumentation schemes define various types of arguments in everyday discourse. The initial list of argumentation schemes proposed by Arthur Hastings (Hastings, 1963) was adopted and extended by other researchers, e.g. Perelman and Olbrechts-Tyteca (1969), Kienpointner (1992) and Grennan (1997). Walton (1996) presents a list including 26 argumentation schemes that was extended to 96 in his later work (Walton et al., 2008). They define an argumentation scheme as follows:

> "*Argumentation Schemes are forms of argument (structures of inference) that represent structures of common types of arguments used in everyday discourse, as well as in special contexts like those of legal argumentation and scientific argumentation.*" (Walton et al., 2008, p. 1)

Accordingly, the focus of argumentation schemes is on the reasoning type and the nature of the inference rather than on the role of argument components. Argumentation schemes also model defeasible arguments which might not be very strong by themselves. A *defeasible argument* is a tentative proof that can be accepted by means of the evidence that is known so far, but it can be discarded if new evidence emerges. Argumentation schemes also describe arguments that are fallacious in the sense of traditional logic and occur frequently in everyday discourse such as argument from expert opinion, argument from ignorance or argument from popular opinion.

An argumentation scheme includes templates of argument components that describe the type reasoning. Most of the schemes include a *major premise*, a *minor premise* and a *conclusion* (claim) (Walton et al., 2008, p. 308–346). In addition, an argumentation scheme includes a set of *critical questions* (CQ) that can be used to evaluate the argument's strength once the argumentation scheme is known.

The following example taken from the textbook of Walton et al. (2008) illustrates the analysis of everyday arguments using argumentation schemes:

> *Helen and Bob are hiking along a trail in Banff, and Bob points out some tracks along the path, saying, "These look like bear tracks, so a bear must have passed along this trail."* (Walton et al., 2008, p. 9)

The claim in this argument represents a generalization from a sign (bear tracks). Although this argument seems to be plausible, there is still a chance that the claim is false. The tracks, for instance, might be from another animal or could be created by tricksters. In both cases the inferred claim would be false. The argumentation scheme of this defeasible argument is known as argument from sign:

ARGUMENT FROM SIGN[3]

Minor Premise: `A` (a finding) is true in this situation.
Major Premise: `B` is generally indicated as true when its sign, `A`, is true.

Conclusion (claim): `B` is true in this situation.

CQ1: What is the strength of the correlation of the sign with the event signified?

CQ2: Are there other events that would more reliably account for the sign?

Once the argumentation scheme is known, the associated critical questions can be used to evaluate the strength of the argument. For the argument of sign, for instance, one could ask how strong the correlation between the sign and the event is or if there are other events that would more reliably account for the claim. Thus, argumentation schemes can also be considered as a tool for manually and critically assessing arguments of a particular argumentation scheme.

However, the framework of argumentation schemes does not provide tools for comparing the strength of different schemes with each other or, more specifically, there are no relations between the argumentation schemes that describe which scheme is stronger than another. Although critical questions provide a tool for assessing an argument of a particular scheme, it will not guide author's to select another argumentation scheme that might be considerably stronger than the current one. Furthermore, argumentation schemes are defined and collected by observing everyday discourse. Thus, it is unknown if the current list of 96 argumentation schemes is complete. Consequently, it is also unknown if all existing arguments can be explained with the current list. Another drawback of argumentation schemes is

[3] The description of the argumentation scheme is taken from (Walton et al., 2008, p. 329).

that a single argument can exhibit several schemes (Reed and Walton, 2003). Thus, not each argument can be assigned uniquely to a single argumentation scheme. Therefore, we will not further consider argumentation schemes in this thesis.

2.1.3 Argument Diagramming

Laying out the structure of arguments is a widely used method in informal logic (Copi and Cohen, 1990, p. 18-45; Govier, 2010, p. 22-56). This technique referred to as *argument diagramming*, aims at transferring arguments in natural language into structured representations for evaluating them in subsequent analysis steps (Henkemans, 2000, p. 447). Although argumentation theorists usually define argument diagramming as a manual activity, the diagramming conventions are a good foundation for designing systems that automatically recognize arguments in natural language texts (Peldszus and Stede, 2013a).

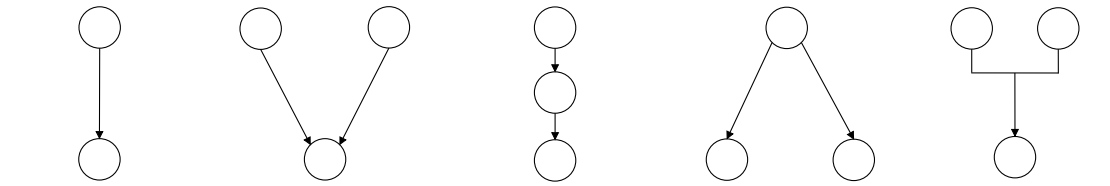


Figure 2.5: Microstructures of arguments proposed by argumentation theorists. Nodes indicate argument components and arrows mark argumentative relations. Nodes at the bottom are the claims of the individual arguments.

An argument diagram is a node-link diagram in which each node represents an argument component (a statement represented in natural language). Each link represents a directed argumentative relation such as a support relation that indicates that the source component is a premise for justifying the target component (Reed et al., 2007, p. 93). Figure 2.5 shows the common argument structures. A *basic argument* includes a claim supported by a single premise. It can be considered as the minimal form of an argument. A *convergent argument* comprises two premises that support the claim individually; an argument is *serial* if there is a reasoning chain and *divergent* if a single premise supports several claims (Beardsley, 1950). Complementary, Thomas (1973) defined *linked arguments* (Figure 2.5e). Like convergent arguments, a linked argument includes two premises. However, neither of the two premises independently supports the claim. The premises are only relevant to the claim in conjunction. More complex arguments can combine any of these elementary structures illustrated in Figure 2.5. In order to model contra positions and opposing reasons, Peldszus and Stede (2013a) proposed another type of argumentative relations which indicates that the source component attacks the target component. We refer to this relation as attack relation.

The following example paragraph taken from a persuasive essay about studying abroad exhibits four argument components. The first sentence is the claim (underlined) and the following two sentences include several premises (wavy underlined).

> "[*Living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet*]$_{Claim}$. [*One who is living*

overseas will of course struggle with loneliness, living away from family and friends]$_{Premise1}$ *but* [*those difficulties will turn into valuable experiences in the following steps of life*]$_{Premise2}$. *Moreover,* [*the one will learn living without depending on anyone else*]$_{Premise3}$."

The identification of argument components and their argumentative roles (claims and premises) is the first step for recognizing the argument structure. The next step focuses on the identification of the targets of each premise and thus on the identification of argumentative relations between the argument components. Finally, the classification of support and attack relations reveals if the argument component is a justification or refutation of the target. Figure 2.6 illustrates the structure of the previous argument.

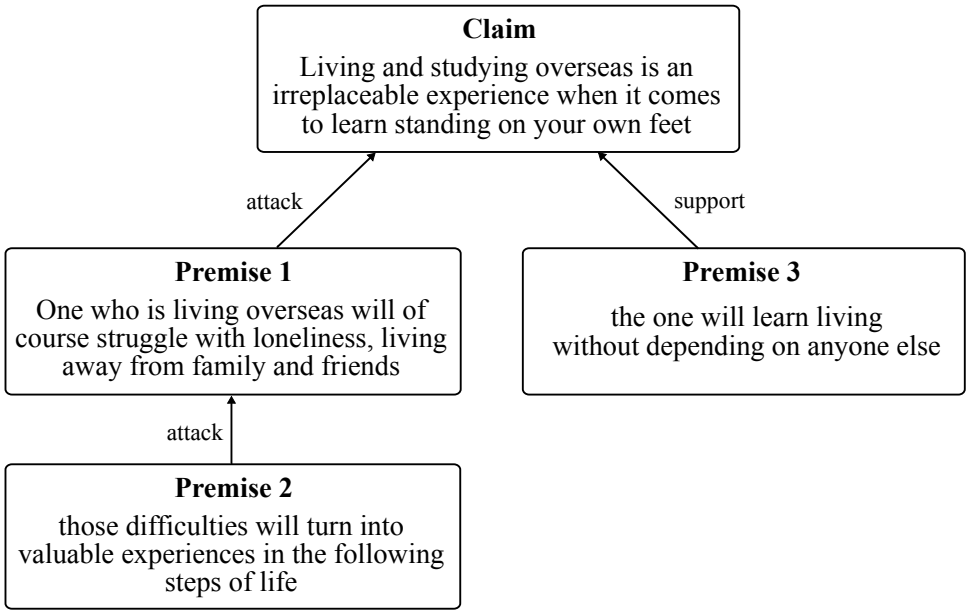


Figure 2.6: Example of an argument diagram.

The argument diagram reveals that the argument includes two attacking premises in a serial structure (Premise 1 and Premise 2). In addition, it includes another premise that supports the claim directly (Premise 3).

The example illustrates that argument diagramming allows for modeling more complex argument structures such as chains of attacking premises. In contrast to Toulmin's model (cf. Section 2.1.1) or argumentation schemes (cf. Section 2.1.2), it is not limited to single arguments and also allows for modeling relations between entire arguments. Argument diagramming focuses on the basic types of argument components (claims and premises) and argumentative relations between them. Compared to the components of Toulmin's argument model, the distinction of these basic components may be less ambiguous (cf. Section 2.1.1). Since argument diagrams also model the precise targets of each premise, they also allow for separating several arguments in texts, whereas Toulmin's model only implicitly models the relations between argument components (Habernal and Gurevych, 2016a). Thus, argument diagramming techniques are a good foundation for modeling the microstructure of arguments in natural language texts.

2.2 Quality of Arguments

Assessing the quality of arguments is a complex task since arguments are hardly ever present in standard form (Damer, 2009; Govier, 2010). In addition, the quality of an argument is a product of many different criteria (Johnson and Blair, 2006). For instance, it depends on the lexical clarity and the phrasing of the argument (representation), the level of trust that the audience has in the arguer (ethos), the emotions and values appealed by the argument (pathos), and many more. The *logical quality* (logos) is, however, independent of other merits, defects and external influence factors (Johnson and Blair, 2006, p. 50). Therefore, it is most suitable for assessing the internal quality of arguments and for providing objective feedback about arguments respectively. There are two different perspectives on the logical quality of arguments: (1) the formal logic perspective, and (2) the informal logic perspective.

The objective of *formal logic approaches* is to formalize the relations between premises and claims (Copi and Cohen, 1990, p. 46). They rely on mathematical formalisms to distinguish between deductive and inductive arguments (van Eemeren et al., 1996, chapter 1.2). Furthermore, they focus on assessing the soundness of arguments by evaluating the truth of premises in deductively valid arguments. We introduce these criteria in Sections 2.2.1 and 2.2.2.

Informal logic approaches aim at evaluating arguments in everyday discourse (Groarke, 2015). Compared to *formal logic approaches*, they are not restricted to formally valid arguments but also enable the evaluation of defeasible and inductive arguments which represent a great deal of arguments in everyday discourse (Damer, 2009, p. 22; Copi and Cohen, 1990; p. 357). Informal logic approaches include fallacy theories and the RAS-criteria. We introduce both in Section 2.2.3 and Section 2.2.4 respectively.

2.2.1 Deductive and Inductive Arguments

Formal logic differentiates two different types of arguments: deductive arguments and inductive arguments (Copi and Cohen, 1990, p. 45-47; Hurley, 2012, p. 33; van Eemeren and Grootendorst, 2004, p. 43).

Deductive arguments are based on mathematical inference rules. The truth of the claim follows necessarily from its premises and it is impossible that the premises are true and the claim is false (Copi and Cohen, 1990, p. 45). Deductive arguments are formal proofs and thus represent arguments with formally correct inference that guarantees the truth of the claim given that the premises are true (Weston, 2000, p. 40). Therefore, these arguments follow formal logical inference rules like e.g. *modus ponens*, *modus tollens* and different types of *syllogisms*.[4] The following argument illustrates a deductive argument and, more specifically, a classical *syllogism*:

> Premise 1: "*All fruits are sweet.*"
> Premise 2: "*All peaches are fruits.*"

[4] More details about formal logical inference rules can be found in textbooks from Copi and Cohen (1990) and Weston (2000).

Claim: "*Therefore, all peaches are sweet.*"

Each argument component in this example is a *categorical proposition* that includes a quantifier, a subject term, a copula[5] and a predicate term (Copi and Cohen, 1990, p. 166). The validity of a syllogism is independent of the subject matter of the argument and depends on the types of its categorical propositions and the form of the syllogism (Copi and Cohen, 1990, p. 194). Due to their quantifiers (all three categorical propositions start with "*all*") and their copula ("*be*"), all three propositions in the example above are of type `A` (Copi and Cohen, 1990, Chapter 5). Consequently, and due to the arrangement of the terms, the form of this syllogism is `AAA-1` which is a valid form of deductive inference (Copi and Cohen, 1990, p. 194). In total, there are 256 distinct forms of syllogisms of which only a few are deductively valid. The description of all categorical propositions and types of formal logic inference is beyond the scope of this thesis. More details about the formal evaluation of deductive arguments can be found in various textbooks about logic like those written by Hurley (2012) or Copi and Cohen (1990).

In contrast to deductive arguments that aim to achieve *certainty*, the objective of inductive arguments is to achieve the *acceptability*[6] of a claim (van Eemeren et al., 1996, p. 33). In *inductive arguments*, the truth of the claim is not a formal logical consequence of its premises. The premises provide reasons for increasing or decreasing the acceptability of the claim (Govier, 2010, p. 91). Inductive arguments are based on probabilistic reasoning (Hurley, 2012, p. 33) or reasoning that "*moves from specific cases to generalizations*" (Freeley and Steinberg, 2009, p. 174). Consequently, inductive reasoning is less strong than deductive reasoning in the view of formal logic. However, inductive reasoning is omnipresent in everyday discourse even in law or science (Walton et al., 2008, p. 1). The following argument illustrates common inductive reasoning in court cases:

Premise 1: "*Bob's fingerprints are on the murder weapon.*"
Premise 2: "*Bob has no alibi.*"

Claim: "*Therefore, Bob should be convicted.*"

Although the two premises count in favor of the claim, it is not proven that Bob is guilty. There is still a chance that a third person wants to set Bob up as a murder and premeditatedly used the weapon with Bob's fingerprints for committing the crime. In addition, not having an alibi is not a guarantor for being guilty, since having an alibi requires that a person has been seen by another person during the time of offense. Thus, the truth of the claim in this inductive argument is not proven. Instead the acceptability of the claim depends on the strength of the premises and needs to be re-evaluated if new evidence emerges.

[5] The *copula* connects the subject with the predicate. Details can be found in (Copi and Cohen, 1990, p. 166).

[6] Acceptability is the informal counterpart of the formal notion of truth.

2.2.2 Validity and Soundness

In the previous section, we defined a deductive argument as an argument in which the claim follows necessarily from its premises. In these arguments, it is impossible that the premises are true and the claim is false. In formal logic, deductive arguments are called *valid* (Hurley, 2012, p. 44). All other types of arguments, and consequently also inductive arguments, are *invalid* arguments.

The second formal logic quality criterion besides the validity of arguments is *soundness*. An argument is sound if it is valid and all premises are formally true (Copi and Cohen, 1990, p. 52). Accordingly, an unsound argument, i.e. an argument that is not deductively valid or that exhibits false premises, fails to deduce the truth of the claim. Figure 2.7 illustrates the relations between validity, truth of premises, and soundness.

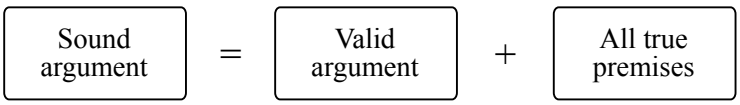


Figure 2.7: A sound argument is a deductively valid argument whose premises are formally true (Hurley, 2012, p. 47).

As a consequence of the interdependence between validity and soundness, merely knowing that an argument is deductively valid is not sufficient for evaluating the formal logical quality of an argument. For instance, the following example taken from page 50 of Copi and Cohen's textbook is a valid deductive argument that illustrates the limitations of the validity criterion:

> Premise 1: "*All spiders have ten legs.*"
> Premise 2: "*All ten-legged creatures have wings.*"
> ———
> Claim: "*Therefore, all spiders have wings.*"

The reasoning of this argument is deductively valid. However, the argument is not sound since the premises are not true. Consequently, the truth of the claim is not proven. Note that the example argument on page 19 is also unsound since the first premise is false (lemons are fruits and are not sweet). However, the inference is valid and even the claim that peaches are sweet is true. Thus, formal logic approaches on argument quality require both, the evaluation of the validity and the assessment of the truth of the premises.

Although sound arguments are the strongest arguments, using formal logic criteria for evaluating the quality of arguments has several drawbacks, especially if the arguments are present in natural language. First of all, verifying the validity of arguments requires a formal representation (e.g. categorical propositions) which is hard to extract from natural language. Although semantic parsers like *Sempre* (Berant and Liang, 2015) or *C&C* and *Boxer* (Curran et al., 2007) promise to encode natural language in logical forms, the formal evaluation of natural language arguments additionally requires the reliable recognition of claims and premises, the reconstruction of enthymemes and other text normalization tasks like resolving paraphrases or coreferences. Besides these challenges, the second and more crucial drawback is

that deductive arguments appear infrequent in everyday discourse (Damer, 2009; Copi and Cohen, 1990). Consequently, formal logic approaches are limited to a particular set of arguments and cannot be employed for assessing the merits or defects of a great many of arguments in everyday discourse (Woods and Walton, 2007; van Eemeren et al., 1996).

2.2.3 Fallacy Theory

Fallacy theories focus on determining and naming common mistakes in argumentation. The resulting lists serve as guidelines for recognizing mistakes in reasoning and for preventing frequent pitfalls when constructing arguments. The study of fallacious arguments dates back to Aristotle in 384-322 B.C. who proposed the first collection of fallacies (van Eemeren et al., 1996, p. 56). Although the initial list has been extended and refined over the centuries, there is still no consensus about the precise definition of fallacies in argumentation theory (Hansen, 2015).

The standard definition defines a fallacy as an argument that is deceptive and seems to be a good argument though it exhibits a common mistake (Hamblin, 1970, p. 12). Some of these mistakes occur so frequently in everyday discourse that they have a particular name (Damer, 2009, p.52). The standard definition is still disputed since researchers disagree if a fallacy is an argument at all and if it appears to be valid (Hansen, 2015; Woods and Walton, 2007). However, for the sake of this section and in order to understand the basic concepts of fallacy theory, it is enough to consider a fallacy as a particular defect of arguments that can be observed in everyday discourse.

Most fallacy theories cover formal as well as informal fallacies. *Formal fallacies* describe a violations of deductive inference rules (Damer, 2009, p. 76). They define fallacious forms of arguments that are invalid in the sense of formal logic like *denying the antecedent* or *undistributed middle term. Informal fallacies* focus on the meaning and the content of arguments rather than on the formal logical form. The following example taken from Bennett (2012, p. 168) illustrates an informal fallacy:

> "*People generally like to walk on the beach. Beaches have sand. Therefore, having sand floors in homes would be a great idea!*"

This fallacy is known as *non-sequitur* which translates to "*it does not follow*". Although people like to walk on beaches that have sand, it does not mean that they like to have sand in their floors. In this type of argument, the given premises are irrelevant to the claim since they do not count in favor of the claim.

Another common fallacy is *begging the question* that is also known as *arguing in a circle* (Damer, 2009, p. 63). The following example is taken from Bennett (2012, p. 82):

> "*Paranormal activity is real because I have experienced what can only be described as paranormal activity.*"

Begging the question arguments assume the truth of their claim in their premises and do not answer the actual question at hand. In the example above, the premise assumes that paranormal activities are real because there is no other explanation

for an observed event. Thus, the argument "begs the question" and does not answer why paranormal activity is real.

The *hasty generalization* fallacy supports a claim with too few samples (Damer, 2009, p. 161). The following example illustrates this type of fallacy:

> "*My neighbor has an academic degree and is the mayor of our town. Therefore, all mayors have an academic degree.*"

The claim in this argument represents a generalization that is inferred from only one particular sample. Obviously, this sample is not enough for supporting the general claim of the argument.

Besides these three examples, there are numerous other types of fallacies. For instance, Copi and Cohen (1990) introduced eighteen informal fallacies. More recent collections like the one proposed by Damer (2009) describe 61 fallacies and include informal as well as formal fallacies. However, fallacy theories are not appropriate for distinguishing good from bad arguments since it is unlikely that current collections are complete. Even if we cannot classify an argument as fallacious on the basis of a certain collection of fallacies, it does not guarantee that the argument is good. There is still a chance that the argument has a defect that is not included in our current collection (van Eemeren et al., 1996, p. 178). Therefore, fallacies theories do not allow to provide positive feedback to students, i.e. highlighting arguments which do not require further elaboration. In addition, the types of fallacies may very across different text types which restricts their applicability to different domains. For instance, it is likely that dialogical communications exhibit considerably more *ad hominem*[7] fallacies than monological texts. Thus, we will not further consider fallacy theories in the current thesis.

2.2.4 Relevance, Acceptability and Sufficiency

A high quality argument is free of fallacies. However, there is little consensus about the optimal set of fallacies and it is unknown if all fallacies are already known (van Eemeren et al., 1996, p. 178). Therefore, Johnson and Blair (1977) propose another framework for evaluating everyday arguments that constitutes an essential groundwork for informal logic (Groarke, 2015).[8] Instead of listing common mistakes, they propose three binary criteria that a logically good argument needs to fulfill (Johnson and Blair, 2006, p. 55):

- *Relevance*: An argument fulfills the relevance criterion, if all of its premises count in favor of the truth (or falsity) of the claim.
- *Acceptability*: An argument fulfills the acceptability criterion if its premises represent undisputed common knowledge or facts.

[7] Arguments attacking the opponent instead of her/his position.

[8] Johnson and Blair's criteria have been widely adopted in argumentation theory and can be found in many of today's textbooks on argument analysis like "*Attacking Faulty Reasoning*" by Damer (2009), "*A Practical Study of Argumentation*" by Govier (2010), "*Thinking Logically: Basic Concepts for Reasoning*" by Freeman (1988), "*Good Reasoning Matters!*" by Groarke and Tindale (2012), and many more.

- *Sufficiency*: An argument complies with the sufficiency criterion if its premises provide enough evidence for accepting or rejecting the claim.

The relevance criterion addresses the relation between each premise and the claim whereas the acceptability criterion focuses on the truthfulness of each individual premise. Both need to be evaluated independently for each premise of the argument. The sufficiency criterion addresses the premises of an argument together. It is fulfilled if the relevant premises of an argument are enough for justifying (or rejecting) the claim.[9]

In contrast to fallacy theories, the RAS-criteria enable to distinguish good from bad arguments with respect to logical quality since each argument that complies with all three criteria is a logically good one (Govier, 2010; Johnson and Blair, 2006). Each of the following examples illustrates the violation of one of these criteria.

Premise: "*Students gain a lot of experience when studying abroad.*"

Claim: "*Students who studied abroad will contribute more in their jobs.*"

The argument violates the relevance criterion since having more experience is not a guarantor for more commitment. Therefore, the premise does not count in favor of the claim and is not relevant. The next example illustrates another violation of the relevance criterion.

Premise: "*Having children brings happiness to our life.*"

Claim: "*Parents are so serendipitous with their kids.*"

The premise is a paraphrase of the claim. Therefore, it cannot be accepted as a relevant justification. A better premise would be that raising children is like having an important goal in your life and that someone is more happy if she/he has a goal. The next argument violates the acceptability criterion:

Premise: "*It will cause social trouble if students are dressed unequal.*"

Claim: "*Therefore, wearing school uniforms should be mandatory.*"

The premise states that not wearing school uniforms will cause social trouble among students. However, there is no evidence that different cloths provoke problems among students. Thus, the premise is an unwarranted assumption and cannot be accepted. A critical thinking person would ask why wearing school uniforms reduces social trouble and would probably not accept the argument. The following argument illustrates a violation of the sufficiency criterion.

[9] The sufficiency criterion presupposes a non-empty set of relevant premises. However, an argument can violate the relevance criterion and comply with the sufficiency criterion at the same time since an argument can have several relevant premises that are sufficient for accepting the claim and additional premises that are not relevant to the claim. This also implies that sufficient arguments have relevant premises but it is unknown if all premises of a sufficient argument are relevant to the claim.

Premise 1: "*Tourists destroyed the Great Barrier Reef.*"
Premise 2: "*They broke corrals as souvenirs and dropped fuel.*"

Claim: "*Therefore, tourism harms the natural habitats of destinations.*"

The argument includes a single sample for supporting a general claim. The fact that a particular attraction was destroyed is not sufficient for claiming that tourism generally harms natural habitats. Thus, the argument does not comply with the sufficiency criterion.

The RAS-criteria can be considered as a relaxed version of the formal logic criteria described above. In particular, the relevance criterion addresses the relation between premises and claims and thus the consequence relation of the argument. The acceptability criterion focuses on the status of the premises similarly to the formal logic criterion of soundness. The sufficiency criterion addresses the completeness of the premises which is implicitly given by deductive validity in formal logic. Table 2.1 shows the dependencies between formal and informal criteria and their objectives.

	formal logic	*informal logic*
objective	certainty of the claim	acceptability of the claim
inference criterion	validity	relevance
status of premises	truth / soundness	acceptability
completeness	deduction	sufficiency

Table 2.1: Comparison of formal and informal quality criteria.

In addition, Figure 2.8 illustrates the relationships between the RAS-criteria and formal logic criteria in a Venn diagram. Each argument that complies with all RAS-criteria is a logically good argument (bold intersection of all RAS-criteria). Note that all fallacies are a violation of one or several of these criteria (Damer, 2009; Govier, 2010) and thus can be located anywhere outside of this intersection. For instance, the non-sequitur and the begging the question fallacies introduced in Section 2.2.3 are violations of the relevance criterion. Both describe particular issues of the relation between premises and claim. The hasty generalization fallacy is a violation of the sufficiency criterion.[10] Not all deductive arguments comply with the acceptability criterion, since deductive arguments do not require that each premise is undisputed common knowledge or a fact (Copi and Cohen, 1990, p. 50). However, all deductive arguments comply with the relevance and sufficiency criteria because the claims of all deductive arguments follow necessarily from their premises. For this reason, deductive arguments are encapsulated in the corresponding intersection. Note that all non-deductive arguments in Figure 2.8 follow other types of reasoning like e.g. inductive reasoning. It is also important to note that an argument can comply with the sufficiency criterion and violate the relevance criterion at the same time. For instance, in an argument with several premises, one premise can be irrelevant and the remaining ones are still sufficient for supporting the claim. Sound arguments are a subset of deductive arguments that intersects with the acceptability criterion since all formally true premises are also acceptable.

[10] More details about the relation between fallacy theories and RAS-criteria can be found in the textbook from Damer (2009).

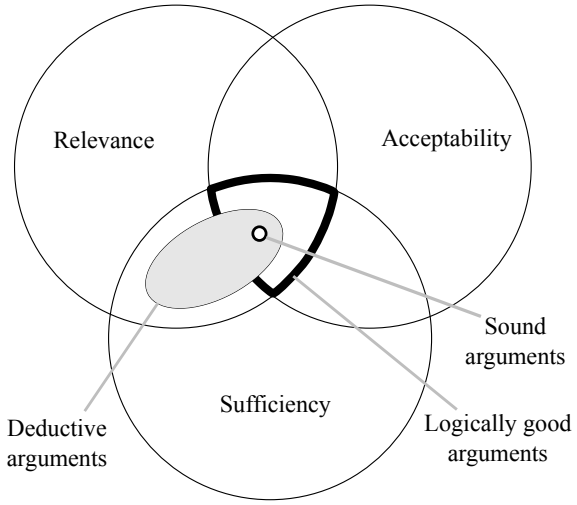


Figure 2.8: Overview and relationships between informal and formal logical quality criteria for argument evaluation.

Compared with formal logic criteria, the RAS-criteria are not restricted to a particular subset of arguments, i.e. deductive arguments. Instead they can be employed for evaluating any type of argument (including arguments that follow deductive inference). In addition, they enable to differentiate between logically good and bad arguments. Thus, the RAS-criteria also enable positive feedback, i.e. pointing out good arguments which comply with all three criteria, whereas fallacy approaches are not appropriate for differentiating between good and bad arguments. In addition, the RAS-criteria allow to attribute a particular defect to the relation between premises and claim (relevance), the premises considered together (sufficiency) or individual premises (acceptability). Therefore, the RAS-criteria enable constructive feedback for resolving particular defects of weak arguments and are particularly suited for our objectives outlined in the introduction.

2.3 Chapter Summary

Argumentation involves to persuade others of a particular standpoint, to exchange ideas about a controversial topic, and to explicitly and implicitly justify opinions. The first studies on argumentation can be traced back to the early work of Aristotle on logic, rhetoric and pragmatic. Today, there are diverse theoretical proposals for formalizing arguments with varying objectives. These approaches range from monological models for structuring argument components over dialogical models addressing the relations between arguments to rhetorical models which consider the perception of the targeted audience.

In this chapter, we reviewed the most prevalent approaches for formalizing the microstructure of arguments. We put a special emphasis on monological models which allow for a fine-grained analysis of arguments in text. By comparing different monological models, we found that argument diagramming is a good foundation for defining annotation schemes. Compared to other monological models, it allows for modeling more complex arguments including serial structures or chains of at-

tacking argument components. Furthermore, argument diagrams explicitly model the targets of argument components by means of argumentative relations. Because of this property, argument diagramming also allows to separate several arguments appearing in the same text.

In the second part of this chapter, we compared several quality criteria for assessing arguments. The quality of natural language arguments is a product of many different criteria which frequently depend on highly subjective factors. We put a special emphasis on logical quality criteria which are mostly independent of external influence factors like ethos, pathos, and kairos (Johnson and Blair, 1977). We found that formal logical approaches are limited to verifying deductively valid forms of reasoning. In contrast, informal approaches such as fallacy theories and the RAS-criteria proposed by Johnson and Blair (1977) are not limited to a particular set of arguments and allow for assessing arguments in everyday discourse. Furthermore, we showed that the RAS-criteria are not restricted to particular defects but also enable the distinction between well-reasoned and fallacious arguments. In addition, the RAS-criteria can be used to attribute a particular defect to specific components of an argument. Because of these features, the RAS-criteria allow for a fine-grained analysis and enable focused feedback about the quality of arguments.

Chapter 3

Computational Argumentation

Computational argumentation is a recent research field in natural language processing that involves the identification and analysis of arguments in natural language text. The first approaches appeared only a few years ago and focused on extracting arguments from legal documents (Mochales-Palau and Moens, 2007). The field receives increased interest in the recent past and is rapidly growing which is clearly evident by international events like the annual *Workshop on Argument Mining*, the *Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, and two Dagstuhl seminars, the first on *Debating Technologies*[1] and the second on *Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments*[2]. Apart from being a highly competitive research field, computational argumentation also bears a high economic potential. For instance, beyond providing formative feedback to students, novel advances could enable innovative applications in information retrieval, decision support, assisted reasoning, automatic argument construction or debating technologies.

Current approaches in computational argumentation focus predominantly on argument mining. These approaches involve several subtasks like separating argumentative from non-argumentative text units (Moens et al., 2007; Florou et al., 2013; Al-Khatib et al., 2016), classifying argument components (Kwon et al., 2007; Rooney et al., 2012; Mochales-Palau and Moens, 2009), and recognizing argumentative relations either between argument components (Mochales-Palau and Moens, 2011; Peldszus, 2014; Peldszus and Stede, 2015) or between full arguments (Cabrio and Villata, 2012a; Ghosh et al., 2014; Boltužić and Šnajder, 2014).

Complementarily, there are also approaches that focus on *argument attribution*. These approaches aim to identifying certain properties of arguments like the type of reasoning (Feng and Hirst, 2011), the argumentation style (Oraby et al., 2015), the sentiment flow (Wachsmuth et al., 2015) or the stance of the author (Somasundaran and Wiebe, 2009; Hasan and Ng, 2012). However, there are hardly any approaches addressing the quality of natural language arguments. The few existing approaches for assessing argument quality either focus on identifying undisputed arguments in a particular community (Cabrio and Villata, 2012b) or address the argumentation strength of an entire document as a holistic score (Persing and Ng, 2015).

In this chapter, we provide a systematic overview of existing work in computa-

[1] `http://www.dagstuhl.de/15512`
[2] `http://www.dagstuhl.de/16161`

tional argumentation.[3] In Section 3.1, we introduce existing corpora and a taxonomy for highlighting the (often subtle) differences between the employed annotation schemes. In Section 3.2, we provide an overview of current approaches on argument mining. In Section 3.3, we focus on argument attribution like stance recognition and approaches on argument quality. Finally, we discuss the similarities and differences between argument mining and discourse analysis (Section 3.4).

3.1 Existing Corpora

Using annotated corpora is a common practice in natural language processing. On the one hand, supervised natural language processing methods learn a particular task by means of labeled training instances in order to generalize the task to unseen texts. On the other hand, annotated corpora are crucial for evaluating and comparing natural language processing methods.

Corpora for computational argumentation are still very rare. Existing resources focus primarily on argument mining and include annotations of arguments, their components and structures. These corpora are usually tailored to a particular task, employ different annotation schemes, contain particular text types or exhibit different granularities of arguments or argument components. In order to illustrate these differences, we investigate existing corpora by means of the following taxonomy:

- *Tasks*: An argument mining system requires several subtasks (Peldszus and Stede, 2013a). However, not all corpora are appropriate for all tasks due to the type of text or the employed annotation scheme. We investigate the suitability of existing corpora for (*i*) argument component identification (CI) that requires the presence of non-argumentative text units, (*ii*) argument component classification (CC) which requires the annotation of different argument component types, (*iii*) structure identification (SI) that presupposes argumentative relations either between argument components or between full arguments, and (*iv*) argument attribution (AT) that postulates annotations of argument's properties.

- *Domain*: Describes the source of the texts and the types of documents in the corpus respectively.

- *Language (Lang)*: Most corpora in computational argumentation are in English. There are only few resources in other languages like German and Greek.

- *Argument granularity (ArgGran)*: Existing corpora exhibit different granularities of arguments. Resources addressing the macro-level include properties of arguments or relations between arguments. Corpora focusing on the micro-level of arguments include fine-grained annotations of argument components such as claims and premises. We refer to this difference as macro-level and micro-level.

[3] Note that we also include approaches that emerged during the writing of the current thesis.

- *Component granularity (CompGran)*: The granularity of argument components differs in existing resources. There are micro-level corpora with clause-level, sentence-level or multi-sentence components. Some corpora also include components of different granularities, e.g. clause-level and multi-sentence components.

- *Single versus multi-document (SiMuDo)*: Most existing corpora contain argumentation structures at the discourse-level and thus annotations of argument components and structures in a single document. However, a few resources include argumentation structures over several documents, i.e. the premises are in other documents than the claim.

- *Reliability*: If the quality of the corpora is evaluated, we report the reliability scores, i.e. inter-annotator agreement (Artstein and Poesio, 2008). These scores show how reliable the annotation scheme can be applied by human annotators and constitute the upper bound for computational methods.

In addition, we report the number of documents (*#Doc*) and the number of argument components (*#Comp*). A tabular summary of existing corpora can be found in Appendix B.

The remainder of this section is structured as follows: In Section 3.1.1, we review existing corpora focusing on the macro-level. These resources include annotation of argument properties or annotations of argumentative relations between full arguments for studying the structure of dialogical communications. In Section 3.1.2, we introduce corpora containing annotations of argument components at the micro-level. In Section 3.1.2, we provide an overview of corpora annotated with argumentation structures at the micro-level, i.e. argumentative relations between argument components. Finally, we introduce existing corpora of student essays that are related to argumentation analysis in Section 3.1.4.

3.1.1 Macro-level Argument Corpora

The following corpora address the macro-level of arguments and do not include annotations of argument components at the micro-level.

A frequently used source for argument corpora are debating portals such as *iDebate.org* or *ProCon.org*. They are particularly suited for studying the macro-level perspective of argumentation since they organize arguments in a predefined structure (tree-like thread structure). Debating portals also encode the stance of each argument, i.e. an attribute that indicates if the argument is for or against the controversial topic.

Somasundaran and Wiebe (2010) collected arguments from several debating portals such as *OpposingViews.com*, *CreateDebate.com* and *ForAndAgainst.com*. In total, they collected 3,921 arguments about six different topics, e.g. "*healthcare*", "*gun rights*", "*abortion*", etc. Similarly, Anand et al. (2011) collected 4,873 arguments about twelve different topics from *ConvinceMe.net*. Each argument in these corpora includes a stance attribute that was automatically extracted from the debating portal. Thus, both corpora are suited for argument attribution and, more specifically, stance recognition. However, the reliability of the automatically extracted stance

attributes is questionable. For instance, Anand et al. (2011) found that human annotators agree only in 78.26% on the automatically extracted stance from debating portals ($\kappa = .27$).

The internet argument corpus (IAC) introduced by Walker et al. (2012) includes 11,800 discussions with more that 390k posts from *4forums.com* organized as quote-response pairs. They employed crowdsourcing for annotating a subset of the corpus with several attributes like cordiality, assertiveness, emotionality, sarcasm, topic, undercutting, etc. Due to this wide range of different properties, the corpus is particularly suited for argument attribution at the macro-level.

Corpus	*Domain*	*Lang*	*Task*	*#Doc*	*Reliability*
Anand et al. (2011)	online discussions	en	AT	4,873	$\kappa = .27$
Boltužić and Šnajder (2014)	online discussions	en	SI	pairs 2,436	$\kappa = .49$
Cabrio and Villata (2012a)	online discussions	en	AT+SI	pairs 200	-
Cabrio and Villata (2014)	several	en	SI	pairs 792	$\kappa = .71$
Florou et al. (2013)	focused web crawl	gr	-	677	-
Somasundaran and Wiebe (2010)	online discussions	en	AT	3,921	-
Walker et al. (2012)	online discussions	en	AT	390k	$.22 \leq \kappa \leq .62$

Table 3.1: Argument corpora at the macro-level.

Other argument corpora include argumentative relations at the macro-level, i.e. relations between complete arguments. These resources can be employed for recognizing argumentation structures between arguments in dialogical communication.

Cabrio and Villata (2012a) exploited the thread structure of debating portals to recognize support and attack relations between arguments. They extracted 200 argument pairs about various topics from *Debatepedia*. Furthermore, they considered arguments that are not attacked as undisputed. Therefore, their corpus can be employed for argument attribution and structure identification at the macro-level.

The *ComArg* corpus links user comments from a debate portal to predefined arguments. Boltužić and Šnajder (2014) extracted 373 user comments about the topics "*Under God in Pledge*" and "*Gay Marriage*" from *ProCon.org*. They linked each user comment to an argument from *iDebate.org* with a support or attack relation. They manually annotated 2,436 pairs and achieved an inter-annotator agreement of $\kappa = .49$.

Cabrio and Villata (2014) introduced the NoDE corpus (natural language arguments in online debates). It includes 792 argument pairs annotated as support or attack. They extracted the pairs from Debatepedia, a script of a play called "12 Angry Men" and Wikipedia revision histories. In their annotation study, they achieved an inter-annotator agreement of $.7 \leq \kappa \leq .74$ depending on the source of the argument pairs. Their data set is adequate for studying the relations between arguments in debates and written dialogs.

Florou et al. (2013) collected 677 Greek text segments crawled with a focused crawler. They manually annotated each text segment as being argumentative or non-argumentative. Therefore, the corpus can be used to study the difference between argumentative texts and non-argumentative texts at the macro-level.

Table 3.1 shows an overview of argument corpora. The resources can be used for argument identification, argument attribution or the identification of argumentative relations between full arguments at the macro-level.

3.1.2 Micro-level Argument Corpora

Corpora annotated with argument components address the microstructure of arguments. They include annotations of different argument components such as claims and premises and allow for a fine-grained analysis of arguments in text. In this section, we highlight the differences in the applied argumentation schemes, i.e. the types of argument components and their granularity.

Kwon et al. (2007) annotated the main claim in English online comments about the emission standard rules proposed by the environmental protection agency (EPA). They annotated claims at the sentence-level in 119 documents and achieved an agreement of $\kappa = .62$ with two annotators. Furthermore, they annotated each claim as support, oppose, or propose with an inter-annotator agreement of $\kappa = .80$. They found that 59% of the claims are opposing claims, 7% are supporting claims and 34% are proposing claims. Their corpus includes non-argumentative sentences and different types of argument components. It is therefore usable for argument component identification and component classification.

The ECHR corpus contains legal cases of the European court of human rights annotated with argument components (Mochales-Palau and Moens, 2008). The annotation scheme includes claims and supporting or opposing premises. In their first annotation study, Mochales-Palau and Moens (2008) annotated 10 documents and obtained an inter-annotator agreement of $\kappa = .58$. In a subsequent study, they extended their experiments to 47 documents and achieved an inter-annotator agreement of $\kappa = .75$ (Mochales-Palau and Moens, 2009). The final corpus includes 1,449 non-argumentative sentence and 1,067 argumentative sentences. The argumentative sentences include 304 claims and 763 premises. This proportion indicates that arguers in legal cases provide several reasons per claim for ensuring a robust standpoint.

Biran and Rambow (2011a) annotated claims and premises (justifications) in 309 blog threads from *LiveJournal.com* (a virtual and informal blog community). The corpus contains 1,377 multi-sentence argument components. They achieve an inter-annotator agreement of $\kappa = .69$ among two annotators. In subsequent work, they applied their annotation scheme to 118 Wikipedia talk pages (Biran and Rambow, 2011b). They annotated 2,404 argument components and obtained an inter-annotator agreement of $\kappa = .75$. Both corpora contain non-argumentative text units and different types of argument components, i.e. claims and premises.

Rosenthal and McKeown (2012) created a corpus of 285 blogposts collected from LiveJournal.com and 51 Wikipedia discussion pages. Two annotators identified claims at the sentence-level and reached an agreement of $\kappa = .53$ ($\kappa = .5$ on LiveJournal and $\kappa = .557$ on Wikipedia discussion forums). The corpus is appropriate for component identification but not for component classification since it includes only a single type of argument component.

Sardianos et al. (2015) annotated argument components in 300 news articles written in Greek. Two annotators labeled claims and premises at the clause-level and achieved an agreement of $F1 = .76$. In total, the corpus contains 1,191 argument components. It can be employed for component identification as well as for component classification.

Goudas et al. (2014) introduced a Greek corpus with 204 documents about re-

newable energy collected from social media. It contains documents from different sources such as news, blogs and microblogs. The corpus comprises 16k sentences of which 760 sentences include claims and premises at the clause-level. The authors do not report agreement scores. The corpus can be applied for identifying and classifying argument components.

All corpora described above contain annotations of argument components in single documents, i.e. all components of an argument are encapsulated in the same document. However, it is worthwhile to identify argument components from several documents, e.g. in order to collect evidence for or against a given claim. Therefore, Aharoni et al. (2014) created a corpus that contains claims and premises at the clause-level over multiple Wikipedia articles. Starting with a set of 33 topics from *iDebate.org*, 20 annotators selected 1,392 related claims from Wikipedia articles with an inter-annotator agreement of $\kappa = .39$. Subsequently, they annotated 1,291 associated premises. They classified each premise as study (quantitative analysis), expert (testimony by a person) or anecdotal (specific events) and achieved an inter-annotator agreement of $\kappa = .40$. The data set is continuously extended in subsequent work at IBM (Rinott et al., 2015). The current version includes 58 topics, 547 documents, i.e. Wikipedia articles, annotated with 2,294 claims and 4,960 associated premises. The corpus is particularly suitable for information retrieval tasks. It can be used to train supervised machine learning models that identify evidence for a given claim in multiple documents. Furthermore, the different argument component types enable the development of component classification methods.

Habernal and Gurevych (2016a) presented a corpus of user-generated web content (blog posts, forum posts, user comments, etc.) annotated with a modified Toulmin model (cf. Section 2.1.1). First, three annotators annotated 990 documents as argumentative (on-topic persuasive) or non-argumentative (non-persuasive) and achieved an inter-annotator agreement of $\kappa = .59$. Subsequently, they annotated 340 argumentative documents with multi-sentence claims, premises, backings, rebuttals and refutations. They achieved an average inter-annotator agreement of $\alpha_U = .48$ across different topics. The corpus is appropriate for identifying arguments at the macro-level and also allows for a more fine-grained analysis of argument components and their types at the micro-level.

A drawback of all corpora described above is that they do not model the targets of argument components, i.e. the target argument component of a particular premise. Consequently, it is not possible to separate several arguments in a document or to model serial argumentation structures (cf. Section 2.1.3). In order to model the targets of premises, Eckle-Kohler et al. (2015) proposed an annotation scheme that indicates if a premise refers to a preceding or following claim. In addition, their scheme distinguishes between supporting and attacking premises. They applied their annotation scheme to 88 German news articles collected with a focused crawler. Three annotators annotated 1,708 multi-sentence argument components (74% of the tokens are argumentative) and reached an agreement of $\alpha_U = .40$. However, the annotation scheme is limited to convergent argument structures and does not model serial argumentation structures. In addition, the annotation scheme fails to model the targets of premises, e.g. if several independent reasons follow two adjacent claims.

Table 3.2 provides an overview of the discussed corpora. They span a wide va-

Corpus	*Domain*	*Lang*	*Task*	*CompGran*	*#Doc*	*#Comp*	*Reliability*
Biran and Rambow (2011a)	blog threads	en	CI+CC	multi	309	1,377	$\kappa = .69$
Biran and Rambow (2011b)	Wikipedia talk pages	en	CI+CC	multi	118	2,404	$\kappa = .75$
Eckle-Kohler et al. (2015)	news	de	CI+CC	multi	88	1,708	$\alpha_U = .40$
Goudas et al. (2014)	social media	gr	CI+CC	clause	204	760	-
Habernal and Gurevych (2016a)	web content	en	CI+CC	multi	340	1,319	$\alpha_U = .48$
Kwon et al. (2007)	online comments	en	CI+CC	sentence	119	240	$.62 \leq \kappa \leq .80$
Mochales-Palau and Moens (2009)	court cases	en	CI+CC	sentence	47	1,067	$\kappa = .75$
Rinott et al. (2015)	Wikipedia articles	en	CI+CC	clause	547	7,254	$\kappa = .39$
Rosenthal and McKeown (2012)	blogs and discussions	en	CI	sentence	336	2,479	$\kappa = .53$
Sardianos et al. (2015)	news	gr	CI+CC	clause	300	1,191	$F1 = .76$

Table 3.2: Corpora annotated with argument components at the micro-level.

riety of text types including news, legal documents, online discussions and different types of user-generated web content. The investigation of the employed annotation schemes shows that the claim-premise-scheme is the most frequent one. The reported reliability scores range from moderate to substantial agreement. Thus, the overview provides some evidence that the claim-premise-scheme can be reliably applied to heterogeneous types of text. A drawback of the applied annotation schemes is that none of the corpora listed above explicitly models the relations between argument components in single documents. However, knowing the targets of argument components and the structure of arguments respectively is crucial for argument analysis (Henkemans, 2000, p. 448; Govier, 2010, p. 22; Sampson and Clark, 2006; Sergeant, 2013). First of all, the structure of arguments is essential for evaluating the quality of arguments since it is not possible to examine how well a claim is justified without knowing which premises belong to it. Second, solely modeling the types of argument components is not sufficient for recognizing more complex argument structure, i.e. serial arguments. Third, the structure is required to separate several arguments within a single text. Without knowing if and how argument components are related to each other, it is not possible to group the components of an individual argument.

3.1.3 Argumentation Structures

Corpora annotated with argument structures at the micro-level are sill very rare. Complementary to the corpora introduced in Section 3.1.2, they link argument components with annotated argumentative relations for modeling the internal microstructure of arguments.

One of the first resources annotated with argument structures is AraucariaDB (Reed et al., 2008). The corpus consists of more than 700 argument analyses and includes heterogeneous text types such as newspaper editorials, parliamentary records, judicial summaries and discussions. The annotation scheme structures arguments as trees and distinguishes between claims and premises. The analysts employed a graphical argument diagramming tool (Araucaria) for the annotation task (Reed and Rowe, 2004). A special feature of this corpus is that it includes implicit argument components that are not present in the original text. These have been added by the annotators during the analysis for reconstructing enthymemes (cf. Section 2.1). In addition, the corpus contains annotations of argumentation schemes and thus the reasoning type of each argument (cf. Section 2.1.2). The original version of AraucariaDB includes approximately 700 argument analyses. In more recent work, the

Center of Argumentation Technology[4] at the University of Dundee provided several web services for creating argument diagrams. However, compared to the original version of the corpus, the provided diagrams do not include the source documents and are thus not usable for identifying argument components (Lippi and Torroni, 2016). In addition, the reliability of the annotations is unknown.

Peldszus and Stede (2016) created a corpus of German documents of controlled linguistic and rhetoric complexity (microtexts). Each document includes a single argument and does not exceed five argument components. The annotation scheme models the argument structure by means of argumentative support, attack and undercut relations. Furthermore, it distinguishes between claims and premises and additional properties of argument components like proponent/opponent, normal/example and rebut/undercut. In a first annotation study, 26 naïve annotators applied the scheme to a subset of 23 microtexts in a classroom annotation experiment yielding an agreement of $\kappa = .38$ (Peldszus and Stede, 2013b). In subsequent work, they extended their corpus and obtained an inter-annotator agreement of $\kappa = .83$ among three expert annotators using the full label set. Recently, they translated the corpus to English resulting in the first parallel corpus in computational argumentation. It includes 112 arguments and documents respectively (Peldszus and Stede, 2016). The corpus is particularly suited for studying the structure of arguments and to differentiate supporting and attacking argument components. In addition, it is the only resource for studying argumentation structures in different languages.

Kirschner et al. (2015) annotated argument structures in introductions and discussion sections of 24 German scientific articles from the educational domain. Their annotation scheme includes four argumentative relations (support, attack, detail and sequence). Considering argumentative relations less distant than 6 sentences, they achieved an agreement of $\kappa = .43$ with four annotators. However, the granularity of the argument components is limited to sentences and it does not include annotations of argument component types.

Corpus	*Domain*	*Lang*	*Task*	*CompGran*	*#Doc*	*#Comp*	*Reliability*
Kirschner et al. (2015)	scientific articles	de	CI+SI	sentence	24	~2,700	$\kappa = .43$
Peldszus and Stede (2016)	microtexts	de/en	CC+SI	clause	112	576	$\kappa = .83$
Reed et al. (2008)	various	en	CI+CC+SI	clause	~700	~2,000	-

Table 3.3: Corpora annotated with micro-level argument structures.

Table 3.3 provides an overview of the existing corpora annotated with micro-level argument structures. The few existing resources either lack non-argumentative text units and are fairly small (Peldszus and Stede, 2016), include sentence-level argument components without argumentative types (Kirschner et al., 2015), or the reliability of the annotations is unknown (Reed et al., 2008).

3.1.4 Student Essays

Student essays are extensively studied in computational linguistics. For instance, there are various corpora in the field of automated essay grading (Shermis and Burstein, 2013a; Attali et al., 2013). The publicly available ETS corpus of non-native

[4] `http://www.arg-tech.org`

written English includes 12,100 English essays written by speakers of eleven different mother tongues (Blanchard et al., 2013). The corpus was originally designed for language identification experiments but also includes holistic scores on a three-grade scale. The international corpus of learner English (ICLE) contains 6,085 essays (Granger et al., 2009) written in response to varying prompts. The essays are written in English by undergraduate students of 16 non-English mother tongues. Most of the essays (91%) are written in response to argumentative prompts (Persing and Ng, 2015). However, none of these corpora includes annotations of argument components or argument structures.

Although the prompts for student essay writing frequently demand an argumentative writing style, there are relatively few essay corpora that include annotations related to argumentation. One of the few examples is the corpus created by Song et al. (2014). They selected three argumentation schemes (cf. Section 2.1.2) related to two specific prompts and derived 16 labels from the associated critical questions. Each label signals a particular critical question about the justifications in arguments. They applied the annotation scheme to 600 essays and achieved an inter-annotator agreement of $\alpha = .33$ to .85 depending on the particular label. Furthermore, they showed that some of the labels contribute significantly to essay scores and thus to the holistic quality of essays.

Persing and Ng (2015) investigated the argumentation strength of student essays. They employed a subset of 1,000 student essays from the ICLE corpus and annotated the argumentation strength as a numerical score from one to four at a half-point scale. The annotators were selected as the six most consistent ones from over 30 applicants. The evaluation of the agreement scores showed that the annotators agree only in 26% on the exact score. Evaluating the agreement scores within 0.5 point yields an agreement of 67% and within 1.0 point 89%. However, holistic scores of the argument strength do not pinpoint particular weaknesses of the individual arguments and thus are only of limited use for providing feedback about arguments.

Falakmasir et al. (2014) annotated thesis statements and conclusions, i.e. restatements of thesis statements at the sentence-level. They employed 432 essays from eight different writing assignments. For ensuring the reliability, they repeated the annotation until the agreement reached $\kappa > 0.6$ among two annotators. However, thesis statements are only one particular type of argument components in essays.

Other works on student essays investigated shell expressions (Madnani et al., 2012), style criteria (Burstein and Wolska, 2003), metaphors (Beigman Klebanov and Flor, 2013), usage of factual knowledge (Beigman Klebanov and Higgins, 2012), thesis clarity (Persing and Ng, 2013), or prompt adherence (Persing and Ng, 2014). However, none of the corpora above include fine-grained annotations of micro-level argumentation structures. We are only aware of one study of argumentation structures in persuasive essays. Botley (2014) analyzed 10 English essays written by Malay L1 students. They manually applied argument diagramming for analyzing different argumentation strategies. However, the data set is too small for computational purposes. In addition, they did not conduct an annotation study with several annotators. Therefore, it is unknown if the applied annotation scheme can be reliably applied to student essays.

Having introduced existing corpora in computational argumentation, we provide an overview of existing methods for recognizing arguments, their components and

structures in the following section.

3.2 Argument Mining

Argument mining focuses on the identification of arguments, their components and relations in natural language texts. In this section, we review existing computational approaches in argument mining and categorize them by means of the following subtasks:

1. *Identification*: The identification task involves filtering of non-argumentative documents, the separation of argumentative from non-argumentative text units, and the identification of argument component boundaries. The results are either a set of argumentative documents (e.g. argumentative comments in online discussions) or a set of argument components of different granularities (e.g. argumentative sentences or clauses). We introduce existing approaches for identifying arguments and their components in Section 3.2.1.

2. *Classification*: The classification task addresses the functional analysis of argument components. It attempts to detect the type of argument components in its argumentative context. Existing approaches aim, for instance, at discriminating claims from premises or recognizing different types of evidence. We review existing approaches in Section 3.2.2.

3. *Linking*: The linking task focuses on recognizing argumentative relations. The objective depends on the considered argument granularity of the underlying argument model. At the micro-level, the linking task addresses relations between argument components and the microstructure of arguments respectively. At the macro-level, it aims at recognizing relations between complete arguments in order to analyze interactions of several interlocutors. Some approaches also discriminate between different relation types such as support, attack or counter-attack. We discuss existing approaches in Section 3.2.3.

All tasks are closely related and there are systems that solve several of these objectives in a single model. For instance, a system could classify sentences as claim, premise or non-argumentative for identifying argument components and classifying their types simultaneously. Accordingly, the categorization of some approaches by means of the common tasks might be ambiguous. However, in order to bring some order to the various proposals, we categorize each approach by means of its "most upstream capability". Thus, we discuss an approach that differentiates several argumentative types in the classification section even if it also separates non-argumentative from argumentative text units. This also implies that approaches categorized as linking task might not solve all downstream tasks required for an end-to-end argument mining system. In all of these cases, we will point out the missing steps.

3.2.1 Identifying Arguments and their Components

The identification of arguments and their components is the first step of an argument mining system. This task is usually considered as a binary classification of

documents or sentences. Approaches that aim at identifying argument components at the clause-level usually apply sequence labeling for recognizing the boundaries of argument components at the token-level.

Moens et al. (2007) separated argumentative from non-argumentative sentences in AraucariaDB. They experimented with lexical features (unigrams, bigrams, verbs and word pairs), structural features (sentence length, number of punctuation marks, and average word length), keywords that signal arguments and syntactic features (parse tree depth and number of subclauses). They obtained the best accuracy of .738 with a multinomial naïve Bayes classifier using word pairs, text statistics, verb, and keyword features.

Florou et al. (2013) proposed an approach for recognizing arguments in a corpus of Greek text segments. They classified each text segment as containing an argument or not. They tried several shallow features such as discourse markers and the tense and mood of verbs. The best performing model based on a C4.5 decision tree achieved an F1 score of .764.

Rosenthal and McKeown (2012) introduced an approach for identifying claims in online discussions. The approach is based on logistic regression implemented in Weka. They experimented with sentiment, committed belief (Prabhakaran et al., 2010), lexical, syntactic, question (presence of a question mark), and social media features. They showed that the predictiveness of features differs in various domains. They showed that sentiment and syntactic features are useful in online comments, while n-grams and committed beliefs are useful in Wikipedia.

Roitman et al. (2016) proposed a claim-oriented document retrieval approach for identifying Wikipedia articles that are relevant to a given controversial topic. They considered this retrieval problem in two consecutive steps. First, they identified articles belonging to the topic using a state of the art information retrieval method. Second, they manually created a lexicon including words which may signal a controversy. For example, the lexicon includes words like "*dispute*", "*prove*", "*justify*", etc. Based on this lexicon, they extracted several features and re-ranked the relevant documents accordingly. They showed that their approach improves the recall of document and claim retrieval by 10%.

Levy et al. (2014) proposed a pipeline including three consecutive steps for identifying the precise boundaries of context-dependent claims in Wikipedia articles. Their first component detects sentences that contain a relevant claim of the given topic. The second component segments the sentences in order to detect the boundaries of the claims. It generates several candidate clauses using a maximum likelihood model and selects the most probable claim with a logistic regression classifier. The third component ranks the identified claims using another logistic regression classifier in order to identify the most probable claims for the given topic.

Instead of using several consecutive steps, Lippi and Torroni (2015) proposed an approach for recognizing claims in Wikipedia articles using a local tree kernel that measures the similarity between two sentence-level parse trees. Though their approach operates locally without considering the controversial topic, they slightly improved the result reported by Levy et al. (2014) and achieved 16.8 F1@200.

Goudas et al. (2014) presented a two step approach for identifying argument components in Greek social media texts. They classified each sentence as argumentative or non-argumentative and achieved an accuracy of .774 using logistic regression. The

second step aims at identifying the boundaries of argument components using an IOB-tagset (Ramshaw and Marcus, 1995). Their best system achieves an accuracy of .424 using conditional random fields.

3.2.2 Classifying Argument Components

The classification task addresses the detection of argument component types, i.e. the argumentative function of argument components in their specific context. Existing approaches differentiate, for instance, between claims and premises or different types of evidence.

Kwon et al. (2007) proposed an approach for identifying supporting, opposing and proposing claims in online comments using two consecutive steps. First, they identified claims at the sentence-level using binary classification. As features they employed n-grams, positive and negative words from the General Inquirer, FrameNet frames, and syntactic features extracted from parse trees. They achieved the best F1 score of .55 by using a boosting algorithm from Schapire and Singer (2000). Second, they employed the same features for classifying the claims as support, oppose or propose and achieved an F1 score of .67.

Rooney et al. (2012) applied kernel methods for classifying argument components in AraucariaDB. They classified each proposition as claim, premise or non-argumentative and report an overall accuracy of .65.

Mochales-Palau and Moens (2011) applied two binary classifiers for classifying sentences of the ECHR corpus as claim or premise. They experimented with domain-dependent key phrases, token counts, location features, information about verbs, and the tense of the sentence. They achieved an F1 score of .741 for claims and .681 for premises using a support vector machine.

Habernal and Gurevych (2016a) proposed an approach for identifying claims, premises, backings, rebuttals and refutations in user-generated web discourse (for a description of the corpus see Section 3.1.2). They first trained a support vector machine for separating argumentative from non-argumentative documents. Their best system achieved an F1 score of .69 using lexical features only. In a second step, they identified the argument components and their boundaries using a sequence model. Similar to the approach from Goudas et al. (2014), they employed an IOB-tagset for identifying the boundaries of components. However, they also encoded the type of the components in the tagset, which yields a total of 11 labels (two labels for marking the beginning (B) and the inside (I) for all five component types and a single label (O) for marking non-argumentative tokens). They employed a structural support vector machine (Joachims et al., 2009) with lexical, structural, syntactic, topic, sentiment, semantic, discourse and embedding features. The approach solves several tasks in one model. It separates argumentative from non-argumentative text units, identifies the boundaries of argument components and recognizes the types of argument components simultaneously. They achieved an F1 score of .251 using a combination of topic, sentiment, semantic, discourse and embedding features.

Sardianos et al. (2015) identified claims and premises in Greek news articles using conditional random fields. They employ unigrams, part-of-speech tags and manually defined cue words. Furthermore, they showed that news articles outperform blogs, comments from facebook and tweets for training word embeddings. They showed

that extending the list of cue words improves the results of their approach to .325 (F1 score).

Rinott et al. (2015) proposed an approach for identifying different types of evidence in Wikipedia. Since each evidence can include several sentences, they first applied a context-independent component that splits each article in segments of up to three sentences. Subsequently, they assess the evidence characteristic of each segment. The second part of the system is context dependent. It considers the topic of the debate and a given claim in order to rank the retrieved evidence. They trained the system for three different types of evidence (study, expert and anecdotal) and achieved macro averaged mean reciprocal ranks between .03 and .2 depending on the type of evidence.

Falakmasir et al. (2014) introduced a system for identifying thesis statements and conclusions in student essays. They consider this task as a multiclass classification problem and experimented with various features. They found that structural features and lexical overlap with the essay prompt are the most predictive features for separating thesis statements and conclusions from other sentences. They achieved the best F1 score of .83 using a decision tree.

3.2.3 Recognizing Argumentative Relations

Arguments (and their components) are usually embedded in a particular situation and cannot be understood in isolation (Peldszus and Stede, 2013a; Sergeant, 2013). However, all approaches presented so far identify arguments and their components without establishing a connection to their context. In this section, we review existing works that address this shortcoming. These approaches focus on the identification of argumentative relations (*i*) between full arguments for analyzing the interactions of several interlocutors or (*ii*) between argument components for recognizing the internal microstructure of arguments. Both tasks are usually considered as pair classification of arguments or argument components.

Cabrio and Villata (2012a) identified argumentative support and attack relations between arguments of a debating portal. They extracted pairs of arguments from Debatepedia and trained an existing textual entailment platform (EDITS)[5]. The system achieved .67 accuracy. They further employed the approach for identifying undisputed arguments. We will discuss this aspect of their work in Section 3.3.2 since it focuses on the quality of arguments.

Ghosh et al. (2014) presented an approach for identifying the targets of arguments in online discussions. They classified argument-target-pairs as agree or disagree using a support vector machine. They experimented with sentiment and several lexical features and achieved an F1 score of .669 and .626 for the agree and disagree category respectively. However, their approach presupposes that the targets and arguments are identified in previous analysis steps.

Boltužić and Šnajder (2014) linked user comments to a set of predefined arguments with support or attack relations. They experimented with textual entailment features extracted with the excitement open platform[6], semantic text similarity and a "stance alignment" feature set to true if the stance of the comment is the same as

[5] `http://edits.fbk.eu`

[6] `http://hltfbk.github.io/Excitement-Open-Platform`

the stance of the predefined argument. They achieved an F1 score of .818 using a support vector machine with textual entailment and semantic similarity features.

The approaches described above identify argumentative relations between arguments at the macro-level for analyzing the interactions between several interlocutors. In the following, we introduce existing approaches for identifying argumentative relations between arguments components which allows for a more fine-grained analysis of argumentation structures at the micro-level.

One of the few approaches focuses on the identification of argument structures in legal argumentation and, more specifically, in legal cases of the European Court of Human Rights. Mochales-Palau and Moens (2009) created a context-free grammar (CFG) based on manually defined rules and lexical indicators. This approach allows for identifying argument structures as trees and is specifically tailored to documents from the legal domain since it relies on the presence of domain-dependent key words (Wyner et al., 2012). However, it achieves a promising accuracy of .60 for recognizing argumentation structures, an F1 score of .673 for classifying claims and .640 F1 score for identifying premises.

Peldszus (2014) encoded the targets of argumentative relations along with additional information like opponent, proponent, support and attack, etc. in a single tagset. It includes, for instance, tags which denote if an argument component at position n is argumentatively related to the preceding argument components $n-1$, $n-2$, etc. or the following argument components $n+1$, $n+2$ etc. This approach achieves a promising accuracy of .48 for the 16 target tags and .39 for the entire tagset including 48 tags. However, the approach is only applicable to relatively small texts since the tagset will increase tremendously if the texts become longer.

Very recently Peldszus and Stede (2015) presented the first approach for globally modeling several aspects of argument structures. Instead of classifying argumentative relations between pairs of argument components, the model aims at globally optimizing the predictions of several local classifiers using a minimum spanning tree (MST) algorithm. They jointly modeled several aspects of the argumentation structure and found that the function (support or attack) and the role (opponent and proponent) are the most beneficial dimensions for improving the argumentation structures. They report an F1 score of .720 for identifying argumentative relations and .869 for recognizing claims in their microtext corpus (cf. Section 3.1.3). However, their approach is not capable of separating several arguments in a text and it presupposes that the boundaries of argument components are known in advance.

3.3 Argument Attribution

Argument attribution focuses on identifying properties of arguments or their components. For instance, Feng and Hirst (2011) proposed an approach for identifying the five most frequent argumentation schemes in AraucariaDB (argument from example, argument from cause to effect, practical reasoning, argument from consequences and argument from verbal classification). They experimented with several classification setups and achieved an accuracy between 62.9% and 97.9% using a binary C4.5 decision tree for each argumentation scheme. However, their approach is based on features extracted from mutual information of claims and premises and thus requires

that the argument components are reliably identified in advance.

Oraby et al. (2015) addressed the argumentation style of arguments. They employed the IAC corpus and classified each argument as factual or emotional for separating arguments with an argumentative merit from those which are based on emotional reasons. They employed a bootstrapping approach for extracting linguistic patterns from unlabeled arguments and achieved .80 accuracy. Although this approach increases precision, it exhibits a significantly lower recall compared to a supervised unigram baseline.

Wachsmuth et al. (2015) focused on the identification of sentiment flows in product reviews. They modeled the sentiment flow of a review as a sequence of local sentiments, i.e. sentiment scores of discourse units, in order to identify common argumentation patterns. They showed that reviews across several domains exhibit similar sentiment flows. In later work they extended their approach to rhetorical moves including sequences of discourse relations, discourse functions and argumentative roles (Wachsmuth and Benno, 2016). Their results suggest that flow patterns generalize well over different text types such as product reviews and student essays. Furthermore, they showed that features derived from flow patterns improve the classification of global sentiments and the scoring of the essay organization.

Besides these approaches, there are several approaches on stance recognition and only a few on the quality of arguments. We introduce both in the following two sections.

3.3.1 Stance Recognition

Stance recognition aims at identifying the stance of an author about a controversial topic. This task is usually considered as labeling an author's comment in an online debate as either for or against.

Somasundaran and Wiebe (2009) proposed an approach for maximizing the overall side-score of a comment by using integer linear programming since a single comment can also include concessions or statements opposing the view of the author. They identified the probability that a particular term is positively or negatively associated with the topic by extracting subjectivity clues and associating them with targets from topic-relevant documents. In addition, they considered concessions recognized with a list of discourse constructs. In their experiments, they achieved accuracies between .611 and 1.0 depending on the four topics of 117 comments in their test set.

In their following work, Somasundaran and Wiebe (2010) experimented with clue words and sentiment features using a supervised classifier. They extracted 3,094 positive and 668 negative words from the annotations of the MPQA corpus (Wilson et al., 2005) and showed that combining them with content words yields promising results. They achieved the best results of .639 accuracy by using a support vector machine and a combination of sentiment and clue word features.

In addition, there are several other approaches on stance recognition. For instance, Anand et al. (2011) experimented with lexical, structural, dependency and context features, and Hasan and Ng (2012) showed that jointly modeling contextual information and author's stances on particular subtopics improves the accuracy. Hasan and Ng (2014) found that combining reason classification and stance

classification yields promising results, and Qiu and Jiang (2013) proposed a novel generative latent variable model to capture the viewpoint, user identity and user interactions.

3.3.2 Argument Quality

Approaches for automatically assessing the quality of arguments are still very rare. Cabrio and Villata (2012a) employed textual entailment to identify accepted arguments in online communities. They built a graph that represents attack and support relations between arguments and applied the abstract argumentation framework proposed by Dung (1995) to identify accepted arguments. They report an accuracy of 75% for identifying accepted arguments. However, their approach focuses on the acceptance of entire arguments (macro-level) instead of the acceptability of individual premises as required by the RAS-criteria (cf. Section 2.2.4). In addition, the reliability of their data set is unknown. They consider an argument as accepted if it is not attacked in the debate, i.e. there is no related contra argument. However, it is unclear if humans agree on these automatically extracted labels.

Park and Cardie (2014) presented an approach for classifying argument components as verifiable or unverifiable. Their best approach based on a support vector machine and various features achieved a macro F1 score of .690. Although they claim that verifiability allows for determining appropriate types of premises and consequently the strength of an argument, it is unclear how the verifiability of propositions relates to the logical quality of arguments.

Persing and Ng (2015) introduced an approach for recognizing the argumentation strength of an entire essay. They report that pos n-grams, prompt adherence features (Persing and Ng, 2014), and predicted argument components perform best. However, their system outputs a single holistic score which summarizes the strength of all arguments in an essay. Consequently, it is only of limited use to pinpoint the weak points of arguments in persuasive essays.

3.4 Argumentation and Discourse Analysis

The identification of argument structures is closely related to discourse analysis. Similarly to the identification of argumentation structures, discourse analysis aims at identifying elementary discourse units (EDU) and discourse relations between them. The major challenge for both fields is to identify implicit discourse relations that are not signaled by lexical indicators such as discourse connectives (Braud and Denis, 2014, p. 1694).

Existing computational approaches on discourse analysis differ mainly in the employed discourse theory and the inventory of discourse relations. The two most common approaches in computational linguistics are rhetorical structure theory (RST) (Mann and Thompson, 1987) and the penn discourse tree bank (PDTB) (Prasad et al., 2008). RST represents the discourse structure of a document as a tree by iteratively linking discourse units (Feng and Hirst, 2014; Hernault et al., 2010). The PDTB captures discourse structures in a more shallow representation by only linking adjacent sentences or clauses (Lin et al., 2014).

Marcu and Echihabi (2002) proposed one of the first approaches for identifying implicit discourse relations. In order to collect large amounts of training data, they exploited several discourse connectives such as "because" or "but". After removing the discourse connectives from the training instances, they found that word pair features are indicative for implicit discourse relations. They achieved an accuracy of 75% for identifying cause-explanation-evidence relations (their most similar relation compared to argumentative relations) using a naïve Bayes classifier. Pitler et al. (2009) identified four different types of implicit discourse relations in the PDTB and achieved F1 scores between .22 and .76 depending on the relation type. They found that using a tailored feature set for each individual relation leads to the best results. Lin et al. (2009) showed that besides lexical features, production rules collected from parse trees yield good results and Louis et al. (2010) found that named-entity features do not perform as well as lexical features.

Although the identification of argument structures and discourse analysis share similar subtasks and challenges, there are some major differences between both. First, most existing approaches on discourse analysis are limited to the identification of discourse relations between adjacent text units (Peldszus and Stede, 2013a), whereas argumentative relations may also hold between non-adjacent text units (Stab and Gurevych, 2014a). Second, the employed discourse relations differ. Discourse analysis usually employs a large set of discourse relations to capture general discourse structures. However, only a subset of these relations is relevant for argument structures. For example, Peldszus and Stede (2013a) introduced support, attack and counter-attack relations for modeling argument structures. This difference is best illustrated by the work of Biran and Rambow (2011a). They argue that only a subset of RST relations are relevant for identifying argumentative relations between justifications and claims and selected only a particular subset of 12 relations from the RST discourse treebank (Carlson et al., 2001). Because of these differences, i.e. the focus of discourse analysis on relations between adjacent discourse units and the different sets of relations, existing discourse parsers such as the PDTB-style end-to-end discourse parser[7] introduced by Lin et al. (2014) or the RST parser[8] from Feng and Hirst (2014) are not sufficient for capturing argumentation structures. Therefore, there is a need to develop novel methods, which consider relations between non-adjacent text units and solely focus on discourse units that are relevant to argumentation in order to recognize argumentation structures.

3.5 Chapter Summary

Computational argumentation is a recent and rapidly evolving field in natural language processing which addresses the analysis and identification of arguments in text. In this chapter, we provided a systematic overview of existing corpora and reviewed current computational approaches for analyzing arguments.

In the first part of this chapter, we introduced a taxonomy that allows for investigating existing corpora with respect to their granularity as well as their domain, language and size. Furthermore, we evaluated the appropriateness of existing cor-

[7] `http://wing.comp.nus.edu.sg/~linzihen/parser`
[8] `http://www.cs.toronto.edu/~weifeng/software.html`

pora for different analysis tasks. We found that most corpora do not suffice to train end-to-end argumentation structure parsers because of the following reasons: First and foremost, most of the employed annotation schemes focus on particular subtasks. We found only three corpora annotated with argumentation structures at the micro-level (cf. Table 3.3). However, they either lack non-argumentative text units and are fairly small, do not include fine-grained annotations of argument components, or the reliability of the annotations is unknown (cf. Section 3.1.3).

In the second part of this chapter, we showed that argument mining approaches do not fulfill the requirements for identifying fine-grained argumentation structures (cf. Section 3.2). Existing approaches are either limited in granularity or omit required subtasks such as identifying argument components or separating several arguments. An approach covering all subtasks for identifying fine-grained argumentation structures is still missing (cf. Section 3.2). Furthermore, most approaches operate locally which is not sufficient for recognizing consistent argumentation structures (cf. Chapter 1). We also showed that existing approaches for assessing the quality of arguments are restricted to coarse-grained scores. However, providing adequate feedback to students requires to answer why an argument is fallacious and to point out the weak components. These requirements are not yet fulfilled by current analysis methods.

These limitations give rise to the three main research questions of this thesis (cf. Section 1). In particular, we will investigate if human annotators agree on the argumentation structures in persuasive essays and if it is possible to create reliable corpora for training end-to-end argumentation structure parsers. We approach this research question (RQ1) in the following chapter of this thesis. In addition, we want to examine the automatic identification of arguments. We will investigate which linguistic features are effective for each task of an argument mining system and if joint modeling can be used to recognize consistent argument structures (RQ2). We deal with this research question in Chapter 5. Finally, we seek to develop approaches for assessing the quality of arguments which are capable of recognizing the weak points of the argumentation in persuasive essays. We address this research question (RQ3) in Chapter 6 of the current thesis.

Chapter 4

Annotating Argumentation Structures

In this chapter, we consider the first research question (RQ1) raised in the introduction. As discussed in Chapter 3, there are only few corpora annotated with fine-grained argumentation structures. These, however, do not suffice to train end-to-end argumentation structure parsers for persuasive essays. In fact, it has been largely unknown at the beginning of this research whether human annotators agree on the argumentation structure of persuasive essays and if it is possible to create annotated corpora of high quality. This shortcoming also implies that there is little empirical evidence about the properties of argumentation structures in persuasive essays. To address these open issues, we seek to find answers for the following questions:

1. How can we model the argumentation structure in persuasive essays?
2. Do humans agree on the argumentation structure in persuasive essays?
3. What are the properties of arguments in persuasive essays?

Answering these questions is a first step towards a better understanding of argumentation structures in persuasive essays. In particular, the answer to the second question is essential for the feasibility of argumentative writing support systems, since it will give us an estimate whether it is possible to create annotated corpora. Answering this question will also allow us to determine the human upper bound, which is an essential part for thoroughly evaluating computational methods. Furthermore, the analysis of the argumentation structure enables to derive the requirements for end-to-end argumentation structure parsers.

To analyze argumentation structures in persuasive essays, we first need to define an annotation scheme. As discussed in Chapter 2, argument diagramming is a good basis for modeling arguments in text. To answer the first question, we introduce an annotation scheme for persuasive essays in Section 4.1. For answering the second question, we conduct an annotation study with three annotators and apply the annotation scheme to persuasive essays.[1] We present the results of this

[1] In the course of this thesis, we conducted two annotation studies. Here we present the results of the second annotation study which is based on an elaborated annotation guideline (cf. Appendix D). The results of the first annotation study are described in (Stab and Gurevych, 2014b).

annotation study along with the evaluation of the inter-annotator agreement and the analysis of disagreements in Section 4.2. Finally, we analyze the properties of the annotated argumentation structures in order to derive the requirements for an end-to-end argumentation structure parser in Section 4.3.

4.1 Annotation Scheme

We already discussed the most prevalent argumentation models in Chapter 2. As we have seen, argument diagramming is a good foundation for modeling argumentation structures in text. On the one hand, it allows for modeling the targets of argument components by means of argumentative relations. Because of this feature, argument diagramming also enables to separate several arguments in a text. On the other hand, argument diagramming allows for modeling more complex argumentation structures such as serial structures or chains of attacking argument components.

In this section, we build upon argument diagramming and derive an annotation scheme for modeling argumentation structures in persuasive essays. Our annotation scheme models the structure of arguments as a node-link diagram. Each node represents an argument component and each link represents a directed argumentative relation. However, on closer inspection there are several ambiguities when applying argument diagramming to real texts. First, the distinction between convergent and linked structures causes ambiguities when analyzing real arguments (Henkemans, 2000, p. 448). Second, it is unclear if the argumentation structure should be modeled as graph or as tree. Third, there is disagreement about how to label the argument components in more complex argumentation structures. We discuss each of these questions in the following three sections.

4.1.1 Distinguishing Linked and Convergent Structures

The first decision is whether the annotation scheme needs to distinguish between linked and convergent arguments which is still an ongoing debate in argumentation theory (van Eemeren et al., 1996, p. 176; Freeman, 2011, p. 89; Yanal, 1991; Conway, 1991). Both linked and convergent arguments include two premises for justifying a single claim. The only difference between them is that the premises in linked arguments are only relevant in conjunction, whereas both premises of a convergent argument independently support the claim (cf. Section 2.1.3). From a traditional logic perspective, linked structures indicate deductive reasoning and convergent structures represent inductive reasoning (Henkemans, 2000, p. 453). Although this distinction is theoretically appropriate, Freeman (2011) illustrates that the traditional definition of linked structures causes ambiguities when analyzing real arguments. He suggests a more precise definition that takes the relevance of each premise into account. Yanal (1991) argues that the distinction is equivalent to separating several arguments, and Conway (1991) argues that it can be safely omitted when modeling single arguments. From computational perspective, the task of distinguishing between linked and convergent structures is equivalent to finding groups among premises or classifying the reasoning type as deductive or inductive. Ac-

cordingly, the distinction between linked and convergent structures can be solved in subsequent analysis steps after the components and relations of arguments have been identified. We leave this task for future work.

4.1.2 Argumentation Structure as Tree

Defining the argumentation scheme as a tree structure is a matter of excluding divergent structures, restricting each premise to support (or attack) only one argument component and omitting circular structures. According to Freeman (2011, p. 16) divergent structures are equivalent to several arguments (one for each claim). As a result of this treatment, a great many of textbooks neglect divergent structures (Henkemans, 2000, p. 447; Reed and Rowe, 2004, p. 972). Although most argument mining approaches consider argument structures as trees (Mochales-Palau and Moens, 2009; Cohen, 1987) or allow only one outgoing relation per argument component (Peldszus and Stede, 2015), we argue that this decision requires a careful investigation of the specific text genre. In particular, modeling argument structures as trees might potentially limit the expressiveness of the approach if a particular genre includes many divergent or circular structures. Usually persuasive essays exhibit a common structure. According to various textbooks about essay writing (Whitaker, 2009; Shiach, 2009; Perutz, 2010; Kemper and Sebranek, 2004), the writing process follows a linear procedure. Starting with the formulation of a *thesis statement* in the introduction, each body paragraph should include a single point expressed in a *topic sentence*. The remaining sentences in each body paragraph should provide justifications for convincing the reader of the idea in the topic sentence. Therefore, it is unlikely that persuasive essays include divergent or circular structures, and we assume that modeling the argumentation structure of persuasive essays as a tree is a reasonable decision. Furthermore, an empirical study of argument structures in political speech (which can be generally assumed to exhibit complex argument structures) shows that only 5.26% of the arguments are divergent (Indrajani and Angeline, 2010).

4.1.3 Argumentation Structures and Argument Component Types

Assigning an argumentative type to argument components is unambiguous if the argumentation structure is shallow. For instance, it is obvious that an argument component a_1 is a premise and an argument component a_2 is a claim, if a_1 supports a_2 in a basic structure (cf. Section 2.1.3). However, if the structure is deeper, assigning types becomes ambiguous. Basically there are three different approaches for assigning an argumentative type to the argument components in deeper structures. First, according to Beardsley (1950) a serial argument structure includes an argument component which is both a claim and a premise for another claim. Therefore, the inner argument component has two different argumentative types (*multi-label-approach*). Second, Govier (2010) distinguishes between "main claim" and "subclaim". Similarly, Damer (2009) differntiates between "premise" and "subpremise" in order to label deeper argument structures. Both approaches define a particular label for each level in the argument structure (*level-approach*). Third, Co-

hen (1987) considers only the root node of an argument as a claim and the remaining nodes in the structure as premise (*one-claim-approach*). In order to define an annotation scheme for persuasive essays, we propose a *hybrid approach* that combines the level-approach and the one-claim-approach.

4.1.4 Argumentation Structure of Persuasive Essays

We model the argumentation structure of persuasive essays as a tree structure and employ a level-approach for modeling the first level of the tree and a one-claim-approach for representing the structure of each individual argument, i.e. the subtrees connected to the root node. Accordingly, we model the first level of the argumentation structure with two different argument component types and the structure of individual arguments using argumentative relations.

The *major claim*, the root node of the argumentation structure, represents the author's standpoint on the topic (Whitaker, 2009, p. 7). It is an opinionated statement that is usually present in the introduction and restated in the conclusion of an essay. The body paragraphs of an essay include the actual arguments. They either support or attack the author's standpoint expressed in the major claim. Each argument consists of a claim and several premises. For differentiating between supporting and attacking arguments, each claim includes a *stance attribute* that can take the values "for" or "against".

We model the structure of each argument using a one-claim-approach. The claim constitutes the central component of each argument. The premises are the reasons of the argument. The actual structure of an argument is constituted by directed argumentative support and attack relations. They link each premise either to a claim or to another premise (serial arguments). Note that the equivocal role of the inner premise in a serial argument is implicitly encoded by the structure of the argument (the inner premise exhibits one outgoing relation and at least one incoming relation). The stance of each premise is indicated by the type of its outgoing relation which can either be a support relation or an attack relation.

We refer to this structure as *argumentation structure* since it does not only model a single argument but several arguments in a single tree structure. In contrast to other models (cf. Section 3.1), our annotation scheme also considers the relations of each argument to the author's standpoint, i.e. relations from claims to major claims. By doing so, our annotation scheme combines micro-level and macro-level perspectives and allows for modeling the entire argumentation structure of an essay. In the following, we illustrate our annotation scheme by applying it to an example essay.[2]

Persuasive essays usually include four to six paragraphs (Botley, 2014). The introduction, i.e. the first paragraph, describes the controversial topic and includes the major claim (Whitaker, 2009). The following paragraph illustrates an introduction of an essay about cloning:

> "*Since researchers at the Roslin Institute in Edinburgh cloned an adult sheep, there is an ongoing debate if cloning technology is morally and eth-*

[2] Note that the example essay was written by the author to illustrate all phenomena of argumentation structures in persuasive essays.

> *ically right or not. Some people argue for and others against and there is still no agreement whether cloning technology should be permitted. However, as far as I'm concerned,* ***[cloning is an important technology for humankind]***$_{MajorClaim1}$ *since [it would be very useful for developing novel cures]*$_{Claim1}$*."*

The first two sentences introduce the topic and do not include any argumentative content. The third sentence includes the major claim (boldfaced) and a claim which supports the major claim (underlined). The following body paragraphs of the essay include arguments which either support or attack the major claim and the author's standpoint respectively. For instance, the following body paragraph includes one argument that supports the positive standpoint of the author on cloning:

> "*First, [cloning will be beneficial for many people who are in need of organ transplants]*$_{Claim2}$*. [Cloned organs will match perfectly to the blood group and tissue of patients]*$_{Premise1}$ *since [they can be raised from cloned stem cells of the patient]*$_{Premise2}$*. In addition, [it shortens the healing process]*$_{Premise3}$*. Usually, [it is very rare to find an appropriate organ donor]*$_{Premise4}$ *and [by using cloning in order to raise required organs the waiting time can be shortened tremendously]*$_{Premise5}$*."*

The first sentence contains the claim of the argument. It is supported by five premises in the remaining three sentences (wavy underlined). The second sentence includes two premises of which Premise_1 supports Claim_2 and Premises_2 supports Premise_1. Premise_3 in the third sentence supports Claim_2. The fourth sentence includes Premise_4 and Premise_5, both support Premise_3. The next paragraph illustrates a body paragraph including two arguments:

> "*Second, [scientists use animals as models in order to learn about human diseases]*$_{Premise6}$ *and therefore [cloning animals enables novel developments in science]*$_{Claim3}$*. Furthermore, [parents with no eggs or sperms can have children that are genetically related]*$_{Premise7}$*. [Even same sex couples can have children without the use of donor sperm or eggs]*$_{Premise8}$*. Consequently, [cloning can help human families to get children]*$_{Claim4}$*."*

The first sentence includes the first argument starting with Premise_6 followed by Claim_3. The next two sentences include a premise which supports another claim in the last sentence. Note that both arguments cover different aspects (development in science and cloning humans) which both support the author's standpoint about cloning. This example illustrates that knowing argumentative relations is important for separating several arguments in a paragraph. The example also shows that argument components frequently exhibit preceding text units that are not relevant to the argument but helpful for recognizing the argument component type. For example, preceding discourse connectors like "*therefore*", "*consequently*", or "*thus*" can signal a subsequent claim. Discourse connectors like "*because*", "*since*" or "*furthermore*" could indicate a premise. We refer to these text units as *preceding tokens*. The third body paragraph illustrates a contra argument and argumentative attack relations:

> "*Admittedly, [cloning could be misused for military purposes]$_{Claim5}$. For example, [it could be used to manipulate human genes in order to create obedient soldiers with extraordinary abilities]$_{Premise9}$. However, because [moral and ethical values are internationally shared]$_{Premise10}$, [it is very unlikely that cloning will be misused for militant objectives]$_{Premise11}$.*"

The paragraph begins with a claim against the stance of the author which is supported by Premise$_9$ in the second sentence. The two premises in the third sentence defend the standpoint of the author. Premise$_{11}$ is an attack of Claim$_5$ and Premise$_{10}$ supports premise$_{11}$. The conclusion, i.e. the last paragraph restates the major claim and summarizes the main aspects of the essay:

> "*To sum up, although [permitting cloning might bear some risks like misuse for military purposes]$_{Claim6}$, I strongly believe that [**this technology is beneficial for humanity**]$_{MajorClaim2}$. It is likely that [this technology bears some important cures which will significantly improve life conditions]$_{Claim7}$.*"

This conclusion begins with an attacking claim followed by the restatement of the major claim. The last sentence includes another claim restating the most important points of the author's argumentation. Figure 4.1 illustrates the entire argumentation structure of this example essay. It shows the argumentation structure of all five paragraphs and illustrates the argumentation structure of each individual argument.

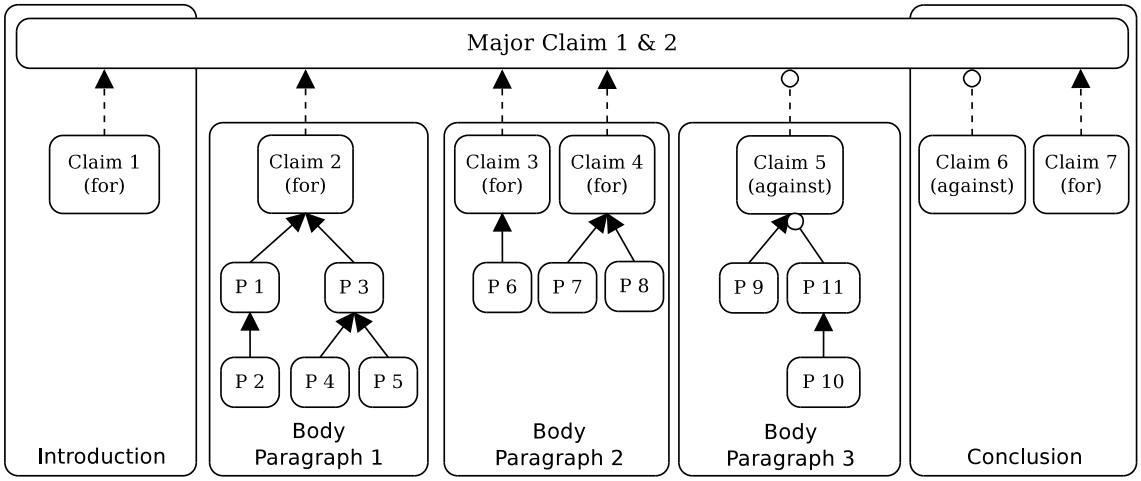


Figure 4.1: Argumentation structure of the example essay. Arrows indicate argumentative relations. Arrowheads denote argumentative support relations and circleheads attack relations. Dashed lines indicate relations encoded in the stance attributes of claims. "P" denotes premises.

In this section, we answered the first question stated in the introduction of this chapter by introducing an annotation scheme for modeling argumentation structures in persuasive essays. In the following, we seek to answer whether human annotators agree on its application to realistic texts.

4.2 Annotation Study

In order to answer the second question stated in the introduction of this chapter, we conduct an annotation study and evaluate the inter-annotator agreement between three human annotators. For creating a corpus of persuasive essays, we randomly selected English essays from *essayforum.com*. This forum is an active community that provides feedback for different texts such as research papers, essays or poetry. For example, students post their essays for receiving feedback about their writing skills while preparing for standardized language tests such as IELTS[3] or TOEFL[4]. We manually reviewed each essay and selected 402 essays that include a sufficiently detailed description of the writing prompt. The corpus comprises 7,116 sentences with 147,271 tokens.

4.2.1 Preliminary Study

We conducted a preliminary study on a corpus of 14 short text snippets (1–2 sentences) for elaborating the annotation guideline. Each text snippet is either collected from example essays or written by the author of this thesis. We asked five untrained annotators to classify each text as argumentative or non-argumentative. If a text is classified as argumentative, the annotators are also asked to identify the claim and the premise. In the first task, we obtained an observed agreement[5] of 58.6% and *multi-*π = 0.171 (Fleiss, 1971)[6]. Using majority vote we found that 7 of the 14 text snippets are argumentative. We identified the markables for evaluating the second task by using majority vote on the token level and considered the maximum overlapping argument component of each annotator as a hit. Following this procedure, we determined 32 text segments and obtained an observed agreement of 55.9% and *multi-*π = 0.291. These results indicate a low reliability. In subsequent discussions with the annotators, we discovered that the primary source of uncertainty was missing contextual information. Since the text snippets are provided without any information about the topic, the annotators found it difficult to decide if a text includes an argument or not. In addition, the annotators reported that the author's standpoint might facilitate the separation of argumentative from non-argumentative text units and to determine the components of arguments.

According to these findings, we define a top-down annotation process starting with the major claim and drill-down to the claims and premises. Therefore, the annotators are aware of the author's standpoint after identifying the major claim. In addition, we ask the annotators to read the entire essay for identifying the topic before starting with the actual annotation task. The resulting annotation process consist of the following steps:

[3] `https://www.ielts.org`

[4] `https://www.ets.org/toefl`

[5] We determine the observed agreement by averaging the percentage agreement of all annotator pairs as described in (Artstein and Poesio, 2008, p. 563).

[6] Although the coefficient was introduced by Fleiss as a generalization of Cohen's κ (Cohen, 1960), it is actually a generalization of Scott's π (Scott, 1955), since it assumes a cumulative distribution of annotations by all annotators (Artstein and Poesio, 2008). We follow the naming proposed by Artstein and Poesio and refer to the measure as *multi-*π.

1. *Topic and stance identification*: We ask the annotators to read the entire essay before starting with the annotation task.

2. *Annotation of argument components*: Annotators mark major claims, claims and premises. They annotate the boundaries of argument components and determine the stance attribute of claims.

3. *Linking premises with argumentative relations*: The annotators identify the structure of arguments by linking each premise to a claim or another premise with argumentative support or attack relations.

Based on this process, we elaborated an annotation guideline which comprises 31 pages. The guideline comprises a brief introduction to argumentation and an overview of the common structure of persuasive essays. Furthermore, it introduces the annotation process described above and includes various example arguments taken from real essays. For ensuring the reproducibility of the annotation study, the guideline is reprinted in Appendix D of this thesis.

4.2.2 Inter-Annotator Agreement

Three non-native speakers with excellent English proficiency participated in our annotation study. One of the three annotators elaborated the annotation guideline (expert annotator). The other two annotators trained the task by independently reading the annotation guideline.[7] For the actual annotation tasks, we used the brat rapid annotation tool (Stenetorp et al., 2012). It provides a graphical web interface for marking text units and linking them.

All three annotators independently annotated a subset of 80 randomly selected essays. The remaining 322 essays are annotated by the expert annotator. We evaluate the inter-annotator agreement of the argument component annotations using two different strategies: First, we evaluate if the annotators agree on the presence of argument components in sentences using *observed agreement*, *multi-π* (Fleiss, 1971) and Krippendorff's α (Krippendorff, 1980). We consider each sentence as a markable and evaluate the presence of each argument component type $t \in \{MajorClaim, Claim, Premise\}$ in a sentence individually. Accordingly, the number of markables for each argument component type t corresponds to the number of sentences N = 1,441, the number of annotations per markable equals with the number of annotators $n = 3$, and the number of categories is $k = 2$ ("t" or "*not* t"). Evaluating the agreement at the sentence level is an approximation of the actual agreement, since the boundaries of argument components can differ from sentence boundaries and a sentence can include several argument components.[8] Therefore, for the second evaluation strategy, we use Krippendorff's α_U (Krippendorff, 2004). In

[7] In our first annotation study, we conducted collaborative training sessions with the annotators (Stab and Gurevych, 2014b). In order to ensure the reproducibility of the results, we omitted these training sessions in our second study. Here we report the results of the second study.

[8] In our evaluation set of 80 essays the annotators identified in 4.3% of the sentences several argument components of different types. There is no argument component spanning more than one sentence. Thus, evaluating the reliability of argument components at the sentence level is a good approximation of the inter-annotator agreement.

contrast to common alpha coefficients, this coefficient allows to evaluate the agreement of unitizing tasks by comparing the boundaries of the annotation units. We use the squared difference δ^2 between any two annotators' sections as proposed by Krippendorff (2004, p. 9) and consider each essay as a single continuum at the token level. Accordingly, the length L of each continuum is the number of tokens of each essay. The number of annotators m that unitize the continuum is 3. We report the average α_U scores over 80 essays. For determining the inter-annotator agreement, we use *DKPro Agreement* whose implementations of inter-annotator agreement measures are well-tested with various examples from literature (Meyer et al., 2014).

	%	*multi-π*	α	α_U
major claim	.979	.877	.879	.810
claim	.889	.635	.635	.524
premise	.916	.833	.833	.824

Table 4.1: Inter-annotator agreement of argument component annotations.

The annotators agree best on major claims (Table 4.1). The observed agreement of 97.9% and the chance-corrected agreement of $\alpha = .879$ indicate that annotators can reliably annotate major claims in persuasive essays. The unitized alpha measure of $\alpha_U = .810$ shows that there are only few disagreements regarding the boundaries of major claims. The agreement scores of $\alpha = .833$ and $\alpha_U = .824$ also indicate good agreement for premises. We obtain the lowest agreement of $\alpha = .635$ for claims which shows that the identification of claims is a more complex task than the identification of major claims and premises. The joint unitized measure for all argument components is $\alpha_U = .767$. Therefore, we conclude that our guideline and annotation scheme guide annotators to substantial agreement.

For determining the agreement of the stance attribute, we treat each sentence containing a claim as "for" or "against". Consequently, the agreement of claims constitutes the upper bound for the agreement of the stance attribute. We obtain an agreement of 88.5% and $\alpha = .623$ for the stance attribute. This is only slightly lower than the agreement of claims. Therefore, human annotators can reliably differentiate between supporting and attacking claims.

We determined the markables for evaluating the agreement of argumentative relations by pairing all argument components in the same paragraph. For each paragraph with argument components $c_1, ..., c_n$, we consider each pair $p = (c_i, c_j)$ with $1 \leq i, j \leq n$ and $i \neq j$ as markable. Thus, the set of all markables corresponds to all argument component pairs that can be annotated according to our annotation guideline. The number of argument component pairs is $N = 4{,}922$, the number of ratings per markable is $n = 3$, and the number of categories $k = 2$.

	%	*multi-π*	α
support	.923	.708	.708
attack	.996	.737	.737

Table 4.2: Inter-annotator agreement of argumentative relation annotations determined on a subset of 80 persuasive essays.

Table 4.2 shows the inter-annotator agreement of argumentative relations. We

obtain for both argumentative support and attack relations chance-corrected agreement scores above .7 which allows tentative conclusions (Krippendorff, 2004). On average the annotators marked only 0.9% of the 4,922 pairs as argumentative attack relations and 18.4% as argumentative support relations. Although the agreement is usually much lower if a category is rare (Artstein and Poesio, 2008, p. 573), the annotators agree more on argumentative attack relations. This indicates that the identification of argumentative attack relations is a simpler task than identifying argumentative support relations.

4.2.3 Analysis of Human Disagreement

In order to analyze the disagreements between the annotators, we determine confusion probability matrices (CPM) (Cinková et al., 2012) for argument components and argumentative relations. Compared to traditional confusion matrices, a CPM also allows to analyze confusion if more than two annotators are involved in the annotation study. In particular, a CPM includes conditional probabilities that an annotator assigns a category in the column given that another annotator selected the category in the row.

	major claim	*claim*	*premise*	*non-arg*
major claim	.771	.077	.010	.142
claim	.036	.517	.307	.141
premise	.002	.131	.841	.026
non-arg	.059	.126	.054	.761

Table 4.3: Confusion probability matrix of argument component annotations ("non-arg" are sentences without argumentative content).

Table 4.3 shows the CPM for argument component annotations. It shows that the highest confusion is between claims and premises. The two main reasons for this confusion are context dependence and ambiguity of argumentation structures (Stab et al., 2014). In particular, if an argument includes a serial structure, the identification of the correct claim requires that the annotators are aware of the context which is difficult if the argumentation structure is more complex. We also observed that one annotator frequently did not split sentences including a claim. For instance, the annotator labeled the entire sentence as a claim though it includes another premise. These errors also explain the lower unitized alpha score for claims compared to the sentence-level agreement scores in Table 4.1. We also observed that concessions before claims were frequently not annotated as an attacking premise. For instance, annotators didn't split sentences similar to the following example:

> *Although [in some cases technology makes peoples' life more complicated]$_{premise}$, [the convenience of technology outweighs its drawbacks]$_{claim}$.*

The distinction between major claims and claims exhibits less confusion. This may be due to the fact that major claims are relatively easy to locate in essays, since they usually occur in introductions or conclusions whereas claims can occur anywhere in the essay.

	support	*attack*	*unlinked*
support	.605	.006	.389
attack	.107	.587	.307
unlinked	.086	.004	.910

Table 4.4: Confusion probability matrix of argumentative relation annotations ("unlinked" indicates argument component pairs that are not argumentative linked).

Table 4.4 shows the CPM of argumentative relations. There is little confusion between argumentative support and attack relations. The CPM also shows that the highest confusion is between argumentative relations (support and attack) and unlinked pairs. This can be attributed to the identification of the correct targets of premises. In particular, we observed that agreement on the targets decreases if a paragraph includes several claims or serial argument structures.

4.2.4 Creation of the Final Corpus

Since not all annotators labeled the entire corpus (cf. Section 4.2.2), we created a partial gold standard including only essays annotated by all annotators. We use this partial gold standard of 80 essays as our test data (20%) and the remaining 322 essay annotated by the expert annotator as our training data (80%).

We merge the annotations of all three annotators in the following way: first, we consolidate the argument components before the annotation of argumentative relations. Consequently, each annotator uses the same argument components when annotating argumentative relations. Second, we merge the argumentative relations to compile our final gold test set. Since the argument component types are strongly related, we didn't merge the annotations by majority voting (for instance the selection of premises depends on the selected claim(s) in a paragraph). Instead, we merged the annotations by discussing the disagreements with all annotators.

4.3 Corpus Statistics and Analysis

In this section, we analyze the final corpus in order to better understand the characteristics of argumentation structures in persuasive essays and to derive the requirements of end-to-end-argumentation structure parsers.

Table 4.5 shows an overview of the size of the corpus. The corpus comprises 147,271 tokens in 7,116 sentences. It includes 6,089 argument components of which 751 are major claims, 1,506 are claims and 3,832 are premises. This proportion between claims and premises is common in argumentative writing since writers usually provide several reasons for ensuring a robust standpoint (Mochales-Palau and Moens, 2011, p. 10). In addition, the less frequent attacking claims and argumentative attack relations confirm that students tend to support their own standpoint instead of considering opposing views (Wolfe and Britt, 2009).

The average essay in our corpus has the following characteristics: it includes five paragraphs, 18 sentences and 366 tokens (Table 4.5). It has approximately 15 argument components and two major claims. On average each essay includes four

		all	*avg. per essay*	*standard deviation*
size	sentences	7,116	18	4.2
	tokens	147,271	366	62.9
	paragraphs	1,833	5	0.6
arg. comp.	argument components	6,089	15	3.9
	major claims	751	2	0.5
	claims	1,506	4	1.2
	premises	3,832	10	3.4
	claims (for)	1,228	3	1.3
	claims (against)	278	1	0.8
rel.	support	3,613	9	3.3
	attack	219	1	0.9

Table 4.5: Size of the final corpus.

claims, three of which support the major claim and one attacks the major claim. It has nine supporting premises and one attacking premise.

Table 4.6 shows the proportion of argumentative and non-argumentative text units. The entire corpus includes 47,474 (32.2%) non-argumentative tokens and 99,797 (67.8%) argumentative tokens, i.e. tokens covered by an argument component. The corpus has 5,485 (77.1%) argumentative and 1,631 (22.9%) non-argumentative sentences. This proportion shows that the corpus can be used for separating argumentative from non-argumentative text units. It also shows that the separation of argumentative and non-argumentative text units is a crucial requirement for computational models that aim to recognize argumentation structures in persuasive essays.

	tokens		*sentences*	
argumentative	99,797	(67.8%)	5,485	(77.1%)
non-argumentative	47,474	(32.2%)	1,631	(22.9%)

Table 4.6: Proportion of argumentative and non-argumentative text units.

The corpus contains 1,506 claims. However, only 1,130 of these claims are supported (or attacked) by at least one premise. Thus, 25% (376) of the claims are unsupported and do not have incoming argumentative relations. Consequently, the corpus includes 1,130 arguments, i.e. claims that are supported (or attacked) by at least one premise. Table 4.7 shows an overview of the number of arguments per essay. Each essay includes at least two arguments. Most essays include three arguments.

#arguments	*essays*
2	144
3	202
4	47
5	7
6	1
7	1

Table 4.7: Frequency of arguments in persuasive essays.

In particular, 144 (35.8%) of the essays contain two arguments, 202 (50.2%) essays

have three arguments, and 47 (11.7%) essays contain four arguments. Only 2.3% of the essays contain more than 4 arguments. Our corpus includes considerably fewer attack relations compared to the microtext corpus from Peldszus and Stede (2016). Whereas 97 of the 112 microtexts (arguments) include attack relations (86.6%), our corpus exhibits only 140 of 1,130 arguments (12.4%) with an attack relation.

All essays together have 1,833 paragraphs. Each essays has 5 paragraphs on average. There are 44 of the 1,029 body paragraphs that include several arguments and 64 body paragraphs containing several claims. Considering all paragraphs, 113 include several claims and 421 of the paragraphs include several claims or major claims. Consequently, 23% of all paragraphs include either argument components without incoming or outgoing argumentative relations, i.e. unlinked argument components or several arguments, i.e. several claims with connected premises. Therefore, methods that link all components in a paragraph are only of limited use for identifying argumentation structures in our corpus and there is a need to group argument components before linking them.

Essays with attacking arguments are relatively rare though essay writing guidelines recommend to include a rebuttal paragraph. In total, 151 essays (37.6%) do not include attacking arguments, and 251 essays (62.4%) include at least one argument against the author's standpoint.

Table 4.8 shows the frequency of major claims per essay. Each essay in our corpus includes at least one major claim. There are 71 (17.7%) essays with one major claim, 313 (77.9%) essays have two major claims, and 18 (4.4%) have three major claims.

#major claims	*essays*
1	71
2	313
3	18

Table 4.8: Frequency of major claims in persuasive essays.

We annotated argument components at the clause-level. Accordingly, a sentence can contain several argument components. There are no sentences that include more than two argument components. Table 4.9 shows an overview of the patterns

	frequency
sentences with several components	583
sentences with different components	302
claim - claim	3
claim - major claim	70
claim - premise	112
major claim - claim	76
major claim - premise	1
premise - claim	51
premise - major claim	2
premise - premise	289

Table 4.9: Analysis of argument components in a single sentence.

found in a single sentence. The most common pattern are two premises in the same sentence. It occurs 289 times. The second most common pattern is a claim followed by a premise. It occurs 112 times and is more frequent than a premise followed by a claim. The third most common pattern is a major claim followed by a claim. In total, there are 583 sentences that include several argument components of which 302 sentences include two argument components of a different type, e.g. a claim followed by a premise. Therefore, 8.2% of all sentences need to be split in order to identify argument components. This shows that classifying sentences as a whole is not sufficient for identifying argument components.

Table 4.10 shows the average length of introductions, body paragraphs and conclusions. Conclusions are the shortest paragraphs. On average they have 48.6 tokens and 2.1 sentences. Introductions have 66.7 tokens on average and 3.3 sentences. Body paragraphs are on average 94.6 tokens long. They include 4.4 sentences on average.

	average tokens	*standard deviation*
introduction	66.7	22.4
body	94.6	32.1
conclusion	48.6	17.7

Table 4.10: Length of paragraphs.

For better understanding the common structure of persuasive essays, we analyzed the positions of argument component types. We segmented each paragraph in equally sized segments and counted the occurrence of each argument component type in a segment (we considered the maximum overlapping segment as 'hit'). Because of the different lengths of the paragraph types, we segmented introductions and conclusions in three segments and body paragraphs in five segments. Table 4.11 shows the results for all argument component types. Rows with the suffix "-all" summarize an entire paragraph. All other rows represent one segment of a particular paragraph type. The analysis shows that major claims occur solely in introductions

			claims			*premises*		
paragraph	*segment*	*major claim*	*claim (for)*	*claim (against)*	*claim (all)*	*premise (for)*	*premise (against)*	*premise (all)*
introduction	1	20	6	2	8	1	-	1
introduction	2	57	19	12	31	5	2	7
introduction	3	261	61	10	71	7	-	7
introduction-all		338	86	24	110	13	2	15
body	1	-	475	90	565	429	22	451
body	2	-	107	30	137	777	37	814
body	3	-	59	15	74	822	36	858
body	4	-	58	6	64	783	38	821
body	5	-	232	25	257	733	62	795
body-all		-	931	166	1,097	3,544	195	3,739
conclusion	1	171	48	54	102	10	4	14
conclusion	2	117	70	23	93	27	11	38
conclusion	3	125	93	11	104	19	7	26
conclusion-all		413	211	88	299	56	22	78

Table 4.11: Position of argument components.

and conclusions. Conclusions include more major claims than introductions. The major claim column also shows that most of the major claims are at the end of the introduction or at the beginning of the conclusion.

Claims occur most frequently in in the first segment of body paragraphs. Premises are more frequent in segments 2-5 of body paragraphs. This shows that stating the claim before premises is a common pattern of writing arguments in persuasive essays. Table 4.11 also shows that introductions include only few argument components, whereas body paragraphs exhibit the highest density of argumentative segments.

Our corpus includes 3,832 argumentative relations of which 3,613 are support relations (94.3%) and 219 are attack relations (5.7%). Each relation is a directed relation and has a target component and a source component. Table 4.12 shows that claims do not occur as sources. Accordingly, there are no claims that have *outgoing*

	frequency
relations	3,832
support	3,613
attack	219
claim as source	0
claim as target	3,108
premise as source	3,832
premise as target	724

Table 4.12: Overview of argumentative relations.

relations. However, claims are frequently the target of relations and exhibit several *incoming relations*. In total, 3,108 relations have a claim as their target (81.1%), whereas 724 relations have a premise as their target (18.9%).

Table 4.13 shows the depth of arguments. By definition, unsupported claims are no arguments. Therefore, we did not include the 376 unsupported claims ("arguments" with depth 0) in Table 4.13. Most of the arguments have a depth of one and are thus convergent arguments (cf. Section 2.1.3). The corpus includes 236 arguments with a $depth > 1$. Thus, 20.9% of all arguments include serial structures.

depth of arguments	*frequency*
1	894
2	210
3	24
4	2

Table 4.13: Depth of arguments.

Table 4.14 shows an overview of the direction of the relations. Most argumentative relations point backward, i.e. the target component appears before the source component of the argumentative relations. This shows that premises frequently follow the claims in persuasive essays. Amongst all relations, 2,540 (66.3%) point backward and 1,292 (33.3%) point forward.

Argumentative relations do not only hold between adjacent argument components but the relations can also cross several argument components. Table 4.15 shows an overview of the distance between linked argument components. The distance shows the number of argument components between the source and target component of an argumentative relation. It shows that 1,758 (45.9%) relations hold

relation type	*frequency*
forward	1,292
backward	2,540
forward support	1,260
backward support	2,353
forward attack	32
backward attack	187

Table 4.14: Direction of argumentative relations.

between adjacent argument components. Thus, the larger proportion of argumentative relations crosses at least one argument component. There are 888 (23.2%) relations crossing one argument component, 586 (15.3%) relations crossing two components, and 316 (8.2%) relations cross three components. The longest relation crosses ten argument components.

distance	*frequency*
0	1758
1	888
2	586
3	316
4	178
5	65
6	26
7	10
8	3
9	1
10	1

Table 4.15: Distance of argumentative relations (distance is the number of argument components between source and target component of a relation).

Table 4.16 shows that the average support relation crosses 125.59 characters, 24.72 tokens, 1.29 sentences and 1.13 argument components. It also shows that the distance of argumentative support relations does not differ considerably from argumentative attack relations.

	metric	*average distance*	*standard deviation*
support	character	125.59	140.48
attack	character	116.62	148.44
support	token	24.72	26.86
attack	token	22.87	28.09
support	sentence	1.29	1.38
attack	sentence	0.98	1.38
support	components	1.13	1.40
attack	components	0.99	1.36

Table 4.16: Average distance of argumentative relations.

4.4 Chapter Summary

In this chapter, we dealt with the research questions whether human annotators agree on the argumentation structure of persuasive essays, and if it is possible to create reliably annotated corpora for training end-to-end argumentation structure parsers. In order to answer these questions, we first defined an annotation scheme based on argument diagramming that consists of three argument components (major claim, claim and premise) and two argumentative relations (support and attack relations).

In the second part of this chapter, we reported the results of an annotation study with three human annotators. We showed that human annotators substantially agree on the argumentation structure in persuasive essays. In particular, we obtained an α_U of .767 for argument components and an average α-score of .723 for argumentative support and attack relations. The result of this annotation study is a novel annotated corpus that represents - to the date of writing this thesis - the largest resource annotated with fine-grained argumentation structures at the micro-level (cf. Table B.3 in Appendix B).

Finally, we thoroughly analyzed the properties of the annotated argumentation structures for deriving the requirements for end-to-end argumentation structure parsers which can be summarized as follows:

- *Presence of non-argumentative text units*: The proportion of non-argumentative text units in our corpus amounts to 22.9% non-argumentative sentence and 32.2% non-argumentative tokens respectively. Therefore, it is necessary to separate argumentative from non-argumentative text units.

- *Several components per sentence*: In our corpus, 8.2% of all sentences contain several argument components. Thus, the sentences need to be segmented into smaller units in order to recognize the boundaries of argument components.

- *Unlinked argument components and several arguments within a paragraph*: The corpus analysis showed that 23% of all paragraphs include at least one unlinked argument component or several arguments, i.e. several tree structures. Therefore, methods that link all argument components in a paragraph are not sufficient for identifying argumentation structures in our corpus.

- *Depth of arguments*: Arguments can exhibit serial structures, i.e. several premises arranged in a chain. Thus, knowing the type of an argument component is not sufficient for recognizing the structure of individual arguments.

- *Non-adjacent argumentative relations*: In our corpus 54.1% of all argumentative relations hold between non-adjacent argument components. For this reason, it is necessary to consider all possible pairs of argument components for recognizing argumentative relations.

In the next chapter, we introduce an end-to-end argumentation structure parser that considers all of these requirements.

Chapter 5

Parsing Argumentation Structures

In the previous chapter, we have introduced a novel corpus of persuasive essays annotated with fine-grained argumentation structures. This corpus lays the foundation for tackling the second major research question of this thesis (RQ2) and for establishing a novel parser that identifies argumentation structures. As discussed in Chapter 3, existing approaches are not yet capable to fulfill the requirements for identifying fine-grained argumentation structures in persuasive essays. They are either limited in granularity or omit important subtask. In this chapter, we approach these open issues by answering the following questions:

1. How can we automatically recognize argumentation structures in essays?
2. Which linguistic features are effective for specific subtasks?
3. Can we jointly model several subtasks to improve accuracy?

To answer these questions, we build upon our empirical findings from Chapter 4. The analysis of the annotations showed that the automatic identification of argumentation structures is subject to several requirements. These include the separation of argumentative from non-argumentative text units, the identification of component boundaries, the separation of several arguments, the identification of non-adjacent argumentative relations, and the identification of serial argument structures. According to these requirements, we define the architecture of our argumentation structure parser which comprises five consecutive analysis steps depicted in Figure 5.1. The *segmentation* model separates argumentative from non-

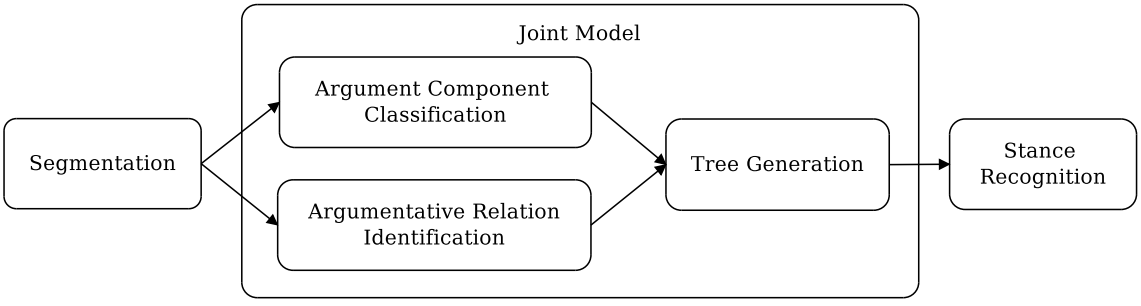


Figure 5.1: Architecture of the argumentation structure parser.

argumentative text units and identifies the boundaries of argument components. Since a single sentence may include several argument components, we consider this task as a sequence labeling task at the token-level. The next three models constitute a joint model for recognizing the argumentation structure. We train two base classifiers. The *argument component classification* model labels each argument component as major claim, claim or premise. The *argumentative relation identification* recognizes if two argument components are argumentatively linked. For finding non-adjacent argumentative relations and serial argumentation structures, we consider each pair of argument components within a paragraph. The *tree generation* globally optimizes the results of the two base classifiers in order to find a tree (or several ones) in each paragraph. Finally, the *stance recognition* model differentiates between argumentative support and attack relations.

The remainder of this chapter is structured as follows: we first describe the evaluation strategy and the evaluation measures used throughout this chapter. In Section 5.2, we introduce the methods used for preprocessing the texts. In Section 5.3, we introduce the segmentation model for recognizing argument components and their boundaries. In Section 5.4, we present the two base classifiers and our joint model for identifying argumentation structures. In Section 5.5, we introduce our stance recognition model. Finally, in Section 5.6, we describe the evaluation results.

5.1 Evaluation Strategy

For preventing overfitting, we strictly separate model selection from model assessment. In order to analyze different features and for selecting the best performing models of each task, we conduct *model selection* using 5-fold cross-validation on our train set (cf. Section 4.2.4). We conduct *model assessment* on our gold test set for comparing the best models to several baselines and human performance. Throughout this chapter, we determine the evaluation scores of cross-validation experiments by accumulating the confusion matrices of each fold into one confusion matrix since it is the less biased method (Forman and Scholz, 2010). We employ macro-averaging as described by Sokolova and Lapalme (2009) and report macro precision (P), macro recall (R) and macro F1 scores (F1) along with accuracy (Acc). We use a two-sided Wilcoxon signed-rank test with $p = 0.01$ for significance testing. Since most evaluation measures for comparing system outputs are not normally distributed (Søgaard, 2013), this non-parametric test is preferable to parametric tests, which make stronger assumptions about the underlying distribution of the random variable. We apply this test to all reported evaluation scores obtained for each of the 80 essays in our test set.

We evaluate the segmentation model independently from the other models. Accordingly, we report the results of the model selection and model assessment of the segmentation model in Section 5.3. For evaluating all other models, we rely on the argument component boundaries annotated in our corpus. For each of the remaining four models, we report the results of the model selection along with the feature analysis and error analysis in its own section. In Section 5.6, we report the results of the model assessment on the test set.

5.2 Preprocessing

We use several modules of the DKPro framework (Eckart de Castilho and Gurevych, 2014) for preprocessing the input texts. We identify tokens and sentence boundaries using the language tool segmenter[1] and identify the paragraphs by checking for line breaks. We lemmatize each token using the mate tools lemmatizer (Bohnet et al., 2013) and apply the Stanford part-of-speech tagger (Toutanova et al., 2003), constituent and dependency parsers (Klein and Manning, 2003), and sentiment analyzer (Socher et al., 2013). We use a PDTB-Parser (Lin et al., 2014) for recognizing general discourse relations, which achieves an overall F1 score of .468 for partial matching and .382 using full automation. We use the DKPro TC text classification framework (Daxenberger et al., 2014) for feature extraction and experimentation.

5.3 Segmentation of Argument Components

We consider the task of identifying argument components and their boundaries as a sequence labeling task at the token-level. We encode the argument components using an IOB-tagset (Ramshaw and Marcus, 1995) and label the first token of an argument component as "Arg-B", the remaining tokens of an argument component as "Arg-I" and tokens which are not covered by an argument component as "O". Table 5.1 shows the class distribution in our train and test set. It indicates that 67.8% of the tokens belong to argument components and 32.2% are non-argumentative.

	train		*test*	
Arg-B	4,823	(4.1%)	1,266	(4.3%)
Arg-I	75,053	(63.6%)	18,655	(63.6%)
O	38,071	(32.3%)	9,403	(32.1%)

Table 5.1: Class distribution of the train and test set for argument component segmentation.

We use the following two baselines. First, we employ a heuristic baseline based on sentence boundaries. Our corpus analysis has shown that the initial sentences in the introductions and the final sentences in the conclusions are frequently non-argumentative (cf. Table 4.11). Therefore, our heuristic baseline selects all sentences as argument components except the first two and the last sentence.[2] Second, we use a majority baseline that classifies all tokens of an essay as "Arg-I".

We employ a conditional random field (CRF) (Lafferty et al., 2001) implemented in *CRFSuite* (Okazaki, 2007) with averaged perceptron training method (Collins, 2002). The learner is particularly suited for sequence labeling tasks since it considers contextual information.

Features

We extract the following features for argument component segmentation (Table 5.2):

[1] `http://www.languagetool.org`
[2] We label each full stop as "O".

Structural Features: For each token, we define two binary features indicating if the token is present in the introduction or conclusion. Since both include less argumentatively relevant information, we expect that these features are effective for filtering non-argumentative content in persuasive essays. In addition, we use six numeric features indicating the absolute and relative position of a token in its sentence, its paragraph and the entire essay. For example, we determine the absolute position of token t_i in its covering sentence with tokens $t_1, t_2, ..., t_n$ as i and the relative position of token t_i as $\frac{i}{n}$. Two further binary features denote if the current token is the first or last word of a sentence. For capturing the position of the covering sentence of each token, we define four features representing the absolute and relative position of the sentence in its paragraph and the entire essay. For example, the absolute position of sentence s_i in a paragraph with sentences $s_1, s_2, ..., s_m$ is i and its relative position is $\frac{i}{m}$. In addition, we define eight binary features which indicate if the token directly follows or precedes any punctuation, a full stop, a comma or a semicolon, since it is more likely that an argument component begins or ends after or before a punctuation. Two additional binary features signal if the token is any punctuation or a full stop.

Syntactic Features: To capture the syntactic characteristic of each token, we extract features based on part-of-speech annotations and constituent parse trees. First, we extract for each token its POS-tag since it is less likely that a verb indicates the beginning or end of an argument component. For example, we extract the POS-tag `NN` for the token "*cloning*" in Figure 5.2 and `JJ` for the token "*medical*". Second, we define two features from the constituent parse tree of the covering sentence of the token. In particular, we measure the length of the path to the *lowest common ancestor* (LCA) of the current token and its preceding and following tokens. We normalize the length according to the total depth of the tree. We define the first feature considering the preceding token as $LCA_p(t_i) = \frac{|lcaPath(t_i,t_{i-1})|}{depth}$, where t_i is the current token, $|lcaPath(u,v)|$ is the length of the path from u to the LCA of u and v, and *depth* is the depth of the constituent parse tree. We define the second feature as $LCA_f(t_i) = \frac{|lcaPath(t_i,t_{i+1})|}{depth}$ considering the current token t_i and its following token t_{i+1}.[3] Figure 5.2 shows a constituency parse tree that illustrates these features. The

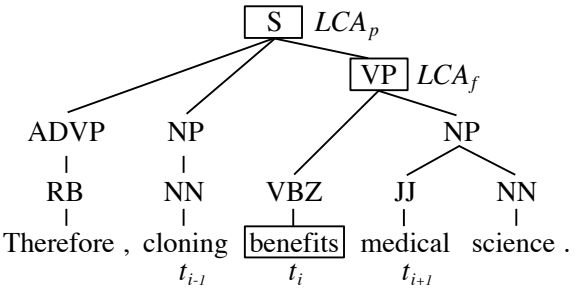


Figure 5.2: Parse tree illustrating the lowest common ancestor features.

LCA of the current token t_i and its preceding token t_{i-1} is the root node `S` of the

[3] Note that we set $LCA_p = -1$ and $LCA_f = -1$ if t_i is the first or last token of its covering sentence.

tree. The depth of the tree is 4 and the length of the path from t_i to the LCA is 3. Accordingly, the feature $LCA_p(t_i)$ is $\frac{3}{4}$. $LCA_f(t_i)$ is calculated accordingly and amounts to $\frac{2}{4}$. Additionally, we add the constituent types of both lowest common ancestors to our feature set. In our example above, we add the constituent type `S` of the LCA of the current token and its preceding token and the constituent type `VP` of the LCA of the current token and its following token. We assume that these features are effective for identifying the boundaries of argument components, since it is less likely that the boundaries of an argument component lays within a noun or verb phrase.

Lexico-syntactic Features (lexSyn): We adopt the lexico-syntactic features introduced by Soricut and Marcu (2003) that have been shown to be effective for segmenting elementary discourse units (Hernault et al., 2010). We use lexical head projection rules (Collins, 2003) implemented in the Stanford tool suite to lexicalize the syntactic constituent parse tree. For each word w, we determine its uppermost node with w as lexical head which as a right sibling. We combine its constituency type with w and denote this feature as N_w. In addition, we consider the parent node of N_w denoted as N_p and the right sibling of N_w denoted as N_r and also combine their lexical heads and constituent types analogous to the approach described by Soricut and Marcu (2003). Figure 5.3 illustrates these features. The upper-most

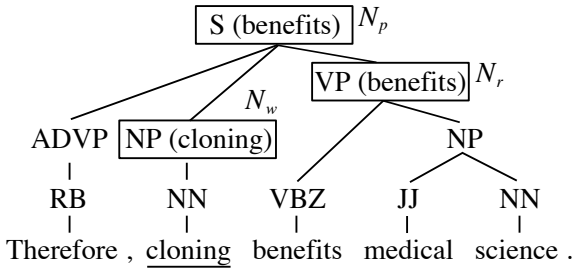


Figure 5.3: Parse tree illustrating lexico-syntactic features.

node of the word "*cloning*" with the same lexical head is the constituent node `NP` which is denoted as N_w in Figure 5.3. Its parent node is the root node of the parse tree with the lexical head "*benefits*" and its right sibling is the verb phrase with the same lexical head. Thus, we extract the features `S (benefits)`, `NP (cloning)` and `VP (benefits)` for the word "*cloning*".

Probability Features (prob): Argument components are frequently embedded in content-independent elements which indicate how argument components are related to each other (Madnani et al., 2012). For instance, argument components frequently have preceding discourse connectives like "*therefore*", "*thus*", "*because*" or phrases like "*to sum up*", "*another reason*" or "*in addition*" which signal the beginning of an argument component. Therefore, we define the conditional probability that the current token t_i is the beginning of an argument component ("Arg-B") given its preceding tokens as an additional feature. We maximize the probability for preceding tokens of a length up to $n = 3$:

$$\underset{n \in \{1,2,3\}}{\operatorname{argmax}} P(t_i = \textit{Arg-B} | t_{i-n}, ..., t_{i-1})$$

To estimate these probabilities, we divide the number of times the preceding tokens $t_{i-n}, ..., t_{i-1}$ with $1 \leq n \leq 3$ precede a token t_i labeled as "Arg-B" by the total number of occurrences of the preceding tokens in our train set. By doing so, we obtain the following probabilities: $P(t_i = \textit{Arg-B}|\textit{"since"}) = .704$, $P(t_i = \textit{Arg-B}|\textit{", since"}) = .696$, and $P(t_i = \textit{Arg-B}|\textit{"medicine, since"}) = 0$. Accordingly, the probability feature of the word "*it*" in the sentence "*Cloning is beneficial for medicine, since it shortens the healing process*" amounts to 70.4%.

Group	***Feature***	***Description***
Structural	Token position	Token present in introduction or conclusion*; token is first or last token in sentence; relative and absolute token position in document, paragraph and sentence
	Punctuation	Token precedes or follows any punctuation, full stop, comma and semicolon; token is any punctuation or full stop
	Position of covering sentence	Absolute and relative position of the token's covering sentence in the document and paragraph
Syntactic	Part-of-speech	The token's part-of-speech
	Lowest common ancestor (LCA)	Normalized length of the path to the LCA with the following and preceding tokens in the parse tree
	LCA types	The two constituent types of the LCA of the current token and its preceding and following tokens
LexSyn	Lexico-syntactic	Combination of lexical and syntactic features as described by Soricut and Marcu (2003)
Prob	Probability	Conditional probability of the current token being the beginning of a component given its preceding tokens

Table 5.2: Features used for segmenting argument components (*indicates genre-dependent features).

Evaluation

For identifying the best performing model and to investigate the effectiveness of each feature group, we conduct several 5-fold cross-validation experiments on our train set. Table 5.3 shows the results of using individual feature groups for the segmentation task. Lexico-syntactic features perform best regarding the macro F1 score,

	F1	*P*	*R*	*F1 Arg-B*	*F1 Arg-I*	*F1 O*
Baseline majority	.259	.212	.333	0	.778	0
Baseline heuristic	.628	.647	.610	.350	.869	.660
CRF only structural	†.748	†.757	†.740	†.542	†.906	†.789
CRF only syntactic	†.730	†.752	†.710	†.638	.868	.601
CRF only lexSyn	†.762	†.780	†.744	†.714	†.873	.620
CRF only probability	.605	†.698	.534	†.520	.806	.217
CRF w/o genre-dependent	†.847	†.851	†.844	†.778	†.925	†.835
CRF all features	†.849	†.853	†.846	†.777	†.927	†.842

Table 5.3: Segmentation model selection using 5-fold CV on the train set († = significant improvement over baseline heuristic).

and they perform particularly well for recognizing the beginning of argument components ("Arg-B"). The second best features are structural features. They yield the best F1 score for separating argumentative from non-argumentative text units ("O"). Syntactic features are useful for identifying the beginning of argument components. The probability feature yields the lowest macro F1 score. Nevertheless, we observe a significant decrease compared to the macro F1 score of the model with all features when evaluating the system without the probability feature ($p = .003$). Since persuasive essays exhibit a particular paragraph structure, which may not be present in other text genres (e.g. user-generated web discourse), we also evaluate the model without genre-dependent features (cf. Table 5.3). This yields a significant difference compared to macro F1 score of the model with all features ($p = 2.24 \times 10^{-54}$). By conducting feature ablation tests, we found that a combination of all features yields the best results on our train set. The model achieves an accuracy of .895 and a macro F1 score of .849 (Table 5.3).

Table 5.4 shows the results of the model assessment on our gold test set. The heuristic baseline achieves a macro F1 score of .642. It achieves an F1 score of .677 for non-argumentative tokens ("O") and .867 for argumentative tokens ("Arg-I"). Thus, the heuristic baseline effectively separates argumentative from non-argumentative text units. However, it achieves a low F1 score of .364 for identifying the beginning of argument components ("Arg-B"). Since it does not split sentences, it recognizes 145 fewer argument components than the number of gold standard components in the test set.

	F1	*P*	*R*	*F1 Arg-B*	*F1 Arg-I*	*F1 O*
Human upper bound	.886	.887	.885	.821	.941	.892
Baseline majority	.259	.212	.333	0	.778	0
Baseline heuristic	.642	.664	.621	.364	.867	.677
CRF all features	†.867	†.873	†.861	†.809	†.934	†.857

Table 5.4: Segmentation model assessment on the test set († = significant improvement over baseline heuristic).

The CRF model with all features significantly outperforms the macro F1 score of the heuristic baseline ($p = 7.85 \times 10^{-15}$). Compared to the heuristic baseline, it performs significantly better in identifying the beginning of argument components ($p = 1.65 \times 10^{-14}$). It also performs better for separating argumentative from non-argumentative tokens ($p = 4.06 \times 10^{-14}$). In addition, the number of identified argument components differs only slightly from the number of gold standard components in our test set. It identifies 1,272 argument components, whereas the number of gold standard components in our test set amounts to 1,266.

We determine the human upper bound by comparing pairs of annotators and averaging the three resulting scores (Table 5.4). The human upper bound yields a macro F1 score of .886 for identifying argument components. The macro F1 score of our model is only .019 less. Therefore, our model achieves 97.9% of human performance.

Error Analysis

For identifying the most frequent errors of our model, we manually investigated the predicted argument components. The most frequent errors are false positives of "Arg-I". The model classifies 1,548 out of 9,403 non-argumentative tokens ("O") as argumentative ("Arg-I"). The reason for these errors is threefold: First, the model frequently labels non-argumentative sentences in the conclusion of an essay as argumentative. These sentences are, for instance, non-argumentative recommendations for future actions or summarizations of the essay topic. Second, the model does not correctly recognize non-argumentative sentences in body paragraphs. It wrongly identifies argument components in 13 out of the 15 non-argumentative body paragraph sentences in our test set. The reason for these errors may be attributed to the high class imbalance in our train set. Third, the model tends to annotate lengthy non-argumentative preceding tokens as argumentative. For instance, it labels subordinate clauses preceding the actual argument component as argumentative in sentences similar to "*In addition to the reasons mentioned above, [actual 'Arg-B'] ...*" (underlined text units represent the annotations of our model).

The second most frequent cause of errors are misclassified beginnings of argument components. The model classifies 137 of the 1,266 beginning tokens as "Arg-I". The model, for instance, fails to identify the correct beginning in sentences like "*Hence, from this case we are capable of stating that [actual 'Arg-B'] ...* " or "*Apart from the reason I mentioned above, another equally important aspect is that [actual 'Arg-B'] ...*". These examples also explain the false negatives of non-argumentative tokens which are wrongly classified as "Arg-B".

5.4 Recognizing Argumentation Structures

The identification of argumentation structures involves the classification of argument component types and the identification of argumentative relations. Both share mutual information (Stab and Gurevych, 2014a, p. 54). For instance, if an argument component is classified as claim, it is less likely to exhibit outgoing relations and more likely to have incoming relations. On the other hand, an argument component with an outgoing relation and few incoming relations is more likely to be a premise. Therefore, it may be valuable to globally optimize both information. To this end, we first train two local base classifiers. One classifier recognizes the type of argument components, and another identifies argumentative relations between argument components. We introduce the argument component classification model in Section 5.4.1 and the argumentative relation identification model in Section 5.4.2. In order to find the optimal argumentation structure, we globally optimize the outcomes of both classifiers using integer linear programming. We introduce the details of this approach in Section 5.4.3. In each of these sections, we present the results of the model selection, parameter estimation and feature analysis. The results of the model assessment and the comparison with the human upper bound are presented at the end of this chapter in Section 5.6.

5.4.1 Classification of Argument Components

The classification of argument components constitutes the first base classifier of our joint model. We consider this task as multiclass classification and label each argument component as "major claim", "claim" or "premise". Table 5.5 shows the class distribution of the train and test set.

	train		*test*	
major claim	598	(12.4%)	153	(12.1%)
claim	1,202	(24.9%)	304	(24.0%)
premise	3,023	(62.7%)	809	(63.9%)

Table 5.5: Class distribution of the train and test set for argument component classification.

We employ the following two baselines for the classification of argument components: first, we use a majority baseline that classifies each argument component as "premise". Second, we use a heuristic baseline according to the following empirical results of the corpus analysis (cf. Section 4.3): major claims appear most frequent in the last segment of the introduction or in the first segment of the conclusion. Claims are frequently the first component in body paragraphs. Premises occur rarely in introductions and conclusions. Therefore, our heuristic baseline classifies the last argument component in introductions and the first argument component in conclusions as major claims and the remaining components in introductions and conclusions as claims. In body paragraphs, it classifies the first component as claim and the remaining components as premises. Given the results of our corpus analysis, we expect that this baseline will yield good results for classifying argument components.

For our model, we use a support vector machine (SVM) (Cortes and Vapnik, 1995) with polynomial kernel implemented in the Weka machine learning framework (Hall et al., 2009). The motivation for selecting this particular learner stems from the results of our previous work in which we found that support vector machines outperform other classifiers such as decision trees, random forests or naïve Bayes (Stab and Gurevych, 2014a, p. 51).

Features

We employ the following features for classifying argument components (Table 5.6):

Lexical Features: We use lemmatized unigram features extracted from each argument component and its preceding tokens. Thus, we ensure that discourse markers which indicate the argumentative types are included in the unigram features. We consider all unigrams as binary features. In addition, we add the 2k most frequent lemmatized dependency pairs, since they capture word dependencies between non-adjacent words. Each dependency pair consists of the lemmatized governor and the lemmatized dependent. For instance, we extract the following four dependency pairs from the example sentence in Figure 5.4: (therefore, cloning), (cloning, benefit), (benefit, science), and (science, medical). We treat each dependency pair as a binary feature and set it to true if a specific dependency pair appears in an argument

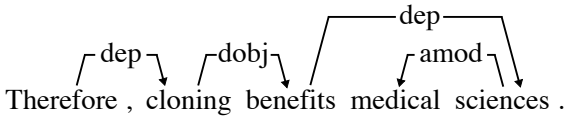


Figure 5.4: Dependency parse tree illustrating the dependency features.

component.

Structural Features: We add two binary features indicating if the current argument component is the first or last argument component in its paragraph. Two binary features represent if the argument component is present in the introduction or conclusion of the essay. In addition, we add the number of argument components in the paragraph, the number of tokens of the argument component and the number of tokens of its covering sentence to our feature set. Another numeric feature indicates the relative position of the argument component in its covering paragraph. For example, the relative position of an argument component c_i in a paragraph with argument component $c_1, c_2, ..., c_n$ is $\frac{i}{n}$. Four additional features encode the number of tokens before and after the argument component in its covering sentence, the ratio between tokens of the argument component and its covering sentence, and if the argument component boundaries match the boundaries of the covering sentence. In addition, we add two features representing the number of argument components preceding and following the current argument component in its covering paragraph.

Indicator Features: We selected four different types of indicators from 30 additional essays that are not included in our corpus. First, we selected 24 indicators that signal *forward reasoning*. These indicators signal that the argument component following the indicator is a result of preceding components. For instance, the list includes indicators like "*therefore*", "*thus*" or "*consequently*". Second, we select 33 indicators that signal *backward reasoning*. For example, indicators like "*for instance*", "*one of the main reasons*" or "*furthermore*" indicate that the argument component following the indicator refers to preceding argument components. Third, we select *rebuttal indicators* which indicate attacking components. However, we found only 10 of these indicators since attacking arguments and relations are rare in persuasive essays (Wolfe and Britt, 2009). Examples are: "*although*", "*admittedly*" or "*but*". Fourth, we found 48 *major claim indicators* indicating the presence of the author's standpoint. These include for instance: "*I think*", "*I totally agree*", or "*in my opinion*". For each argument component, we extract four binary features signaling if one indicator of the four categories is present in the component or its preceding tokens. The complete lists of all four categories are provided in Table C.4 in the Appendix of this thesis. The following example sentence illustrates these features.

> "*To sum up, although* [*permitting cloning might bear some risks like misuse for military purposes*]$_{Component1}$, *I strongly believe that* [*this technology is beneficial for humanity*]$_{Component2}$."

The sentence includes two argument components (both in brackets). The first component has two preceding indicators (underlined). The indicator "*To sum up*"

may signal a major claim and the indicator "*although*" is part of the rebuttal indicator list (cf. Table C.4). Therefore, we set two of the four indicator features of the first component to `true`. The second component has one preceding indicator which is part of the major claim list. Accordingly, we set one of the four features to `true`.

An additional binary feature indicates if the argument component or its preceding tokens include a reference to the first person. In particular, we check the presence of the five words "*I*", "*me*", "*my*", "*mine*" and "*myself*". We expect that this feature can help to identify major claims, since the sentence expressing the standpoint of the author frequently includes phrases like "*I think that*", "*in my opinion*", or "*from my perspective*" (cf. Section 4.1.4).

Group	***Feature***	***Description***
Lexical	Unigrams	Binary and lemmatized unigrams of the component and its preceding tokens
	Dependency tuples	Lemmatized dependency pairs (2k most frequent)
Structural	Token statistics	Number of tokens of component, covering paragraph and covering sentence; number of tokens preceding and following the component in its sentence; ratio of component and sentence tokens
	Component position	Component is first or last in paragraph; component present in introduction or conclusion*; Relative position in paragraph; number of preceding and following components in paragraph
Indicators	Type indicators	Forward, backward, thesis or rebuttal indicators present in the component or its preceding tokens
	First person indicators	"I", "me", "my", "mine", or "myself" present in component or its preceding tokens
Contextual	Type indicators in context	Forward, backward, thesis or rebuttal indicators present in the paragraph preceding or following the component
	Shared phrases*	Shared noun phrases or verb phrases with the introduction or conclusion (number and binary)
Syntactic	Subclauses	Number of subclauses in the covering sentence
	Depth of parse tree	Depth of the parse tree of the covering sentence
	Tense of main verb	Tense of the main verb of the component
	Modal verbs	Modal verbs present in the component
	POS distribution	POS distribution of tokens of the component
Probability	Type probability	Conditional probability of the component being a major claim, claim or premise given its preceding tokens
Discourse	Discourse triples	PDTB-discourse relations overlapping with the current component
Embedding	Combined word embeddings	Sum of the word vectors of each token of the component and its preceding tokens

Table 5.6: Features of the argument component classification model (*indicates genre-dependent features).

Contextual Features: Contextual information plays a major role for identifying the type of argument components (Mochales-Palau and Moens, 2007). For instance, it is likely that argument components close to a known claim serve as evidence and are presumably premises. Therefore, we define several contextual features based on

the indicators defined above. We add for each argument component eight binary features representing the presence of a forward, backward, rebuttal or thesis indicator preceding or following the argument component in its paragraph. We assume that these features are useful for modeling the local context of an argument component if indicators are present in a paragraph. For instance, if it is known that a forward indicator follows an argument component, it is less likely that the current component is a claim. In addition, we add six features representing content overlap with the introduction and the conclusion since claims frequently restate entities or phrases from the major claim or the general topic of the essay. We determine the number of noun phrases and verb phrases of the current component shared with the introduction and conclusion. Additionally, we add four binary features that indicate if the argument component shares any noun or verb phrase with the introduction or conclusion.

Syntactic Features: We adopt two features proposed by Mochales-Palau and Moens (2009) for capturing the syntactic characteristic of argument components: the number of sub-clauses in the covering sentence and the depth of the parse tree. Premises often refer to previous events, and claims are often in the present tense. Thus, we capture the tense of the main verb of an argument component as a binary feature (past or present). Since claims frequently exhibit modals like "*should*", "*can*" or "*could*" to express uncertainty, we use the POS-tags generated during preprocessing to identify modals and define a binary feature that indicates if an argument component contains a modal verb. Finally, we add the number of each POS-tag present in an argument component, hypothesizing that argument components of different types exhibit varying pos distributions. We illustrate these features by means of the following example sentence:

> "$\lceil$*Cloning can be beneficial for medical purposes*$\rfloor_{Component1}$*, since* $\lceil$*scientists cloned organs using stem cells.*$\rfloor_{Component2}$."

The sentence includes two sub-clauses and the depth of the parse tree is 12.[4] The first component is in present tense and the second in past tense. Accordingly, we set the past tense feature of the first component to `false` and the past tense feature of the second component to `true`. The first component in this example includes the modal verb "*can*", whereas the second component does not contain a modal verb. Thus, the modal feature of the first component is `true` and the modal feature of the second component is `false`. The pos distribution of the first component is `NN=1`, `MD=1`, `VB=1`, `JJ=2`, `IN=1`, `NNS=1` and the pos distribution of the second component is `NNS=3`, `VBN=1`, `VBG=1`, `VBP=1`.

Probability Features are the conditional probabilities of the current component being assigned the type $t \in \{MajorClaim, Claim, Premise\}$ given the sequence of tokens p that directly precede the component. To estimate $P(t|p)$, we divide the number of times the preceding tokens p appear before a component tagged as t by the total number of occurrences of p in our train set. We add the probability for major claim, claim and premise to our feature set for each argument compo-

[4] All feature values are determined using the Stanford parser (cf. Section 5.2).

nent. The following sentence includes two components, each with preceding tokens (underlined).

> "*Although, [cloning bears some risks]$_{Component1}$, I believe [it will be beneficial for humankind]$_{Component2}$.*"

The first component has the preceding tokens "*although ,*". We obtain the following probabilities: $P(MajorClaim|$"*although,*"$) = .017$, $P(Claim|$"*although,*"$) = .342$, and $P(Premise|$"*although,*"$) = .191$. For the second component and the preceding tokens "*, I believe*" we obtain $P(MajorClaim|$"*, I believe*"$) = .621$, $P(Claim|$"*, I believe*"$) = .103$, and $P(Premise|$"*, I believe*"$) = .053$.

Discourse Features: Cabrio et al. (2013) showed that discourse relations can be useful for identifying argument components. Therefore, we add features based on the output of the PDTB-style discourse parser (Lin et al., 2014) which achieves an overall F1 score of .468 for partial matching and .382 using full automation. In particular, we add a set of binary features that combine the type of the relation, if the current argument component overlaps with the first or second elementary discourse unit of the discourse relation, and if the discourse relation is implicit or explicit. For instance, we add the feature `Contrast_imp_Arg1` if the argument component overlaps with the first elementary discourse unit of an implicit contrast relation, or `Cause_exp_Arg2` if the argument component overlaps with the second elementary discourse unit of an explicit cause relation. Figure 5.5 shows a sentence with two argument components and an explicit causal discourse relation. In this example, we

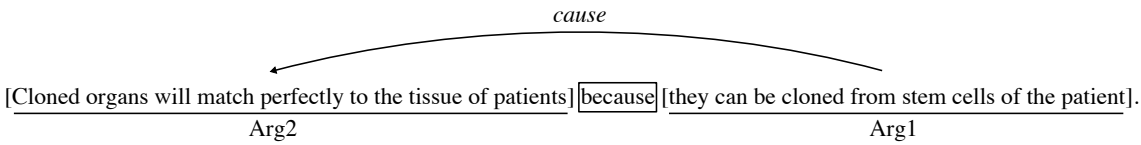


Figure 5.5: Causal discourse relation illustrating the discourse features.

extract a binary feature for each argument component. In particular, we extract the features `Cause_exp_Arg2` for the first component and `Cause_exp_Arg1` for the second component.

Embedding Features: We employ the word embeddings trained on a part of the Google News data set (Mikolov et al., 2013)[5] We sum the word vectors of each argument component and its preceding tokens resulting in a single vector with 300 dimensions. In contrast to common bag-of-word models, embedding features have a continuous feature space that improved the results in several NLP tasks (Socher et al., 2013).

Model Selection

Table 5.7 shows the results of the model selection on our training set. The heuristic baseline sets a challenging accuracy of .776 and performs well for identifying major

[5] Although it might be preferable to train the embeddings on student essays, due to the lack of training data it is common practice to use embeddings trained on texts from another domain.

claims and premises. It significantly outperforms the majority baseline ($p = 2.38 \times 10^{-54}$) and achieves an F1 score of .740 for major claims, an F1 score of .560 for claims and an F1 score of .870 for premises. For selecting the best model, we investigated each feature group individually and experimented with different feature combinations. Structural features perform well for classifying argument components.

	F1	*P*	*R*	*F1 MC*	*F1 Cl*	*F1 Pr*
Baseline majority	.257	.209	.333	0	0	.771
Baseline heuristic	.724	.724	.723	.740	.560	.870
SVM only lexical	.591	.603	.580	.591	.405	.772
SVM only structural	†.746	.726	†.767	†.803	.551	.870
SVM only contextual	.601	.603	.600	.656	.248	.836
SVM only indicators	.508	.596	.443	.415	.098	.799
SVM only syntactic	.387	.371	.405	.313	0	.783
SVM only probability	.561	.715	.462	.448	.002	.792
SVM only discourse	.521	.563	.484	.016	.538	.786
SVM only embeddings	.588	.620	.560	.560	.355	.815
SVM all w/o prob & emb	†.771	†.771	†.772	†.855	.596	.863
SVM w/o genre-dependent	†.742	†.745	.739	†.819	.560	.847
SVM all features	†.773	†.774	†.771	†.865	.592	.861

Table 5.7: Results of model selection and feature analysis of argument component classification model († = significant improvement over baseline heuristic; MC = major claim; Cl = claim; Pr = premise).

They are the most effective features for identifying major claims since they encode whether an argument component is present in the introduction or conclusion of an essay. Using only structural features significantly outperforms the macro F1 score of the heuristic baseline ($p = 4.04 \times 10^{-6}$) and yields .746 macro F1 score. We also found that none of the remaining features significantly outperforms the heuristic baseline when employed alone. Discourse features are the second best features for identifying claims (F1 score of .538). Therefore, we can confirm the assumption that general discourse relations are predictive for classifying argument components (Cabrio et al., 2013). Lexical features perform relatively well. They yield a macro F1 score of .591 and are informative for identifying major claims. Embedding features do not perform as well as common lexical features. They yield lower F1 scores for major claims and claims though they achieve better results for classifying premises. Contextual features are effective for identifying major claims since the shared noun and verb phrases implicitly capture the appearance of an argument component in the introduction and conclusion. They are also informative for identifying claims. Syntactic features are only effective for identifying major claims and premises. Using only indicator features yields a macro F1 score of .508. They are effective for identifying major claims but are less predictive for claims. The probability features yield an F1 score of .448 for identifying major claims and .799 for premises but achieve only a low F1 score for claims. One reason why the indicator and probability features are not effective for recognizing claims could be the ambiguous argumentative role of inner premises in serial argument structures. In particular, inner premises may exhibit claim indicators which signal their local derivation from another premises though they have the role of a premise in the global context of the entire argument.

We also evaluate our system without genre-dependent features. In particular, we removed the structural features that indicate if an argument component is present in the introduction and conclusion. We also removed the shared noun and verb phrases since they implicitly encode the same information. The resulting model (SVM all w/o genre-dependent) yields a macro F1 score of .742 and significantly outperforms the macro F1 score of the heuristic baseline ($p = .008$).

We achieve the best accuracy by omitting the probability and embedding features. However, we select the best performing system by means of the macro F1 measure. In contrast to accuracy, it assigns equal weights to classes and not to individual instances and thus the macro F1 measure is more appropriate for imbalanced data sets. Accordingly, we select the model with all features as our best performing system (Table 5.7).

Error Analysis

For analyzing frequent errors of the argument component classification model, we manually investigated the classification results. The confusion matrix (Table 5.8) reveals that the most frequent confusion is between claims and premises. The system classifies 410 actual premises as claim and 422 claims as premise. We found that

		predictions		
		major claim	claim	premise
actual	major claim	**514**	80	4
	claim	68	**712**	422
	premise	8	410	**2,605**

Table 5.8: Confusion matrix of the argument component classification determined with "SVM all features" on the training set and 5-fold cross-validation.

most of these errors are due to reasoning chains and co-occurring premises in the same sentence. For instance, the model tends to label premises that are part of a reasoning chain as claims, since those are frequently marked by claim-indicating discourse markers which signal their local derivation from another premise. The model, for instance, wrongly classifies the second premise in the following paragraph as claim because of the discourse marker "*therefore*".

> "*First of all, [students who study outside their countries can gain a lot of experience]$_{Claim}$. For example, [students might face many challenges in the host country]$_{Premise1}$. Therefore, [they will learn to better overcome obstacles during their semester abroad]$_{Premise2}$. [Overcoming these problems teaches the students how to be more mature and confident]$_{Premise3}$.*"

We also observed several cases in which the classifier wrongly identifies claims in sentences that include two premises. For example, the classifier wrongly classifies a claim if a sentence includes two premises connected with discourse connectors like "*because*" or "*since*".

The confusion matrix also shows that our model confuses major claims most frequently with claims. It wrongly classifies 68 claims as major claim and 80 actual major claims as claim. One cause of these errors is that the model learns predominant patterns which frequently appear in persuasive essays. It is a common pattern,

for instance, to start the conclusion with an attacking claim before restating the major claim (cf. example essay in Section 4.1.4). In these cases the model tends to classify a claim followed by a major claim if the first sentence of the closing paragraph includes two argument components. Similarly, the model tends to wrongly classify an argument component in the introduction or conclusion as major claim if it includes first person indicators.

5.4.2 Argumentative Relation Identification

The relation identification model constitutes the second base classifier of our joint model. It classifies ordered pairs of argument components as argumentatively "linked" or "unlinked". In this analysis step we consider both argumentative support and attack relations as "linked". The distinction between support and attack relations will be described in Section 5.5.

For each paragraph with argument components $c_1, ..., c_n$, we consider each $p = (c_i, c_j)$ with $i \neq j$ as argument component pair. An argument component pair is linked if our corpus contains an argumentative relation with c_i as source component and c_j as target component. The class distribution is skewed towards unlinked pairs (Table 5.9).

	train		*test*	
unlinked	14,227	(82.5%)	4,113	(83.5%)
linked	3,023	(17.5%)	809	(16.5%)

Table 5.9: Class distribution of linked and unlinked argument component pairs.

We use the following two baselines for evaluating the argumentative relation identification model: first, we define a heuristic baseline that exploits the structure of persuasive essays. It classifies argument component pairs as "linked" if both components appear in the same body paragraph and if the target component is the first argument component of the paragraph. This heuristic baseline correctly identifies convergent argument structures if the claim is the first argument component of a body paragraph. However, it does not recognize serial arguments and fails if several arguments appear in the same paragraph. Second, we employ a majority baseline that classifies each argument component pair as "unlinked".

As a learner for our model, we employ a support vector machine implemented in the Weka machine learning framework (Hall et al., 2009) since it performed best for argumentative relation identification in a comparison of several classifiers (Stab and Gurevych, 2014a).

Features

We use the following features for argumentative relation identification (Table 5.11):

Lexical Features: We extract binary and lemmatized unigrams of both components to capture the lexical information of an argument component pair. Since the preceding tokens of each argument component can include discourse markers, we add all preceding tokens of the source and target components. We limit the number

of the unigrams for both components to the 500 most frequent unigrams in our train set to prevent a too sparse feature set.

Syntactic Features: We extract binary POS features from the source and target components. For instance, for the sentence depicted in Figure 5.6, we set the POS features `RB`, `NN`, `VBZ`, and `JJ` to `true`. In addition, we extract the 500 most frequent production rules from the parse trees of all sentences in our train set and consider each production rule as a binary feature. Each production rule specifies a substitution of a constituent node in the parse tree. For example, we extract the

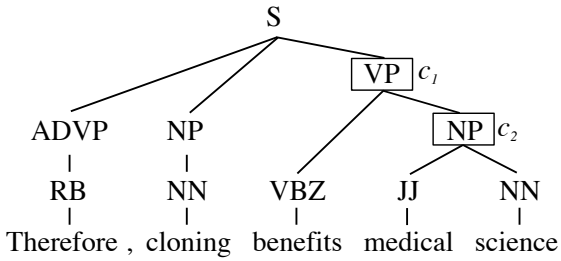


Figure 5.6: Parse tree for illustrating the syntactic features for relation identification.

production rule "`VP→VBZ,NP`" for the highlighted constituent node c_1 in Figure 5.6 and the production rule "`NP→JJ,NN`" for the constituent node c_2.

Structural Features: We add the number of tokens of the source and target components (`#tokens`), the number of argument components between the source and target components (`#componentsBetween`), a binary feature encoding if the target component appears before the source component (`targetBeforeSource`) and the number of argument components in the covering paragraph of the current pair (`#components`) to our feature set. We use a binary feature that encodes if both argument components appear in the same sentence (`sameSentence`) since it is likely that those exhibit an argumentative relation. Since claims appear frequently as the first or last component in a paragraph, we add four binary features encoding if the source or target component is the first or last argument component in the paragraph (`firstComponent` and `lastComponent`). Two additional binary features indicate if the current pair is present in the introduction or conclusion of the essay (`inIntroduction` and `inConclusion`). We illustrate the structural features by means of the following body paragraph taken from the example essay in Section 4.1.4. The source and target components of the current argument component pair are in boldface.

> "*First, [**cloning will be beneficial for many people who are in need of organ transplants**]$_{target}$. [Cloned organs will match perfectly to the blood group and tissue of patients] since [they can be raised from cloned stem cells of the patient]. In addition, [**it shortens the healing process**]$_{source}$. Usually, [it is very rare to find an appropriate organ donor] and [by using cloning in order to raise required organs the waiting time can be shortened tremendously].*"

The source component consists of five tokens and the target component consists of 14 tokens. The number of components between both components of the pair is 2 and the target component appears before the source component. The number of all components in the paragraph amounts to six and the source and target components are not in the same sentence. The target component is the first component in the

structural feature	*source*	*target*	*pair*
`#tokens`	`5`	`14`	`-`
`#componentsBetween`	`-`	`-`	`2`
`targetBeforeSource`	`-`	`-`	`true`
`#components`	`-`	`-`	`6`
`sameSentence`	`-`	`-`	`false`
`firstComponent`	`false`	`true`	`-`
`lastComponent`	`false`	`false`	`-`
`inIntroduction`	`-`	`-`	`false`
`inConclusion`	`-`	`-`	`false`

Table 5.10: Structural features extracted from an argument component pair.

paragraph and neither the source component nor the target component is the last component in the paragraph. Table 5.10 summarizes the structural features of this argument component pair.

Group	***Feature***	***Description***
Lexical	Unigrams	Binary lemmatized unigrams of the source and target components including preceding tokens (500 most frequent)
Syntactic	Part-of-speech	Binary POS features of the target and source components
	Production rules	Production rules extracted from the constituent parse tree (500 most frequent)
Structural	Token statistics	Number of tokens of source and target
	Component statistics	Number of components between source and target; number of components in covering paragraph
	Position features	Source and target present in same sentence; target present before source; source and target are first or last component in paragraph; pair present in introduction or conclusion*
Indicator	Indicator source/target	Indicator type present in source or target
	Indicators between	Indicator type present between source or target
	Indicators context	Indicator type follows or precedes source or target in the covering paragraph of the pair
Discourse	Discourse triples	PDTB-relations overlapping with the source component and target component
PMI	Pointwise mutual information	Ratio of tokens positively or negatively associated with incoming or outgoing relations; Presence of negatively or positively associated words with incoming or outgoing relations
ShNo	Shared noun phrases	Shared noun phrases between the target and source components (number and binary)

Table 5.11: Features used for argumentative relation identification (*indicates genre-dependent features).

Indicator Features: We employ the same set of indicators as used for argument

component classification (cf. Section 5.4.1). In particular, we assume that the indicators are helpful for modeling the direction of argumentative relations and the local context of the current component pair. We define eight binary features indicating if the current source or target component and their preceding tokens exhibit a forward, backward, thesis or rebuttal indicator as described in Section 5.4.1. In order to model the local context of the current pair, we define four binary features encoding the presence of an indicator between the target and the source component and 16 features indicating if an indicator type is present preceding or following the source and target components.

Discourse Features: Although the PDTB parser considers only adjacent discourse relations (Lin et al., 2014), we expect that the types of general discourse relations can be helpful for identifying argumentative relations. We extract for each source and target component the type of the general discourse relation, if the component is the first or second text unit of the discourse relation and if the relation is implicit or explicit analogous to the features described in Section 5.4.1. Note that we also experimented with features capturing PDTB relations between the target and source components. However, those were not effective for capturing argumentative relations.

PMI Features are based on the assumption that particular words indicate incoming or outgoing relations. For instance, words like "*therefore*", "*thus*", or "*hence*" can signal incoming relations, whereas tokens such as "*because*", "*since*", or "*furthermore*" might indicate outgoing relations. To capture this information, we use pointwise mutual information (PMI) which has been successfully used for determining word associations (Turney, 2002; Church and Hanks, 1990). Instead of using PMI between two words, we estimate the PMI between a lemmatized word w and the direction of a relation $d = \{incoming, outgoing\}$ as $PMI(w, d) = log\frac{p(w,d)}{p(w)\,p(d)}$. Here, $p(w, d)$ is the probability that word w occurs in an argument component with either incoming or outgoing relations. The ratio between $p(d, w)$ and $p(w)\,p(d)$ indicates the dependence between a word and the direction of a relation. We estimate $PMI(w, d)$ for each lemmatized word in our train set. We extract for both components the ratio of words that are positively associated with incoming or outgoing relations. In addition, we extract four binary features that indicate if any word has a positive or negative association with either incoming or outgoing relations. Table 5.12 shows an excerpt of the PMI values extracted from our train set. It shows, for instance, that

w	$PMI(w, outgoing)$	$PMI(w, incoming)$
therefore	-1.406	.348
hence	-.337	.147
thus	-.547	.209
consequently	-.461	.185
because	.869	-1.441
since	1.011	-3.299
furthermore	.247	-.165

Table 5.12: Excerpt of PMI values.

the words "*therefore*", "*hence*", "*thus*", and "*consequently*" are positively associated

with incoming relations, whereas words like "*because*", "*since*", and "*furthermore*" have a positive association with outgoing relations. Figure 5.7 illustrates the PMI values for all words of an example sentence. The ratio of words positively associated

Therefore	,	cloning	benefits	medical	science	.
.35 / -1.41		- / -	.285 / -.911	- / -	- / -1.03	

Figure 5.7: Illustration of the PMI values for all words of a sentence (the values below the words indicate $PMI(w, incoming)$ / $PMI(w, outgoing)$).

with incoming relations amounts to $\frac{2}{5}$ and the ratio of words that are positively associated with outgoing relations is 0.[6] Furthermore, the sentence contains words that are positively associated with outgoing relations and words that are negatively associated with incoming relations. Accordingly, we set two of the four binary features to `true` as described above.

Shared Noun Features (shNo): We expect that argument components are more likely connected if they share the same noun phrases. For instance, both premises in classical syllogisms share one subject with the claim (Govier, 2010, p. 199). Therefore, we define a binary feature indicating if the source and target components share any nouns. In addition, we add the number of nouns that both components have in common to our feature set.

Model Selection

Our heuristic baseline yields a macro F1 score of .660 and outperforms the majority baseline by .205. For analyzing the effectiveness of each feature group, we report the results of feature ablation tests in Table 5.13 since none of the feature groups yield remarkable results when used individually. Structural features are the most effective features for recognizing argumentative relations since removing them from the feature set yields the highest decrease of the macro F1 score. However, even without structural features, our system significantly outperforms the heuristic baseline by .055 macro F1 score ($p = 6.96 \times 10^{-5}$). The second and third most effective feature groups are indicator features and PMI features. Both improve the macro F1 score when combining them with other features by .014 and .013 respectively. This result shows that lexical clues are informative for identifying argumentative relations. Syntactic and discourse features are not as effective as our indicator and PMI features. However, both contribute to the identification of argumentative relations and yield a slight improvement when combining them with other features. Removing the shared noun features does not yield a significant difference in accuracy ($p = .626$) or macro F1 score ($p = .730$) compared to SVM with all features. However, we observe a decrease of .002 macro F1 score when removing the feature from our best performing model (Table 5.13). Therefore, we keep the shared noun feature in the feature set of the best performing model.

[6] Since the words "*cloning*" and "*medical*" do not appear in our corpus their PMI values are undefined. The word "*science*" appears only in components with outgoing relations. Therefore, $PMI(science, incoming)$ is also undefined.

	F1	*P*	*R*	*F1 unlinked*	*F1 linked*
Baseline majority	.455	.418	.500	.910	0
Baseline heuristic	.660	.657	.664	.885	.436
SVM all w/o lexical	†.736	††‡.762	†.711	††‡.917	†.547
SVM all w/o syntactic	†.729	††‡.764	†.697	††‡.917	†.526
SVM all w/o structural	††‡.715	††‡.740	‡.692	††‡.911	‡.511
SVM all w/o indicators	††‡.719	††‡.743	††‡.697	††‡.912	††‡.520
SVM all w/o discourse	†.732	†.755	†.709	†.915	†.540
SVM all w/o pmi	††‡.720	††‡.745	††‡.697	††‡.912	††‡.521
SVM all w/o shNo	†.733	†.756	†.712	†.915	†.545
SVM w/o genre-dependent	†.722	†.750	†.700	†.913	†.520
SVM all features	†.733	†.756	†.711	†.915	†.544

Table 5.13: Results of model selection and feature analysis of the argumentative relation identification († = significant improvement over baseline heuristic; ‡ = significant difference compared to SVM all features).

We achieve the best result by removing lexical features from our feature set. This model yields .860 accuracy and .736 macro F1 score. It also exhibits the highest score for linked and unlinked argument component pairs. Note that increasing the number of lemma unigrams improved the accuracy only when using lexical features without other features.[7] However, employing more lexical features did not improve the overall results when using the entire feature set.

Error Analysis

By analyzing the predicted relations, we observed that our model identifies to few linked argument component pairs. The model identifies only 2,319 linked pairs, though our train set includes 3,023 linked argument component pairs. We also observed that the model does not recognize any relation in 15.7% of all paragraphs that include at least one premise.[8] On the other hand, it wrongly identifies argumentative relations in only 3.7% of all paragraphs that do not have argumentative relations in our train set. This indicates that the model effectively identifies unlinked argument component pairs.

Moreover, we observed that the results of the relation identification model strongly deviate from the targeted tree structures. First, the model does not link 37.1% of the premises. It correctly identifies one outgoing relation for only 55.6% of the premises and several outgoing relations for 7.3% of the premises. Second, the model recognizes only 20.9% valid trees in our train set and thus fails to identify the correct tree of 79.1% of all arguments. Although these results differ considerably from our targeted argumentation structures, we show in the next section that they are valuable for identifying the structure of arguments.

5.4.3 Jointly Modeling Argumentation Structures

The classification of argument components and the identification of argumentative relations are closely related. Knowing the type of argument components is a strong

[7] We experimented with 500, 1000, 2000 and 4000 most frequent unigrams in our train set.
[8] Note that each premise should have one outgoing relation.

indicator for identifying argumentative relations and information about the argumentative structure facilitates the classification of argument components (Stab and Gurevych, 2014a, p. 54). For instance, if an argument component is classified as claim, it is less likely that it exhibits outgoing relations and more likely that it has incoming relations. Moreover, the predicted argumentative relations can be exploited to infer information about the argument component types. An argument component with several incoming and few outgoing relations is more likely to be a claim, whereas an argument component with few incoming relations is likely to be a premise. Therefore, it is reasonable to combine argument component types and argumentative relations for finding the tree structure which optimizes the results of the previous analysis steps.

We formalize this task as an integer linear programming (ILP) problem. Given a paragraph including n argument components, we define the following objective function

$$\underset{x}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_{ij} \tag{5.1}$$

with variables $x_{ij} \in \{0, 1\}$ representing an argumentative relation from argument component i to argument component j.[9] Each coefficient $w_{ij} \in \mathbb{R}$ is a weight for a relation and is determined by incorporating the results of previous analysis steps. In order to ensure that the resulting structure is a tree (or several ones), we define the following constraints:

$$\forall i : \sum_{j=1}^{n} x_{ij} \leq 1 \tag{5.2}$$

$$\forall i : x_{ii} = 0 \tag{5.3}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \leq n - 1 \tag{5.4}$$

Since each premise should exhibit exactly one outgoing argumentative relation and claims do not have outgoing relations, equation 5.2 ensures that each argument component i has one or zero outgoing relations. Equation 5.3 prevents that an argumentative relation has the same source and target component. Each paragraph with argumentative content needs to include a claim and there might be cases where several arguments and therefore several claims appear in the same paragraph. Thus, equation 5.4 ensures that a paragraph includes at least one root node without outgoing relation. For preventing cycles, we follow the approach described by Kübler et al. (2008, p. 92) and include a set of auxiliary variables $b_{ij} \in \{0, 1\}$ in our objective function (5.1). Here, $b_{ij} = 1$ if there is a directed path from argument component i to argument component j. For ensuring non-cyclic structures, we add the following constraints:

$$\forall i \forall j : x_{ij} - b_{ij} \leq 0 \tag{5.5}$$

$$\forall i \forall j \forall k : b_{ik} - b_{ij} - b_{jk} \leq -1 \tag{5.6}$$

$$\forall i : b_{ii} = 0 \tag{5.7}$$

[9] We use the lpsolve framework (`http://lpsolve.sourceforge.net`) and set each variable in our objective function to "*binary mode*" for ensuring the upper bound of 1.

The first of these constraints ties the variables x_{ij} to the auxiliary variables b_{ij}. It ensures that there is a path from i to j if there is a direct relation between the argument components i and j. The second constraint covers all paths of length greater than 1 in a transitive way. It states that if there is a path from argument component i to j ($b_{ij} = 1$) and another path from argument component j to k ($b_{jk} = 1$) then there is also a path from argument component i to k. It iteratively covers paths of length $l+1$ by having covered paths of length l. The third constraint restricts any cycle in the graph by preventing all directed paths starting and ending with the same argument component.

Having defined the ILP model, we consolidate the results of the argumentative relation identification and argument component classification models. We consider this task as determining the *weight matrix* $W \in \mathbb{R}^{n \times n}$ which includes the coefficients $w_{ij} \in W$ of our objective function. This matrix can be considered as an adjacency matrix. A greater weight for a particular relation denotes a higher likelihood that the relation is included in the optimal tree found by the ILP-solver.

We start by incorporating the results of the argumentative relation identification model whose result can be considered as an adjacency matrix $R \in \{0,1\}^{n \times n}$. For each pair of argument components (i, j) with $0 < i \leq n$ and $0 < j \leq n$, each $r_{ij} \in R$ is 1 if the relation identification model predicts an argumentative relation from argument component i (source) to argument component j (target). It is 0 if the model does not predict an argumentative relation. Accordingly, the first approach of determining the relation weights is using matrix R as weight matrix W. We refer to this approach as "*ILP-naïve*" and set $w_{ij}^{(ILP\text{-}naïve)} = r_{ij}$.

However, as mentioned above, the results of the argumentative relation identification model bear more valuable information which can be exploited for determining more elaborate weights. For incorporating this information into the weight matrix W, we first determine for each argument component i the *claim score* (cs_i) by means of the predicted relations represented in R:

$$cs_i = \frac{relin_i - relout_i + n - 1}{rel + n - 1} \tag{5.8}$$

where $relin_i = \sum_{k=1}^{n} r_{ki}[i \neq k]$ is the number of predicted incoming relations of argument component i, $relout_i = \sum_{l=1}^{n} r_{il}[i \neq l]$ is the number of predicted outgoing argumentative relations of argument component i and $rel = \sum_{k=1}^{n} \sum_{l=1}^{n} r_{kl}[k \neq l]$ is the total number of predicted relations in the given paragraph. Note that cs_i is bigger for argument components with many incoming argumentative relations and fewer outgoing argumentative relations. It becomes smaller for argument components which exhibit few incoming and more outgoing argumentative relations. By normalizing the score with the total number of predicted relations and argument components, it also accounts for context information in the current paragraph and prevents over optimistic scores. For instance, if all the predicted argumentative relations point to an argument component i which has no outgoing relations, cs_i is exactly 1. If there is an argument component j with no incoming and one outgoing argumentative relation in a paragraph with 4 argument components and 3 predicted relations in R, cs_j is $\frac{1}{3}$.

It is more likely that an argumentative relation links an argument component with a lower claim score to an argument component with a higher claim score.

Therefore, we determine the weight for each potential argumentative relation as:

$$cr_{ij} = cs_j - cs_i \tag{5.9}$$

By treating the claim score cs_j of the target component as a positive term, we assign a higher weight to argumentative relations pointing to argument components which are likely to be a claim. By subtracting the claim score cs_i of the source component i, we assign smaller weights to relations outgoing argument components with a higher claim score. Accordingly, we define our second model as $w_{ij}^{(ILP\text{-}relation)} = \frac{1}{2}r_{ij} + \frac{1}{2}cr_{ij}$ and refer to it as "*ILP-relation*" since it uses only information from our argumentative relation identification model.

Next, we incorporate the predicted types of argument components. Since it is more likely that an argumentative relation points to a claim, we assign a higher score to the weight w_{ij} if the target component j is predicted as claim. Accordingly, we define our third model as $w_{ij}^{(ILP\text{-}claim)} = c_{ij}$ where $c_{ij} = 1$ if argument component j is predicted as claim and $c_{ij} = 0$ if argument component j is not predicted as claim. Note that we also experimented with subtracting the type information of the source component, however, it didn't improve the results of the final model.

Finally, we combine the information of the relation identification and component classification model as

$$w_{ij} = \phi_r r_{ij} + \phi_{cr} cr_{ij} + \phi_c c_{ij} \tag{5.10}$$

Each ϕ represents a hyperparameter that indicates a weight for the particular information in our model. We tune the hyperparameter on our train set before assessing the best model on our gold test set. We experiment with the following proportions. The "*ILP-equal*" model uses $\phi_r = \phi_{cr} = \phi_c = \frac{1}{3}$ and thus uses an equal proportion of all scores. The "*ILP-same*" model uses the same amount of information from our two base classifiers which is realized by setting the coefficients to $\phi_r = \phi_{cr} = \frac{1}{4}$ and $\phi_c = \frac{1}{2}$. The "*ILP-balanced*" model balances the information of component types and relations by using $\phi_r = \frac{1}{2}$ and $\phi_{cr} = \phi_c = \frac{1}{4}$. Note that we incorporate our heuristic baseline in the weight calculation of all models if the base classifiers do neither recognize claims nor relations in a paragraph. In these cases, we set $w_{i1} = 1$ for $1 < i \leq n$ and the remaining $w_{ij} = 0$. For ensuring that an improvement can be attributed to the joint model and not to the incorporated baseline, we investigate the results of the base classifiers combined with the heuristic baseline and refer to it as "*Base+Heuristic*".[10]

Finally, we adapt the argumentative relations and argument component types according to the results of the ILP-solver. We revise each relation according to the determined x_{ij} scores of our objective function, set the types of each root node of the identified trees to claim and the types of all remaining components in the tree to premise.

[10] Base+Heuristic does not use the ILP-model but only incorporates the heuristic baseline in the results of our base classifiers if a paragraph does not exhibit predicted relations or claims.

Model Selection

We determine the best configuration of the joint model by conducting experiments on our train set.[11] Table 5.14 shows the results of the model selection. *Base+heuristic* shows the result of applying the baseline to all paragraphs in which the base classifiers identify neither claims nor argumentative relations. The heuristic baseline is triggered in 31 paragraphs, which results in 3.3% more trees identified compared to the base classifiers. As a consequence, Base+Heuristic identifies 3.3% more correct trees than the base classifier. However, the difference between Base+heuristic and the base classifiers is not statistically significant. For this reason, we can attribute any further improvements to the joint modeling approach.

	parameters			*component classification*				*relation identification*			*statistics*		
	ϕ_r	ϕ_{cr}	ϕ_c	F1	F1 MC	F1 Cl	F1 Pr	F1	F1 NoLi	F1 Li	Cl→Pr	Pr→Cl	Trees
Base heuristic	-	-	-	.724	.740	.560	.870	.660	.885	.436	-	-	100%
Base classifier	-	-	-	†.773	†.865	.592	.861	†.736	†.917	†.547	-	-	20.9%
Base+heuristic	-	-	-	†.776	†.865	.601	.861	†.739	†.917	†.555	0	31	24.2%
ILP-naïve	1	0	0	†.765	†.865	†.591	.761	†.732	††‡.918	†.530	206	1,144	100%
ILP-relation	$\frac{1}{2}$	$\frac{1}{2}$	0	††‡.809	†.865	††.677	‡.875	††.759	††‡.919	††.598	299	571	100%
ILP-claim	0	0	1	†.740	†.865	.549	.777	.666	.894	.434	229	818	100%
ILP-equal	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	††‡.822	†.865	††.699	††.903	††.751	†.913	††.590	294	280	100%
ILP-same	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	††‡.817	†.865	††.687	††.898	††.738	†.908	††.569	264	250	100%
ILP-balanced	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	††‡.823	†.865	††.701	††.904	††.752	†.913	††.591	297	283	100%

Table 5.14: Results of model selection of the ILP joint model († = significant improvement over baseline heuristic; ‡ = significant improvement over base classifier; Cl→Pr = number of claims converted to premises; Pr→Cl = number of premises converted to claims; Trees = percentage of correctly identified trees).

The ILP-naïve model does not yield an improvement over the base classifiers. Since a great many of argument components are not linked by the base classifier, the model converts 1,144 premises to claims. It identifies 78% more claims than present in our train set.[12] The ILP-relation model significantly outperforms the macro F1 score of both base classifiers ($p = 6.43 \times 10^{-12}$ for relations and $p = 7.23 \times 10^{-13}$ for components), but converts a large number of premises to claims. It improves the argument component classification by .036 macro F1 score and the argumentative relation identification by .023 macro F1 score. In particular, this model yields an improvement of .085 F1 score for the claim identification and yields the highest score for identifying argumentative relations. Thus, the claim score derived from the outcomes of the relation identification model effectively models the dependency between argumentative relations and component types. However, the model still converts a large number of premises to claims and identifies 22.46% more claims than present in our train set. The ILP-claim model uses only the outcomes of the argument component base classifier and improves neither component classification

[11] Note that our joint model does not consider major claims. For ensuring a realistic evaluation scenario, we rely on the major claim predictions of our argument component classification model. Consequently, the upper bound is .934 and .996 macro F1 score for argument components and argumentative relations respectively, since the false positives of major claims are excluded in the joint modeling approach and false negatives are included.

[12] Note that the model anyhow identifies 100% correct trees since we consider also claims without linked premises, i.e. unsupported claims, as valid trees.

nor relation identification. Without the predicted relations a great many of relation weights in W are 0 which induces a conversion of 818 premises to claims and thus the identification of 49.1% more claims. In addition, the model identifies 19.22% fewer argumentative relations than present in our train set because of the smaller number of identified premises. Furthermore, the relation identification exhibits no significant difference compared to the heuristic baseline.

Combining the results of the component classification model and the argumentative relation identification model yields a considerably more balanced proportion of argument component type conversions (lower part of Table 5.14). On average, the ILP-equal, ILP-same and ILP-balanced model identify only 1.16% fewer claims and consequently 0.73% more argumentative relations compared to the total number in our train set. Therefore, the combination of all three scores yields more accurate results than using the individual scores. Although all three models lead to a significant improvement of the component classification task over the base classifier, none of the three models significantly outperforms the relation base classifier though the ILP-balanced model improves the macro F1 score by .016.

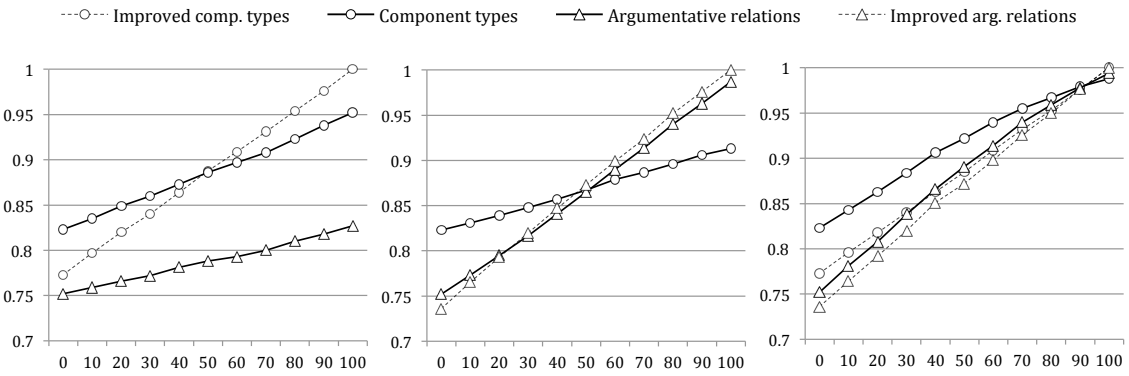
We identify the best performing model by averaging the macro F1 scores of the argument component classification and argumentative relation identification task. Accordingly, we select ILP-balanced as our best performing system. It achieves a macro F1 score of .823 and .752 for the classification of argument components and the identification of argumentative relations. In particular, it improves the F1 score for identifying claims by .109 ($p = 2.21 \times 10^{-23}$) and the F1 score for linked component pairs by .044 ($p = 2.22 \times 10^{-16}$) over the base classifiers. This indicates that jointly modeling argument components and argumentative relations considerably improves the performance and additionally leads to a consistent identification of the targeted tree structures.

Error Analysis and Influence of Base Classifiers

We observe that the model tends to identify more shallow trees compared to the structures in our train set. To be more specific, the model identifies only 34.7% of the serial arguments correctly. This can be attributed to the claim-centered weight calculation in our objective function. In particular, the predicted relations in the adjacency matrix R only include information about serial arguments if the argumentative model correctly classifies serial structures. The two other scores (c_{ij} and cr_{ij}) primarily assign higher weights to relations pointing to claims. We also observe that the model correctly separates arguments in only 57.5% paragraphs that include several arguments.

In order to further analyze the approach, we simulate the effects of improving the base classifiers analogous to the approach presented by Peldszus and Stede (2015). The dashed lines in Figure 5.8 show the performance of the artificially improved base classifiers and continuous lines indicate the performance of the relation identification and argument component classification after applying the joint modeling approach (ILP-balanced). The x-axes show the percentage of improved predictions and the y-axes the macro F1 score.

Figure 5.8a+b depict the effect of improving the argument component types and the argumentative relations respectively. It shows that a less accurate prediction of



(a) Improve Types (b) Improve Relations (c) Improve Both

Figure 5.8: Influence of improving the base classifiers: (a) illustrates the effect of improving the argument component types, (b) the improvement of argumentative relations. and (c) the improvement of both base classifiers.

argumentative relations has a more detrimental effect on the argument component types (Figure 5.8a). In contrast, a less accurate prediction of argument component types has less influence on the outcomes of the argumentative relation identification (Figure 5.8b). Therefore, it is reasonable to focus on the improvement of the relation identification model in future work for improving the overall result. Figure 5.8c shows the effect of improving both base classifiers. It illustrates that the joint model improves the argument type classification more effectively than the argumentative relation identification. It also shows that accurate predictions of the base classifiers improve the outcomes of both tasks. Thus, we conclude that the joint model successfully captures the dependency between component types and argumentative relations.

5.5 Stance Recognition

The stance recognition model differentiates between support and attack relations. We model this task as a binary classification of the source components of each argumentative relation since we expect that the source component exhibits more substantial information for differentiating between support and attack relations than the target component. We classify each premise and claim as either "support" or "attack". Note that the stance of each premise is encoded in the type of its outgoing argumentative relation whereas the stance of each claim is represented in its stance attribute.

Table 5.15 shows the class distribution of the train and test sets. Since authors tend to support their own view instead of providing opposing arguments (Wolfe and Britt, 2009), the class distribution is skewed towards support relations. In total, our corpus includes 90.7% support relations and only 9.3% attack relations.

For our experiments, we employ the following two baselines: first, we use a majority baseline that classifies each argument component as support. Second, we use a heuristic baseline that classifies each argument component in the second last paragraph as attack. This baseline is motivated by essay writing guidelines which

	train		*test*	
support	3,820	(90.4%)	1,021	(91.7%)
attack	405	(9.6%)	92	(8.3%)

Table 5.15: Distribution of supporting and attacking components.

recommend including opposing arguments in the paragraph before the conclusion.

Features

For the stance recognition task, we use the following features (Table 5.16):

Lexical Features: We use binary and lemmatized unigram features including preceding tokens. We assume that preceding tokens like "*but*", "*in contrast*", or "*on the other hand*" are valuable information for identifying attacking components.

Sentiment Features: For identifying the polarity of each argument component, we use the subjectivity lexicon provided by Wilson et al. (2005). We use the number of negative, positive and neutral words, and one binary feature that indicates the presence of negative words. In addition, we determine the overall polarity of each argument component by adding the number of positive words and subtracting the number of negative words. Complementary, we determine five sentiment scores of the covering sentence of each argument component using the Stanford sentiment analyzer (Socher et al., 2013). These include scores for very negative, negative, neutral, positive and very positive sentences. We illustrate the sentiment features by means of the following sentence:

"*Cloning could be misused for military purposes.*"

The sentence includes one neutral word ("*could*") and one negative word ("*misused*"). Therefore, we set the presence of negative words to `true`, the number of negative words to one, the number of neutral words to one, and the number of positive words to zero. Accordingly, the overall polarity of the argument component amounts to minus one. The five sentiment scores determined with the Stanford sentiment analyzer are very negative = .061, negative = .399, neutral = .387, positive = .125, and very positive = .0273.

Syntactic Features: We use the POS distribution as described in Section 5.4.1 and binary production rules (cf. Section 5.4.2) for capturing syntactic properties of each argument component.

Structural Features: We use the number of argument components in the covering paragraph, the number of tokens in the covering sentence, the ratio between the sentence and component tokens, and the number of preceding and following tokens of the component in its covering sentence. In addition, we add the relative position of the argument component in its covering paragraph to our feature set. In particular, we define the relative position of an argument c_i in a paragraph with argument component $c_1, c_2, ..., c_n$ as $\frac{i}{n}$, i.e. the position of the current component divided by

Group	***Feature***	***Description***
Lexical	Unigrams	Binary and lemmatized unigrams of the component and its preceding tokens
Sentiment	Subjectivity clues	Presence of negative words; number of negative, positive, and neutral words; number of positive words subtracted by number of negative words
	Sentiment scores	Five sentiment scores of covering sentence (Stanford sentiment analyzer)
Syntactic	POS distribution	POS distribution of tokens of the component
	Production rules	Production rules extracted from the constituent parse tree
Structural	Token statistics	Number of tokens of covering sentence; number of preceding and following tokens in covering sentence; ratio of component and sentence tokens
	Component statistics	Number of components in paragraph; number of preceding and following components in paragraph
	Component position	Relative position of the argument component in paragraph
Discourse	Discourse triples	PDTB-discourse relations overlapping with the current component
Embedding	Combined word embeddings	Sum of the word vectors of each token of the component and its preceding tokens

Table 5.16: Features used for stance recognition.

the total number of components in its paragraph. We also add the number of components preceding and following the current argument component to our feature set.

Discourse Features: We employ the same discourse features as for the component classification task (cf. Section 5.4.1). Since PDTB also includes contrast and concession relations, we expect that these features will be useful for identifying attacking components.

Embedding Features: We use the word embedding features described in Section 5.4.1. We sum the word vectors of each word in the argument component and its preceding tokens which results in a single vector with 300 dimensions.

Model Selection

Table 5.17 shows the results of the model selection. The heuristic baseline achieves a macro F1 score of .521 and outperforms the majority baseline by .046. For finding the best learner, we compared naïve Bayes (John and Langley, 1995), random forests (Breiman, 2001), multinomial logistic regression (le Cessie and van Houwelingen, 1992), C4.5 decision trees (Quinlan, 1993) and SVMs (Cortes and Vapnik, 1995). We found that the SVM considerably outperforms all other classifiers. Therefore, we report the results using the SVM only.

Using sentiment, structural and embedding features individually does not yield an improvement over the majority baseline. This indicates, on the one hand, that the polarity of words is not sufficient for capturing the argumentative stance of an argument component and, on the other hand, that there is no common alignment

	F1	*P*	*R*	*F1 Support*	*F1 Attack*
Baseline majority	.475	.452	.500	.950	0
Baseline heuristic	.521	.511	.530	.767	.173
SVM only lexical	†.663	†.677	†.650	†.941	†.383
SVM only syntactic	.649	†.725	†.587	†.950	†.283
SVM only discourse	.630	†.746	†.546	†.951	.169
SVM all w/o lexical	†.696	††‡.719	†.657	††‡.948	††‡.439
SVM all w/o syntactic	†.687	††‡.691	††‡.684	††‡.941	††‡.433
SVM all w/o sentiment	†.699	††‡.710	†.688	††‡.945	††‡.451
SVM all w/o structural	†.698	††‡.710	†.686	††‡.946	††‡.449
SVM all w/o discourse	†.675	††‡.685	††‡.666	††‡.941	††‡.408
SVM all w/o embeddings	†.692	††‡.703	††‡.682	††‡.944	††‡.439
SVM all features	†.702	†.714	†.690	†.946	†.456

Table 5.17: Results of model selection and feature analysis of the stance recognition model († = significant improvement over baseline heuristic; ‡ = significant difference compared to SVM all features).

of supporting and attacking argument components in persuasive essays. Lexical features yield a significant improvement over the macro F1 score of the heuristic baseline when used individually ($p = 8.02 \times 10^{-10}$). By ranking the lexical features using information gain, we found that the best ranked unigrams include words such as "*although*", "*however*", "*though*", "*admittedly*" and "*oppose*". Syntactic features significantly improve precision ($p = 1.81 \times 10^{-30}$), recall ($p = 1.95 \times 10^{-47}$), F1 Support ($p = 1.01 \times 10^{-27}$) and F1 Attack ($p = 1.53 \times 10^{-54}$) over the heuristic baseline, but do not yield a significant improvement over the macro F1 score of the heuristic baseline. Discourse features significantly outperform the heuristic baseline regarding precision ($p = 3.68 \times 10^{-28}$), recall ($p = 3.43 \times 10^{-49}$) and F1 Support ($p = 1.06 \times 10^{-32}$).

We identified the best performing model by conducting feature ablation tests. Since omitting any of the feature groups yields a lower macro F1 score compared to SVM all features, we select the model with all features as the best performing model. It achieves a macro F1 scores of .702. It achieves an F1 score of .946 and .456 for supporting and attacking components respectively.

Error Analysis

Table 5.18 shows the confusion matrix of the best performing model on the train set. It shows that the model wrongly classifies 232 actual attacking components as support (false negatives) and 178 actual supporting components as attack (false positives). We observed that the false positives are frequently due to wrongly in-

		predictions	
		support	attack
act.	support	**3,642**	178
	attack	**232**	173

Table 5.18: Confusion matrix of the stance recognition experiement determined with "SVM all features" on the train set using 5-fold cross-validation.

terpreted discourse markers. For instance, the model tends to classify argument

components as attack if they include phrases like "*although*", "*however*" or "*nevertheless*" though they are actually meant to support another argument component. In addition, the model wrongly classifies supporting components as attack if the essay exhibits a strong opposition against the topic. In particular, if the author strongly disagrees with the prompt of an essay, the arguments frequently include negative words. Therefore, considering the overall stance of the essay could be useful to improve the stance classification. For instance, it could be worthwhile to use stance recognition methods (cf. Section 3.3.1) and to normalize the local stance of an argument component with respect to the overall stance of the essays.

Among the false negatives, we observe several argument components which include references to the first person like "*I*", "*my*" or "*myself*". Most of these argument components are rebuttals of contra arguments meant to defend the standpoint of the author on the topic. For instance, these cases include phrases like "*In my opinion*", "*I believe*" or "*from my viewpoint*" without including attacking clue words. Since these phrases usually occur in supporting argument components, it is hard to correctly recognize them as attack. Additionally, some false negatives are due to missing lexical indicators. In these cases, the coupling of content words with subjectivity clues as proposed by Somasundaran and Wiebe (2009) could be useful. However, this approach requires a corpus with sufficient essays of the same topic (prompt) in order to achieve a good coverage of content words for a specific domain. Our corpus, however, includes various topics and only few essays about the same prompt. Therefore, such an approach is hardly applicable to our experiment.

5.6 Evaluation

As described in Section 5.1, we strictly separate model selection from model assessment to prevent overfitting and to ensure a correct assessment of the model performance on unseen data. So far, we have analyzed different features and parameters for each analysis step and selected the best performing models by conducting 5-fold cross-validation on our train set. In this section, we report the results of the model assessment on unseen data. To this end, we evaluate the models on our gold test set and compare the results to the human upper bound. In addition, we evaluate our models on the English microtext corpus from Peldszus and Stede (2016) which allows us to compare our model to previous work.

Table 5.19 shows the F1 scores of the classification, relation identification, and stance recognition tasks using our test set. The ILP joint model significantly outperforms the macro F1 score of the heuristic baselines for component classification ($p = 1.49 \times 10^{-4}$) and relation identification ($p = .003$). It also significantly outperforms the macro F1 score of the base classifier for component classification ($p = 7.45 \times 10^{-4}$). However, it does not yield a significant improvement over the macro F1 score of the base classifier for relation identification. The results show that the identification of claims and linked component pairs benefit most from the joint model. Compared to the base classifiers, the ILP joint model improves the F1 score of claims by .071 ($p = 1.84 \times 10^{-4}$) and the F1 score of linked component pairs by .077 ($p = 6.95 \times 10^{-5}$). The stance recognition model significantly outperforms the heuristic baseline by .118 macro F1 score ($p = .008$). It yields .947 F1 score for

	components				*relations*			*support/attack*			
	F1	F1 MC	F1 Cl	F1 Pr	F1	F1 NoLi	F1 Li	F1	F1 Sup	F1 Att	Avg F1
Human upper bound	.868	.926	.754	.924	.854	.954	.755	.844	.975	.703	.855
Baseline majority	.260	0	0	.780	.455	.910	0	.478	.957	0	.398
Baseline heuristic	.759	.759	.620	.899	.700	.901	.499	.562	.776	.201	.674
Base classifier	.794	†.891	.611	.879	.717	.917	.508	†.680	†.947	†.413	.730
ILP joint model	††‡.826	†.891	‡.682	‡.903	†.751	†.918	††‡.585	†.680	†.947	†.413	.752

Table 5.19: Model assessment on persuasive essays († = significant improvement over baseline heuristic; ‡ = significant improvement over base classifier).

supporting components and .413 for attacking components.

We determine the human upper bound for all three tasks by averaging the scores of all pairs of annotators on our test set. For the argument component classification task, we consider the maximum overlapping annotations since the boundaries of argument components can differ in the annotations of several annotators. The human upper bound yields .868 macro F1 score for component classification, .854 macro F1 score for relation identification, and .844 macro F1 score for stance recognition. The ILP joint model almost achieves human performance for classifying argument components. Its F1 score is only .042 lower compared to the human upper bound. Regarding relation identification and stance recognition, the F1 scores of our model are .103 and .164 less than human performance. Thus, our model achieves 95.2% human performance for component classification, 87.9% for relation identification, and 80.5% for stance recognition.

In order to verify the effectiveness of our approach, we also evaluated the ILP joint model on the English microtext corpus (cf. Section 3.1.3). For ensuring the comparability to previous results, we used the same repeated cross-validation setup as described by Peldszus and Stede (2015). Since the microtext corpus does not include major claims, we removed the major claim label from our component classification model. Furthermore, it was necessary to adapt several features of the base classifiers, since the microtext corpus does not include non-argumentative text units. Therefore, we did not consider preceding tokens for lexical, indicator and embedding features and removed the probability feature of the component classification model. Additionally, we removed all genre-dependent features of both base classifiers.

	components			*relations*			*support/attack*			
	F1	F1 Cl	F1 Pr	F1	F1 NoLi	F1 Li	F1	F1 Sup	F1 Att	Avg F1
Baseline heuristic	.712	.536	.888	.618	.856	.380	.542	.773	.293	.624
Base classifier	†.830	†.712	.937	†.650	†.841	†.446	†.745	†.855	†.628	.742
ILP joint model	††‡.857	††‡.770	†.943	††‡.683	††‡.881	††‡.486	†.745	†.855	†.628	.762
Best EG	.869	-	-	.693	-	.502	.710	-	-	.757
MP+p	.831	-	-	.720	-	.546	.514	-	-	.688

Table 5.20: Model assessment on microtext corpus from Peldszus and Stede (2015) († = significant improvement over baseline heuristic; ‡ = significant improvement over base classifier).

Table 5.20 shows the evaluation results of our model on the microtext corpus. Our ILP joint model significantly outperforms the macro F1 score of the heuristic baselines for component classification ($p = 2.10 \times 10^{-10}$) and relation identification

($p = 1.5 \times 10^{-8}$). The results also show that our model yields significantly better macro F1 scores compared to the two base classifiers ($p = .002$ for component classification and $p = 7.52 \times 10^{7}$ for relation identification). The stance recognition model achieves .745 macro F1 score on the microtext corpus. It significantly improves the macro F1 score of the heuristic baseline[13] by .203 ($p = 7.55 \times 10^{-10}$).

The last two rows in Table 5.20 show the results reported by Peldszus and Stede (2015) on the English microtext corpus. The *Best EG* model is their best model for component classification, and *MP+p* is their best model for relation identification. Compared to our ILP joint model, the Best EG model achieves better macro F1 scores for component classification and relation identification. However, since the outcomes of their systems are not available to us, we cannot determine if this difference is significant. The MP+p model achieves a better macro F1 score for relation identification, but yields lower results for component classification and stance recognition compared to our ILP joint model. This difference can be attributed to the additional information about the function[14] and role attribute[15] incorporated in their joint models (cf. Section 3.2.3). They showed that both have a beneficial effect on the component classification and relation identification in their corpus (Peldszus and Stede, 2015, Figure 3). However, the role attribute is a unique feature of their corpus and the arguments in their corpus exhibit an unusually high proportion of attack relations. In particular, 86.6% of their arguments include attack relations, whereas the proportion of arguments with attack relations in our corpus amounts to only 12.4%. Therefore, we assume that incorporating function and role attributes will not be beneficial using our corpus.

In summary, the evaluation results show that our model simultaneously improves the component classification and relation identification on both corpora. Therefore, we conclude that our approach successfully models the natural relationship between argument component types and argumentative relations and represents a robust model for identifying argumentation structures on two different discourse types, i.e. persuasive essays and microtexts.

5.7 Chapter Summary

In this chapter, we introduced a novel approach for parsing argumentation structures. We focused on the following three research questions:

First, we investigated in which way argumentation structures can be recognized automatically. To this end, we defined several consecutive analysis tasks for segmenting argument components, classifying their argumentative function, recognizing argumentative relations between them, and differentiating between support and attack relations. For separating argumentative from non-argumentative text units, we modeled the segmentation task at the token-level and encoded each argument component using an IOB-tagset. In this way, the model is also capable of recognizing

[13] The heuristic baseline for stance recognition on the microtext corpus classifies the fourth component as "attack" and all other components as "for".

[14] The function denotes whether an argument component has a supporting or attacking function.

[15] The role attribute denotes if an argument component presents a statement of the proponent or the opponent.

fine-grained boundaries of argument components and separating several argument components within the same sentence. For ensuring that the model recognizes serial structures and argumentative relations between non-adjacent components, we defined the relation identification model as pair classification task and considered all possible argument component pairs within a paragraph.

Second, we addressed the research question which linguistic features are effective for each subtask. To answer this question, we defined multiple feature groups for each analysis step and thoroughly investigated their effectiveness by experimenting with different feature combinations and conducting feature ablation tests. The results showed that structural features are most effective for filtering non-argumentative text units in persuasive essays, and lexico-syntactic and syntactic features are useful for recognizing the beginning of argument components. The most effective features for classifying the argument component types are structural features, contextual features and lexical features. We also found that automatically recognized PDTB-relations are effective for identifying claims. For the relation identification task we found that lexical clues (indicator features and PMI features) and structural features are the most predictive features in our feature set. Furthermore, we showed that the polarity of words and sentences is not sufficient for differentiating between argumentative support and attack relations.

Third, we investigated whether jointly modeling argument component types and argumentative relations improves the performance and the consistency of the recognized structures. To this end, we defined a joint model based on integer linear programming that globally optimizes the results of the argument component classification and argumentative relation identification. We defined several constraints that allow for recognizing several arguments, i.e. several trees, within the same paragraph. The evaluation results indicated a simultaneous improvement of component classification and relation identification on two different types of discourse (persuasive essays and microtexts). This illustrates that the model successfully models the natural relationship between argument component types and argumentative relations. Furthermore, we showed that the model correctly identifies valid tree structures in all paragraphs and thus recognizes more consistent results compared to the local base classifiers. Finally, the comparison with the human upper bound showed that the model achieves 97.6% human performance for segmenting argument components, 95.2% for component classification, 87.9% for relation identification, and 80.5% for stance recognition in persuasive essays.

Chapter 6

Quality Assessment of Natural Language Arguments

In this chapter, we consider the third research question (RQ3) of this thesis which focuses on the automatic assessment of argument quality. As discussed in Chapter 2, the quality of an argument is a product of many different criteria that often depend on highly subjective factors. We have also shown that the logical quality of arguments is independent of external influence factors. Due to this reason, logical quality criteria are most suitable for assessing the internal quality of arguments and for providing objective feedback about arguments respectively. However, empirical studies about the applicability of theoretically motivated quality criteria to real arguments are missing and it is largely unknown whether human annotators can reliably apply them to real arguments. Furthermore, we have shown in Chapter 3 that computational approaches for assessing the quality of arguments are restricted to coarse-grained scores which do not suffice to provide adequate feedback about the quality of arguments. In order to approach these gaps, we want to answer the following questions:

1. Do humans agree on the logical quality of arguments?
2. How can we automatically assess natural language arguments?
3. Which features are effective for assessing arguments?

For answering these questions, we first investigate the logical quality of arguments in persuasive essays. We conduct an annotation study with three annotators and apply the sufficiency criterion introduced in Chapter 2 to 1,029 arguments taken from persuasive essays. In addition, we compare feature-rich SVMs with neural networks for automatically detecting insufficiently supported arguments. We present the results of the annotation study, the description of the model, and the analysis of features in Section 6.1 of this chapter.

Furthermore, we introduce an approach for recognizing myside biases in persuasive essays, which is a tendency to ignore opposing standpoints. We model this task as a binary document classification and experiment with different linguistically motivated features. We introduce the approach and the feature analysis in Section 6.2 of this chapter.

6.1 Identifying Insufficiently Supported Arguments

In Chapter 2, we have already introduced the RAS-criteria for assessing the quality of arguments. In this section, we investigate the applicability of the sufficiency criterion to arguments in persuasive essays. An argument fulfills the sufficiency criterion if its premises provide enough evidence for accepting or rejecting the claim. The following example illustrates an argument that violates the sufficiency criterion:

> **Example 1**: "*It is an undeniable fact that tourism harms the natural habitats of the destination countries. As Australia's Great Barrier Reef has shown, the visitors cause immense destruction by breaking corals as souvenirs, throwing boat anchors or dropping fuel and other sorts of pollution.*"

The premise of this argument represents a particular example (second sentence) that supports a general claim in the first sentence. The argument is a generalization from one sample to the general case. However, a single sample is not enough to support the general case. Therefore, the argument does not comply with the sufficiency criterion.

> **Example 2**: "*Cloning will be beneficial for people who are in need of organ transplants. Cloned organs will match perfectly to the blood group and tissue of patients since they can be raised from cloned stem cells of the patient. In addition, it shortens the healing process.*"

Example 2 illustrates a sufficiently supported argument. It is reasonable to accept that transplantation patients will benefit from cloning if it enables a better match and an accelerated healing process.

Our primary motivation is to achieve a better understanding of the sufficiency criterion. To this end, we investigate whether human annotators can reliably differentiate between sufficiently and insufficiently supported arguments and if it is possible to create annotated data of high quality. In addition, we address the automatic recognition of insufficiently supported arguments. We investigate if, and how accurately, insufficiently supported arguments can be identified by computational techniques.

6.1.1 Corpus Creation

We conduct our annotation on the corpus of 402 argumentative essays that has been previously annotated with argumentation structures (cf. Chapter 4). By analyzing the annotated argumentation structures, we found that each body paragraph contains at least one argument and only 4.3% of all body paragraphs include several arguments, i.e. claims supported by premises. Therefore, we considered each body paragraph as an individual argument. This approximation has additional practical advantages for the identification of insufficiently supported arguments since it does not require the identification of argumentation structures in advance and prevents potential error propagation. Following this procedure, we extracted 1,029 arguments with an average length of 94.6 tokens and 4.5 sentences per argument.

Annotation Study

Three non-native annotators with excellent English proficiency independently annotated the arguments as sufficient or insufficient. We used 64 arguments from the corpus for elaborating the annotation guideline (cf. Appendix E) and 20 arguments for collaborative training sessions with the annotators. In these sessions, all three annotators collaboratively analyzed arguments for resolving disagreements and obtaining a common understanding of the annotation guideline. For the actual annotation task, we used the freely available *brat rapid annotation tool* (Stenetorp et al., 2012).

Inter-Annotator Agreement

All three annotators independently annotated an evaluation set of 433 arguments. We evaluated the agreement between the annotators using several inter-annotator agreement measures implemented in DKPro Agreement (Meyer et al., 2014). We used observed agreement and the two chance-corrected measures *multi-π* (Fleiss, 1971) and Krippendorff's α with nominal distance function (Krippendorff, 2004). The three annotators agreed on 91.07% of all 433 arguments (observed agreement). The chance-corrected agreement scores *multi-π* = .7672 and α = .7673 show substantial agreement between the annotators, which allows "*tentative conclusions*" Krippendorff (1980). Therefore, we conclude that human annotators can reliably identify insufficiently supported arguments in persuasive essays.

Analysis of Disagreements

In order to identify the reasons for the disagreements, we manually investigated all arguments on which the annotators disagreed. We found that a high proportion of these arguments include modal verbs in their claims. The following example illustrates such an argument:

> "*Watching television too often can have a negative effect on communication abilities. For instance, children often prefer watching cartoons or movies instead of meeting their classmates and thus they will not learn how to communicate properly.*"

Due to the modal verb "*can*" in the claim of this argument (first sentence), it is sufficient to provide one specific example as premise. However, annotators tend to overlook modal verbs and over-hastily annotate these arguments as insufficient.

The second most frequent cause of disagreements is due to the length of the arguments. In particular, one annotator labeled remarkably fewer arguments as insufficient. These arguments exhibit a comparatively large number of premises. This indicates that longer arguments are more likely to be perceived as sufficient than shorter arguments.

We also observed that several disagreements are due to hard cases. For instance, assessing the sufficiency of the following argument depends on the subjective interpretation of the undetermined quantification "*many*" in the claim:

> "*Living in big cities provides many opportunities. First of all, it will be easier to find a job in a city. Also there are various bars and clubs where you can meet new people. Above all there are shopping malls and cinemas for spending your free time.*"

We also found that annotators do not agree on arguments including terms such as "*some*", "*various*", or "*large number*". Thus, extending the annotation guideline with an explanation of how to handle modal verbs, the number of premises and undetermined qualifiers could further improve the agreement between the annotators in future annotation studies.

Creation of the Final Corpus

We merged the annotations of the three annotators on the evaluation set using majority voting. The remaining arguments have been annotated by the two annotators with the highest pairwise agreement on the evaluation set ($\alpha = .815$). Disagreements on the remaining arguments have been manually resolved in discussions among the two annotators. Table 6.1 shows an overview of the corpus.

Tokens	97,370
Sentences	4,593
Arguments	1,029
Sufficient	(66.2%) 681
Insufficient	(33.8%) 348

Table 6.1: Statistics of the corpus and class distribution of sufficiency annotations.

The class distribution is skewed towards sufficiently supported arguments. However, the proportion of 33.8% insufficiently supported arguments indicates that students frequently do not support their claims with sufficient evidence.

6.1.2 Approach

We consider the identification of insufficiently supported arguments as a binary classification task and label each body paragraph as "sufficient" or "insufficient". For preventing errors in model assessment due to a particular data splitting (Krstajic et al., 2014), we use a repeated 5-fold cross-validation setup and ensured that arguments from the same essay are not distributed over the train, test and development sets. We repeat the cross-validation 20 times which yields a total of 100 folds. As evaluation scores, we use accuracy and macro F1 score as well as the F1 score, precision and recall of the class "insufficient". Whereas the precision indicates the performance of the model to identify arguments that are really in need of revision, recall shows how well the model recognizes all insufficiently supported arguments in an essay. All evaluation scores are reported as average including the standard deviation over the 100 folds. In order to determine the macro F1 score, we employ macro-averaging as proposed by Sokolova and Lapalme (2009, p. 430). For model selection and hyperparameter tuning, we randomly sampled 10% of the training set of each fold as a development set. For significance testing, we use Wilcoxon signed-rank test on macro F1 scores with a significance level of $\alpha = .005$.

We employ several modules from the DKPro Framework (Eckart de Castilho and Gurevych, 2014) for preprocessing. We use the language tool segmenter[1] for tokenization and sentence splitting. We employ the Stanford parser (Klein and Manning, 2003) and named entity recognizer (Finkel et al., 2005) for constituency parsing and recognizing organizations, persons and locations. Note that only the model with manual features requires all preprocessing steps. All other models use only the tokenization of the language tool segmenter.

Baselines

For our experiments, we use the following two baselines: First, we employ a majority baseline that classifies each argument as sufficient. Second, we use a support vector machine with polynomial kernel implemented in the Weka framework (Hall et al., 2009). We employ the 4,000 most frequent lowercased words as binary features and refer to this model as SVM-bow.

Manually Created Features (SVM)

Our first system is based on manually created features. As a learner, we use the same support vector machine as for SVM-bow. For feature extraction and experimentation, we use the DKPro TC text classification framework (Daxenberger et al., 2014). We tried various features which have been used previously for assessing the quality or the persuasiveness of arguments. For instance, we experimented with argument structures (Stab and Gurevych, 2014a), transitional phrases (Persing and Ng, 2015), semantic roles (Das et al., 2014) and discourse relations (Lin et al., 2014). However, we found that only the following features are effective for recognizing insufficiently supported arguments:

Lexical: To capture lexical properties, we employ the 4,000 most frequent lowercased words as binary features analogous to SVM-bow.

Length: We use the number of tokens and the number of sentences as features since sufficiently supported arguments might exhibit more premises than insufficiently supported arguments (cf. Section 6.1.1).

Syntax: For capturing syntactic properties, we extract binary production rules from the constituent parse trees of each sentence of the argument as described in Section 5.4.2.

Named Entities (ner): We assume that arguments with insufficient support refer to particular entities in order to justify more general claims. Thus, we add the number of named entities appearing in the argument and the average occurrence of named entities per sentence to our feature set. We consider organizations, persons and locations separately. Thus the named entity features comprise six features in total, i.e. three binary and three numeric features.

[1] `http://www.languagetool.org`

Convolutional Neural Network (CNN)

Our second model is a convolutional neural network with max-over time pooling (Collobert et al., 2011). We use the implementation provided by Kim (2014). The selection of this model is motivated by the excellent performance that the model achieves in many different classification tasks like sentiment classification or question classification. We found in our experiments that instead of using several convolutional layers with different window sizes, a single convolutional layer with a window size of 2 and 250 feature maps performs best. For representing each word of an argument, we use the 300-dimensional word embeddings trained on the Google news data set by Mikolov et al. (2013). In order to adapt these vectors to the identification of insufficient arguments, we use non-static word vectors as proposed by Kim (2014). We train the network with stochastic gradient descent over shuffled mini-batches with the Adadelta update rule (Zeiler, 2012), a dropout rate of .5 and a mini-batch size of 50. For finding the best model, we apply early stopping on the development sets.

6.1.3 Evaluation

Table 6.2 shows the results of the model assessment on the test sets. The SVM-bow model with unigram features achieves .755 macro F1 score and .785 accuracy. It significantly outperforms the majority baseline by .357 macro F1 score which indicates that lexical features are informative for identifying insufficiently supported arguments. The support vector machine with manually created features significantly outperforms both the majority baseline and SVM-bow. It achieves .798 accuracy and .770 macro F1 score and thus outperforms the SVM-bow model by .015 macro F1 score. We obtain the best performance by using the CNN model. It significantly outperforms all other models with respect to all evaluation scores and achieves .827 macro F1 score and .843 accuracy.

	Accuracy	*Macro F1*	*F1 Insufficient*	*Precision*	*Recall*
Human Upper Bound*	.911±.022	.887±.026	.940±.015	.863±.058	.808±.109
Baseline Majority	.662±.033	.398±.012	0	0	0
Baseline SVM-bow	.785±.029	†.755±.034	.661±.051	.709±.067	.624±.067
SVM	.798±.028	††.770±.032	.681±.047	.731±.060	.641±.061
CNN	.843±.025	††.827±.027	.770±.039	.762±.054	.784±.068

Table 6.2: Results of model assessment on the test sets and comparison to human upper bound († significant improvement over baseline majority; ‡ significant improvement over baseline SVM-bow; *determined on a subset of 433 arguments).

The results also show that the SVM model with manually created features achieves a considerably lower recall compared to precision. Thus, the model is less suitable for exhaustively finding all insufficiently supported arguments. In contrast, the CNN model is more balanced with respect to precision and recall and considerably outperforms the recall of the SVM model. Therefore, the CNN model outperforms the SVM model in finding insufficiently supported arguments in persuasive essays and performs better for recognizing arguments that are really in need of revision.

We determine the human upper bound by averaging the evaluation scores of all three annotator pairs on the 433 independently annotated arguments (cf. Section 6.1.1). Human annotators achieve an accuracy of .911. The CNN model yields only .068 less accuracy compared to the human upper bound and thus achieves 92.5% of human performance.

Feature Analysis

Although the CNN model outperforms the support vector machine with manual features, we analyzed the features for gaining a better understanding of insufficiently supported arguments and to investigate which linguistic properties are informative for recognizing arguments with insufficient support. Table 6.3 shows the macro F1 scores of the support vector machine using individual features and the results of feature ablation tests on the development sets.

	Macro F1	*F1 Insuf.*	*F1 Suf.*
BS Majority	.396 ± .020	0	.793 ± .041
only lexical	.749 ± .048	.649 ± .070	.835 ± .040
only length	.397 ± .023	.002 ± .015	.792 ± .040
only syntax	.640 ± .063	.502 ± .101	.767 ± .047
only ner	.681 ± .059	.410 ± .114	.823 ± .039
all w/o lexical	.658 ± .059	.529 ± .093	.776 ± .045
all w/o length	.766 ± .049	.674 ± .068	.847 ± .040
all w/o syntax	.755 ± .049	.659 ± .070	.839 ± .040
all w/o ner	.760 ± .050	.666 ± .069	.843 ± .041
all features	.768 ± .049	.677 ± .068	.848 ± .040

Table 6.3: Results of the SVM using individual features and feature ablation tests on the development sets.

The results show that lexical features are most effective for identifying insufficiently supported arguments. They achieve the best macro F1 score of .749 when used individually. Removing lexical features from the feature set also yields the highest decrease in macro F1 score compared to the other features. The second best features are named entities. Using only named entity features yields a macro F1 score of .681. Thus, we can confirm our assumption that named entities are informative features for assessing the sufficiency of arguments. Syntactic features are also effective for recognizing insufficiently supported arguments. They yield .640 macro F1 score when used individually. The results also show that the length of an argument is only marginally informative for assessing the sufficiency of arguments. Using the length features individually yields only a slight improvement of the macro F1 score over the majority baseline. However, removing the length from the entire feature set causes a slight decrease of .002 in the macro F1 score compared to the system which uses all features. We achieve the best results by combining all features.

For gaining further insights into the characteristics of insufficiently supported arguments, we ranked all unigrams using information gain. The top ten words are "*example*", "*my*", "*was*", "*instance*", "*i*", "*for*", "*me*", "*friend*", "*he*", and "*did*". This might be an indication that *examples* (signaled by the terms "*example*" and "*instance*") or *personal experiences* (signaled by terms such as "*me*", "*my*", "*friend*" or "*he*") are not sufficient for developing strong arguments.

6.1.4 Error Analysis

In order to analyze the most frequent errors of the convolutional neural network, we manually investigated all arguments which are wrongly classified in each run of the repeated cross-validation experiment. In total, we found 41 sufficient arguments which are consistently misclassified as insufficient (false positives) and 28 insufficient arguments that are always misclassified as sufficient (false negatives).

Among the false positives, we observed that 35 arguments include examples as evidence which are signaled by terms like "*example*" or "*instance*". Thus, the model tends to overemphasize the presence of particular lexical indicators. Most of these arguments either refer to an example in addition to other premises which are already sufficient to support the claim or include an example for specifying another premise. However, we also found several false negatives which include examples as evidence. Thus, the model does not solely rely on these lexical clues.

Among the 28 false negatives, we found 8 arguments that refer to multi-word named entities which are not captured by word embeddings. Another 5 false negatives support the claim by means of personal experience and 3 ones cite numbers, i.e. previous studies or empirical evidence.

6.1.5 Discussion

Although the convolutional neural network achieves promising results, the sufficiency criterion is only one of three criteria that a logically good argument needs to fulfill. Thus, our approach is not yet able to separate logically good from illogical arguments. In our experiments, we also analyzed arguments with respect to the relevance and acceptability criterion. In particular, we conducted several annotation studies with varying guidelines and two annotators on a set of 100 arguments. For annotating the relevance criterion, we presented the annotated structure of each argument to the annotators and asked them to assess the relevance of each premise for the claim individually. In order to evaluate the acceptability criterion, we asked the annotators to mark each premise as acceptable if it represents undisputed common knowledge or a fact. However, we found that human annotators hardly agree on these criteria. We obtained low agreement scores of *multi-*$\pi = .435$ for the relevance criterion and *multi-*$\pi = .259$ for the acceptability criterion, which is not sufficient for creating a reliable corpus. In addition, we found that the violations of the relevance and acceptability criteria are less frequent than violations of the sufficiency criterion in argumentative essays. We observed that only 15% of the arguments include a premise that violates the relevance criterion and 14% of all premises violate the acceptability criterion.[2]

Although we didn't obtain adequate agreement scores for the acceptability and relevance criteria, we implemented a system that identifies insufficiently supported arguments in persuasive essays with a reasonable accuracy. Given that sufficiency flaws are the most frequent quality defects in argumentative essays, our system represents an important milestone for realizing argumentative writing support systems.

[2] We determined this proportion by averaging the ratio of acceptability and relevance violations among all arguments annotated by both annotators.

6.2 Myside Bias Recognition

A frequent mistake when writing argumentative texts is to consider only arguments that support the own standpoint and to ignore opposing arguments (Wolfe and Britt, 2009). This tendency to ignore opposing arguments is known as *myside bias* or *confirmation bias* (Stanovich et al., 2013). It has been shown that guiding students to include opposing arguments in their writings significantly improves the argumentation quality, the precision of claims and the elaboration of reasons (Wolfe and Britt, 2009). Therefore, it is likely that a system which automatically recognizes the absence of opposing arguments would effectively guide students to improve their argumentation. For the same reason, the writing standards of the Common Core Standards[3] require that students are able to clarify the relation between their own standpoint and opposing arguments on a controversial topic.

Existing approaches to argument analysis like e.g. the argumentation structure parser introduced in Chapter 5 or the approach introduced by Peldszus and Stede (2015) recognize the internal microstructure of arguments. Although these approaches can be exploited for identifying opposing arguments, they require several consecutive analysis steps like separating argumentative from non-argumentative text units, recognizing the boundaries of argument components and classifying individual arguments as support or attack. Certainly, an advantage of argumentation structure parsers is that they recognize the position of opposing arguments in text. However, knowing the position of opposing arguments is only relevant for positive feedback to the author and irrelevant for negative feedback, i.e. pointing out that opposing arguments are missing. Therefore, it is reasonable to model the recognition of missing opposing arguments as a document classification task.

The remainder of this section is structured as follows: first, we derive document-level annotations for myside bias recognition from our annotated essay corpus and evaluate their reliability by comparing three independent annotators. Second, we propose a feature set for detecting the absence of opposing arguments in persuasive essays and evaluate their effectiveness by conducting a systematic feature analysis. We show that our features significantly outperform a strong heuristic baseline and the argument structure parser introduced in Chapter 5. Third, we show that our model achieves 84% of human performance.

6.2.1 Corpus

For our experiments, we employ the argument structure annotated essay corpus introduced in Chapter 4. To the best of our knowledge, this corpus is the only available resource that exhibits an appropriate size and class distribution for detecting myside biases at the document-level (cf. Section 3.1). Each essay in this corpus is annotated with argumentation structures that allow to derive document-level annotations. The argumentation structures include arguments supporting or opposing the author's stance. Accordingly, we consider an essay as "negative" if it solely includes supporting arguments and as "positive" if it includes at least one opposing argument. Note that the manual identification of opposing arguments is a subtask of the argumentation structure identification. Both tasks require that the annotators

[3] `http://www.corestandards.org`

identify the author's stance, the individual arguments and if an argument supports or opposes the author's stance. Thus, deriving document-level annotations from argumentation structures is a valid approach since the decisions of the annotators in both tasks are equivalent.

To verify that the derived document-level annotations are reliable, we compare the annotations derived from the argumentation structure annotations of three independent annotators. In particular, we determine the inter-annotator agreement on a subset of 80 essays. The comparison shows an observed agreement of 90%. We obtain substantial chance-corrected agreement scores of *multi-*π = .786 (Fleiss, 1971) and Krippendorff's α = .787 (Krippendorff, 2004). Thus, we conclude that the derived annotations are reliable, since the agreement scores are only slightly below the "*good reliability threshold*" proposed by Krippendorff (2004).

Tokens	147,271
Sentences	7,116
Documents	402
Negative	(62.4%) 251
Positive	(37.6%) 151

Table 6.4: Size and class distribution of the corpus.

Table 6.4 shows the size of the corpus and the class distribution. The corpus includes 251 (62.4%) essays that do not include opposing arguments ("negative") and 151 (37.6%) essays that include at least one opposing argument ("positive").

6.2.2 Approach

We consider the recognition of myside biases as a binary document classification task. Due to the size of the corpus and to prevent errors in model assessment stemming from a particular data splitting (Krstajic et al., 2014), we employ a stratified and repeated 5-fold cross-validation setup. We report the average evaluation scores and the standard deviation over 100 folds resulting from 20 iterations. For model selection, we randomly sampled 10% of the training set of each run as a development set. We report accuracy, macro precision, macro recall and macro F1 scores as described by Sokolova and Lapalme (2009). We employ Wilcoxon signed-rank significance test on macro F1 scores (significance level = .005).

We preprocess the essays using several modules from DKPro (Eckart de Castilho and Gurevych, 2014). For tokenization, sentence and paragraph splitting, we employ the LanguageTool segmenter[4] and check for line breaks. We lemmatize each token using the Mate Tools lemmatizer (Bohnet et al., 2013) and apply the Stanford parser (Klein and Manning, 2003) for constituency and dependency parsing. Finally, we use a PDTB-Parser (Lin et al., 2014) and sentiment analyzer (Socher et al., 2013) for identifying discourse relations and sentence-level sentiment scores.

For model assessment, we use the following two baselines: first, we employ a majority baseline that classifies each essay as "negative" (not including opposing arguments). Second, we employ a rule-based heuristic baseline that classifies an

[4] `http://www.languagetool.org`

essay as "positive" if it includes the case-sensitive term "*Admittedly*" or the phrase "*argue that*" which often indicate the presence of opposing arguments.[5]

As a learner, we choose a support vector machine (Cortes and Vapnik, 1995) with a polynomial kernel implemented in Weka (Hall et al., 2009). For extracting features, we use the DKPro TC framework (Daxenberger et al., 2014).

Features

We experiment with the following features:

Unigrams (uni): In order to capture the lexical characteristics of an essay, we extract binary and case sensitive unigrams.

Dependency triples (dep): The binary dependency features are triples consisting of the lemmatized governor, the lemmatized dependent and the dependency type.

Production rules (pr): We employ binary production rules extracted from all sentences of the essay as illustrated in Section 5.4.2.

Adversative transitions (adv): We assume that opposing arguments are frequently signaled by lexical indicators. We use 47 adversative transitional phrases that are compiled as a learning resource[6] and grouped into the following categories: concession (18), conflict (12), dismissal (9), emphasis (5) and replacement (3). For each of the five categories, we add two binary features set to true if a phrase of the category is present in the surrounding paragraphs (introduction or conclusion) or in a body paragraph.[7] Note that we consider lowercase and uppercase versions of these features which results in a total of 20 binary features.

Sentiment Features (sent): We determine for each sentence five sentiment scores using the Stanford sentiment analyzer (Socher et al., 2013). These consist of scores for very negative, negative, neutral, positive and very positive sentiment. We average these five scores for all sentences for capturing the overall sentiment of the essay. In addition, we count the number of negative sentences and define a binary feature that indicates the presence of a negative sentence.

Discourse relations (dis): The binary discourse features include the type of the discourse relation and indicate if the relation is implicit or explicit. For instance, `Contrast_imp` indicates an implicit contrast relation. We only consider discourse relations of body paragraphs since the introduction frequently includes a description of the controversy which is not relevant to the author's argumentation and whose discourse relations could be misleading for the learner.

[5] We found these indicators by ranking n-grams with information gain.

[6] `http://www.msu.edu/~jdowell/135/transw.html`

[7] We identify paragraphs by checking for line breaks and consider the first paragraph as introduction, the last as conclusion and all remaining ones as body paragraphs.

6.2.3 Evaluation

In order to select a model and to analyze our features, we conduct feature ablation tests (lower part of Table 6.5) and evaluate our system with individual features. The

	Accuracy	*Macro F1*	*Precision*	*Recall*	*F1 Biased*	*F1 Unbiased*
Model assessment on test data						
Human Upper Bound*	.900±.010	.894±.011	.895±.011	.014±.892	.865±.016	.921±.008
Baseline Majority	.624±.001	.384±.000	.312±.001	.500±.000	.769±.001	0
Baseline Heuristic	.711±.039	.679±.050	.715±.059	.646±.045	.797±.027	.497±.083
SVM uni+pr+adv	.756±.044	†.734±.048	.747±.049	.721±.050	.814±.034	.639±.075
Model selection and feature analysis on development data						
SVM only uni	.734±.070	.709±.081	.727±.087	.693±.079	.801±.053	.591±.121
SVM only dep	.688±.045	.657±.097	.756±.153	.589±.056	.798±.027	.306±.148
SVM only pr	.601±.082	.569±.092	.570±.095	.568±.089	.685±.071	.444±.129
SVM only adv	.760±.074	.750±.076	.751±.077	.749±.077	.803±.065	.687±.097
SVM only sent	.625±.000	.385±.000	.312±.000	.500±.000	.769±.000	0
SVM only dis	.619±.054	.520±.111	.528±.164	.527±.064	.745±.040	.214±.171
SVM all w/o uni	.733±.060	‡.708±.087	.768±.110	.660±.073	.817±.038	.496±.151
SVM all w/o dep	.765±.077	.745±.087	.762±.092	.731±.086	.822±.059	.649±.125
SVM all w/o pr	.760±.062	.738±.082	.781±.097	.701±.074	.830±.042	.583±.138
SVM all w/o adv	.736±.066	‡.709±.090	.756±.108	.670±.079	.816±.044	.524±.151
SVM all w/o sent	.756±.064	.733±.085	.778±.100	.696±.076	.828±.043	.572±.146
SVM all w/o dis	.757±.061	.734±.082	.780±.097	.696±.075	.829±.041	.571±.143
SVM uni+pr+adv	.770±.071	.750±.081	.767±.086	.735±.080	.825±.055	.656±.118
SVM all features	.755±.064	.732±.086	.776±.102	.695±.077	.827±.044	.569±.149

Table 6.5: Evaluation Results: upper part shows the results of the best performing model on the test sets; lower part shows the model selection results on the development sets († = significant improvement over baseline heuristic; ‡ = significant difference compared to SVM all features; *determined on a subset of 80 essays).

adversative transitions and unigrams are the most informative features. Both show the best individual performance and a significant decrease if removed from the entire feature set. Thus, we conclude that lexical indicators are the most predictive features in our feature set. The sentiment features do not perform well. Individually they do not achieve better results than the majority baseline and the accuracy increases slightly when removing them from the entire feature set. By experimenting with various feature combinations, we found that combining unigrams, production rules and adversative transitions yields the best results (SVM uni+pr+adv).

For model assessment, we evaluate the best performing model on our test data (upper part of Table 6.5). The heuristic baseline considerably outperforms the majority baseline and achieves 71.1% accuracy. Our best system significantly outperforms this challenging baseline with respect to all evaluation measures. It achieves 75.6% accuracy and a macro F1 score of .734. We determine the human upper bound by comparing pairs of annotators and averaging the results of the 80 independently annotated essays (cf. Section 6.2.1). Compared to the upper bound, the accuracy of our model is 14.4% lower. Our model achieves 84% of human performance.

In order to evaluate if the document-level approach performs better than an argument structure parser, we compare the approach to the parser introduced in Chapter 5. To this end, we train both the argumentation structure parser and the model for myside bias recognition on the train set and evaluate their performance on

the test set of the essay corpus introduced in Chapter 4. If the argumentation structure parser recognizes an attacking claim or an argumentative attack relation, we consider the essay as positive. We consider all essays as negative in which the parser predicts only supporting claims and supporting argumentative relations. This yields a macro F1 score of .648 whereas the document-level approach considerably outperforms the parser with .710 macro F1 score. Hence, we can confirm that modeling the task as document classification outperforms argument parsing approaches.

6.2.4 Error Analysis

To analyze frequent errors of our system, we manually investigated essays that are misclassified in all 100 runs of the repeated cross-validation experiment on the development set. In total, 29 positive essays are consistently misclassified as negative. As reason for these errors, we found that the opposing arguments in these essays lack lexical indicators. In addition, we found 14 negative essays which are always misclassified as positive. Among these essays, we observed that the majority includes opposing indicators (e.g. "*but*") which are used in another sense (e.g. expansion). To sum up, the evaluation of both false negatives and false positives shows that the classifier tends to wrongly interpret lexical signals. Consequently, word-sense disambiguation for identifying senses or the integration of contextual information in the absence of lexical signals could further improve the results in future work.

6.3 Chapter Summary

In this chapter, we focused on the assessment of argumentation quality and addressed the following three research questions:

First, we investigated whether human annotators agree on the logical quality of arguments in persuasive essays. To this end, we conducted an annotation study with three annotators and applied the RAS-criteria proposed by Johnson and Blair (1977). The results indicate that human annotators substantially agree on the sufficiency criterion, whereas the relevance and acceptability criteria are too subjective to create reliable corpora. As a result of this annotation study, we introduced a corpus of arguments annotated with the sufficiency criterion.

Second, we addressed the question how natural language arguments can be assessed automatically. In order to approach this question, we experimented with several models for distinguishing between sufficiently supported arguments and insufficiently supported arguments. Our results showed that convolutional neural networks significantly outperform several strong baselines. In particular, the model achieves an accuracy of 84.3% for recognizing insufficiently supported arguments. The comparison to the human upper bound indicated that the model achieves 92.5% of human performance. Furthermore, we introduced the novel task of recognizing myside biases in persuasive essays. We modeled this task as a binary document classification and classified persuasive essays as positive if it includes at least one opposing argument and as negative if it does not contain any opposing arguments. We experimented with different features and proposed a novel feature set that achieves 77% accuracy and 84% of human performance.

Third, we investigated which features are effective for assessing natural language arguments. To this end, we thoroughly analyzed the features of both approaches by conducting feature ablation tests and experiments with different feature combination. In this way, we found that lexical features and named entity features are most effective for distinguishing sufficiently supported from insufficiently supported arguments. We also found indications that insufficiently supported arguments frequently include examples or personal experiences. Furthermore, we found that lexical features, adversative transitions and production rules are effective for identifying myside biases in persuasive essays.

In summary, we have presented two novel task for assessing argument quality in persuasive essays along with a reliable corpus, classification results and novel insights how argument quality can be modeled.

Chapter 7

Summary

The study of argumentation is a multifaceted research field that ranges from logical formalisms, over the study of monological and dialogical discourse, to research in cognitive science and rhetoric. In this thesis, we focused on argumentation in persuasive essays which was motivated by the need for argumentative writing support systems that provide feedback about written arguments. Within this work, we have proposed several approaches for automatically analyzing natural language arguments in persuasive essays. In particular, we introduced an end-to-end argumentation structure parser that recognizes fine-grained argument components and argumentative relations between them. Furthermore, we presented an approach for recognizing insufficiently supported arguments and an approach for recognizing myside biases. In the following, we summarize the main contributions and findings of this work.

First, we addressed the research questions whether theoretical argumentation models are applicable to persuasive essays and if it is possible to create reliable corpora for training argumentation structure parsers (RQ1). For answering these questions, we analyzed existing theoretical models for formalizing arguments. Within this analysis, our main focus was on monological models which allow for modeling the fine-grained microstructure of arguments. By comparing different monological models, we found that argument diagramming is a good foundation for modeling argumentation structures in text. Compared to other monological models, it explicitly models the targets of argument components by means of argumentative relations. Because of this feature, argument diagramming allows for modeling complex arguments including serial structures or chains of attacking argument components. Furthermore, the modeling of argumentative relations allows to separate several arguments in the same text. Therefore, we built upon argument diagramming and derived an annotation scheme that consists of three argument component types (major claim, claim and premise) and two argumentative relations (support and attack relations). We showed that our annotation scheme can be successfully applied to persuasive essays with substantial agreement. In particular, we obtained a unitized alpha score of .767 for argument components and an average alpha score of .723 for argumentative support and attack relations among three annotators. As a result of this study, we introduced a corpus of 90 persuasive essays reliably annotated with fine-grained argumentation structures in 2014 and published the second significantly extended version in 2016 which consists of 402 persuasive essays. The impact

of these resources is best illustrated by the following works: For instance, these corpora have been recently used in cross-domain experiments for identifying arguments (Al-Khatib et al., 2016), training argument parsers (Persing and Ng, 2016), studying topic-independent linguistic indicators for argumentation (Nguyen and Litman, 2016), studying the impact of persuasive argumentation in political debates (Cano-Basave and He, 2016), recognizing the strength of argumentation (Persing and Ng, 2015), and training methods for context-independent claim detection (Lippi and Torroni, 2015). Within this thesis, the annotated corpora laid the foundation for approaching the second major research question.

Second, we have addressed the automatic identification of argumentation structures (RQ2) and proposed an end-to-end argumentation structure parser. Based on a thorough analysis of the annotations in our corpus, we defined several consecutive analysis steps for segmenting argument components, classifying their argumentative function, recognizing argumentative relations between them, and distinguishing between argumentative support and attack relations. As a result of the corpus analysis, we have shown that a considerable part of persuasive essays is non-argumentative and a single sentence can contain several argument components. This is why we modeled the segmentation task at the token-level which allows for recognizing the fine-grained boundaries of argument components. In order to ensure that the model identifies serial argument structures and long distance relations, we considered the identification of argumentative relations as a pair classification task. For each task, we proposed multiple feature groups and analyzed their effectiveness. To this end, we conducted comprehensive feature ablation tests and experimented with various feature combinations. In the course of this analysis, we showed that syntactic and lexico-syntactic features are useful for recognizing the boundaries of argument components. Moreover, we found that structural features benefit the identification of non-argumentative text units in persuasive essays. With respect to the classification of argument components, we showed that structural, contextual as well as lexical features are effective. Our analysis also indicated that claims can be identified by leveraging automatically predicted discourse relations and that modeling the direction of argumentative relations by means of lexical indicators is effective for finding argumentative relations. Furthermore, we showed that sentiment features are not sufficient for distinguishing between support and attack relations.

Since local classifiers are not sufficient for recognizing consistent argumentation structures, we proposed a novel joint model based on integer linear programming that globally optimizes the results of local base classifiers. To this end, we used the argument component classification and argumentative relation identification models as base classifiers and defined several constraints that allow for recognizing several arguments, i.e. several tree structures, within the same paragraph. We showed that our model not only successfully models the natural relationship between argument component types and argumentative relations but also simultaneously improves the performance of both tasks in two different types of discourse (persuasive essays and microtexts). Moreover, the analysis of the predicted structures showed that the model identifies valid tree structures and more consistent results than the local base classifiers. Finally, we compared our approach to the human upper bound and showed that it achieves promising results. In particular, it achieves 97.5% of human performance for segmenting argument components, 95.2% for component

classification, 87.9% for relation identification and 80.5% for stance classification.

Third, we focused on the automatic assessment of natural language arguments (RQ3). To answer the question which quality criteria are appropriate for argumentative writing support, we first investigated theoretical approaches for evaluating arguments. The quality of arguments depends on various criteria such as ethos, pathos, logos and kairos. However, the feedback of argumentative writing support system should be established on objective criteria that are unequivocal and easily comprehensible to ensure the best possible learning effect. Therefore, we focused on logical quality criteria and investigated formal as well as informal approaches for assessing the quality of arguments. We found that formal logic approaches are not applicable to the various kinds of reasoning in natural language arguments. In contrast, informal approaches are not restricted to a particular set of arguments and allow to assess the quality of arguments in everyday discourse. By comparing different informal approaches, we found that the RAS-criteria allow for a fine-grained analysis of the argument quality and to attribute a particular defect to specific components of an argument. In addition, the RAS-criteria allow to separate well-reasoned arguments from illogical arguments. Therefore, we built upon the RAS-criteria and investigated their applicability to arguments in persuasive essays. To this end, we conducted an annotation study with three annotators and showed that humans considerably agree on the sufficiency criterion while the agreement on the relevance and acceptability criterion was too weak to create a reliable corpus. As a result of this annotation study, we introduced for the first time a corpus of arguments reliably annotated with the sufficiency criterion.

We built upon this corpus to address the automatic identification of insufficiently supported arguments. In order to gain a better understanding of insufficiently supported arguments, we experimented with several linguistically motivated features. By doing so, we found that insufficiently supported arguments frequently include lexical cues that correspond with personal experience or specific examples. Furthermore, we found that convolutional neural networks considerably outperform feature-rich SVMs and strong baselines. In particular, the model achieves a promising accuracy of 84.3% for recognizing insufficiently supported arguments which corresponds to 92.5% of human performance. Besides the automatic identification of insufficiently supported arguments, we introduced the novel task of recognizing myside biases in persuasive essays. We modeled this task as a binary document classification and considered an essay as biased if it does not include opposing arguments. We experimented with different models and proposed a novel feature set consisting of lexical features, adversative transitions and syntactic production rules that achieves 77% accuracy and 84% of human performance.

In summary, we introduced three approaches for automatically analyzing arguments in persuasive essays. Our end-to-end argumentation structure parser allows for recognizing fine-grained argumentation structures, whereas the two approaches on argument quality enable the recognition of myside biases and insufficiently supported arguments. Together, these approaches represent a solid basis for establishing novel argumentative writing support systems[1] and may promote the development of novel enabling technologies in areas such as information retrieval, decision support

[1] We introduce an argumentative writing support system that converts the analysis results to human understandable feedback in Appendix A.

systems or intelligent personal assistants.

Future Research Directions

Computational argumentation is still in its infancy and there are various open research questions in this field. The following list represents a subjective overview of the most crucial research directions for future work:

- *Extension to other text types*: In this work, we have shown that our annotation scheme can be reliably applied to persuasive essays. However, persuasive essays exhibit a common structure and it will be even more difficult to apply the annotation scheme to text types with less explicit argumentation structures such as social media data, product reviews or dialogical debates. Nevertheless, we believe that our annotation scheme can be successfully applied to other text types with minor adaptations. Although other text types may not include major claims, previous work has already demonstrated that claims and premises can be reliably annotated in legal cases (Mochales-Palau and Moens, 2011), written dialogs (Biran and Rambow, 2011b) and even over multiple Wikipedia articles (Aharoni et al., 2014). Additionally, it is unknown if our tree assumption generalizes to other text types. Although most previous work considered argumentation structures as trees, other text types may include divergent arguments and even cyclic argumentation structures.

- *Structured machine learning*: Our argumentation structure parser is a pipeline consisting of several consecutive steps. Therefore potential errors of the upstream models are propagated and negatively influence the results of the downstream models. For example, errors of the identification model can result in flawed argumentation structures if argumentatively relevant text units are not recognized or non-argumentative text units are identified as relevant. Another potential issue of the pipeline architecture is that wrongly classified major claims will decrease the accuracy of the model due to the fact that they are not integrated in the joint modeling approach. For this reason, it is worthwhile to experiment in future work with structured machine learning methods that incorporate several tasks during training (Moens, 2013). For example, recent end-to-end deep learning models like proposed by Miwa and Bansal (2016) prevent error propagation and could possibly lead to better results.

- *Reconstruction of enthymemes*: Structural approaches to argument analysis recognize argumentation structures in text assuming that the argument components are explicitly stated. Although this assumption may hold true in text types like persuasive essays, arguments in other text types may be less explicit. Habernal and Gurevych (2016a, p. 27) showed, for instance, that 48% of the claims in user-generated web discourse are implicit. Therefore, another research direction is to automatically reconstruct enthymemes and to derive the standard form from written arguments respectively.

- *Relevance and acceptability*: The results of our annotation study on logical quality of arguments showed that human annotators can reliably recognize

insufficiently supported arguments. However, the results also showed that human annotators agreed only marginally on the relevance and acceptability criterion. In particular, the disagreements about the acceptability of premises is a matter of previous knowledge. For improving the inter-annotator agreement, it may be reasonable to focus on a narrow topic and to provide factual knowledge to the annotators to support the annotation process.

- *Ethos and pathos for argument quality*: Our contribution to quality assessment focused on the logical dimension of arguments. However, the quality of arguments is a product of many different criteria. Future research should also consider "external" factors of argument quality such as ethos or pathos. Another research direction could be to empirically determine which criteria contribute to good arguments. For instance, Habernal and Gurevych (2016b) recently attempted to empirically determine these criteria by rating pairs of arguments. These results could be further enhanced in future work to investigate which types of arguments are perceived stronger than others.

- *Extrinsic evaluation*: In this thesis, we presented an argumentation structure parser as well as two approaches for automatically assessing the quality of arguments in persuasive essays. We integrated our models in an interactive argumentative writing support system for generating human understandable feedback (cf. Appendix A). However, it is still an open question if the proposed feedback types can successfully guide students to improve their arguments and their argumentation skills respectively. In future research, it is necessary to conduct extensive user studies with the system and to investigate if the feedback types are effective for guiding students.

Appendix

A Argumentative Writing Support

The objective of argumentative writing support is to automatically analyze argumentative texts and to provide formative feedback to authors for improving written arguments. In the current thesis, we introduced several models for analyzing arguments in persuasive essays. However, the question how to provide feedback to authors is not yet answered. In this part, we introduce an argumentative writing support system that incorporates the argumentation structure parser introduced in Chapter 5 as well as the models for recognizing myside biases and insufficiently supported arguments introduced in Chapter 6. In Section A.1, we describe the architecture and the software frameworks used for implementing the system. In Section A.2, we introduce different feedback types and describe how the analysis results of the models are converted to human understandable feedback. In Section A.3, we present the user interface of the system and describe the interaction design of the application.

A.1 System Architecture

We implemented the argumentative writing support system as a web application using *Apache Tomcat 7*[2]. The argument analysis models are implemented as *Apache UIMA*[3] pipeline using *DKPro*[4] (Eckart de Castilho and Gurevych, 2014).

Figure A.1 shows the architecture and the components of the system. The web client consists of an *input form* which takes an essay as input, and an *argumentative feedback* page that shows the feedback generated on the server. The server comprises two components: (1) the *argument analysis* component and (2) the *feedback generation* component. The argument analysis component preprocesses the essay using several modules of DKPro (cf. Chapter 5 and Chapter 6). The structural analysis identifies argument components, argument component types and argumentative relations as described in Chapter 5, whereas the quality-based analysis models recognizes myside biases (cf. Section 6.2) and insufficiently supported arguments (cf. Section 6.1). The analysis results are stored on the local file system of the server. In order to convert the analysis results into a human understandable form, the *feedback generation* executes several rules for generating formative feedback (cf. Section A.2). The feedback generation consists of two components: the

[2] `https://tomcat.apache.org`
[3] `https://uima.apache.org`
[4] `https://dkpro.github.io`

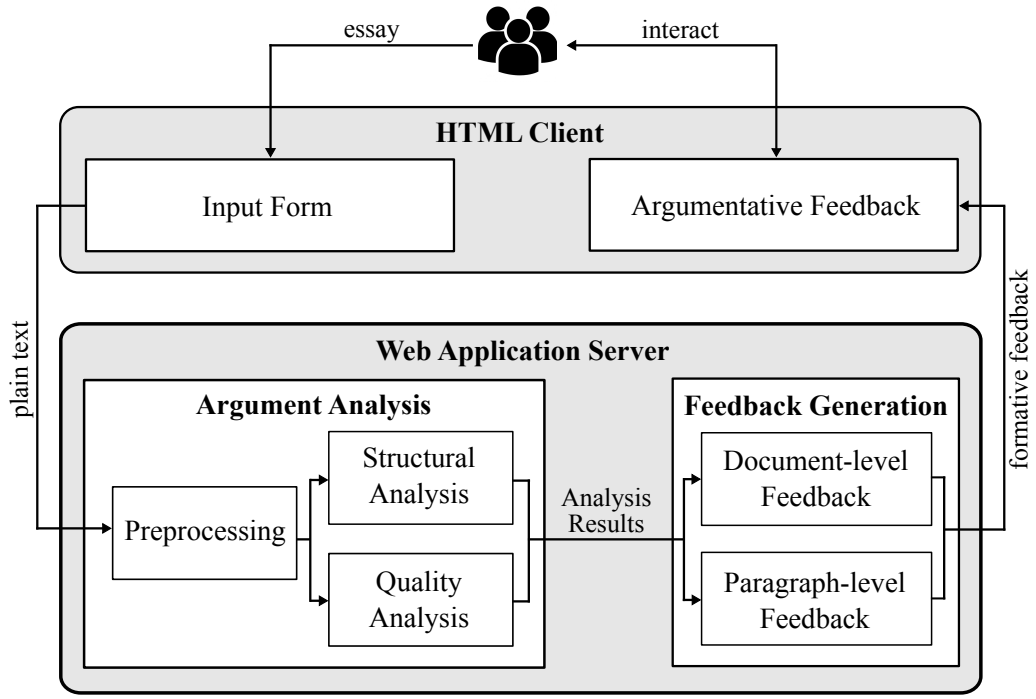


Figure A.1: System architecture of the argumentative writing support system.

document-level component generates feedback about the entire essay, whereas the paragraph-level component generates more detailed feedback about the arguments in individual paragraphs. Both feedback types (document-level and paragraph-level) are serialized in *JavaScript Object Notation* (JSON) which is sent to the client along with the identified argumentation structure.

A.2 Generating Feedback

The feedback generation uses several *test criteria* for deriving feedback from the analysis results. These test criteria are, for instance, "*Is an argument present in body paragraph?*", "*Does the introduction include a major claim?*", or "*Are all claims supported?*". We distinguish between document-level feedback and paragraph-level feedback. Document-level feedback checks the presence of the title, the paragraph structure of the essay which is motivated by writing guidelines for persuasive essays[5] and the presence of myside biases (Wolfe and Britt, 2009). The description of document-level feedback types is given in Table A.1.

Feedback type	***Test criteria***
Title present	There are two succeeding line breaks within the first 20 tokens.
At least four paragraphs	The essay consists of at least four paragraphs.
Opposing arguments present	The myside bias recognition introduced in Section 6.2 didn't classify the essay as biased.

Table A.1: Document-level feedback.

Table A.2 shows the test criteria for the implemented paragraph-level feedback

[5] e.g. those written by Whitaker (2009) or Perutz (2010).

types. In the introduction, we check for the presence of the thesis statement as suggested by Whitaker (2009), the presence of a non-argumentative introduction of the topic and the presence of arguments. The feedback types for body paragraphs check the presence of arguments, the presence of unsupported claims, the ordering of argument components[6], the number of premises per claim and if the arguments are sufficiently supported. In conclusions, we check for a restatement of the major claim as suggested by Whitaker (2009).

Feedback type	***Paragraph***	***Test criteria***
Thesis statement present	Introduction	The structural analysis found at least one major claim in the introduction of the essay.
Topic introduction present	Introduction	The introduction begins with at least two non-argumentative sentences.
No argument in introduction	Introduction	The introduction does not include any arguments, i.e. a claim supported by a premise.
Argument present	Body	The current body paragraph includes at least one argument, i.e. there is one claim supported by at least one premise.
No unsupported claim present	Body	All claims of the current paragraph have incoming relations.
Claim is first component	Body	The first argument component in the body paragraph is a claim.
Appropriate number of reasons	Body	Each claim in the current body paragraph has at least two supporting (or attacking) premises, i.e. each claim has at least two incoming relations.
Sufficiently supported	Body	The sufficiency classification model (cf. Section 6.1) classified the current body paragraph as sufficient.
Restatement of thesis statement	Conclusion	There is a major claim present in the conclusion of the essay.

Table A.2: Paragraph-level feedback.

A.3 User Interface and Interaction Design

The user interface ("argumentative feedback" in Figure A.1) is based on a *check box metaphor*. It shows whether the submitted essay fulfills (green) or violates (red) the feedback types introduced in Section A.2. The user interface also visualizes the identified argumentation structure and provides a detailed description of each feedback type explained. It consists of the following components (Figure A.2):

1. *Essay and Argument Components*: This component shows the submitted essay split into its paragraphs. It also highlights the argument components.

2. *Selection of Feedback Types*: This menu enables to toggle between (1) document-level feedback, (2) paragraph-level feedback, and (3) visualization of the argumentation structure.

[6] Britt and Larson (2003) showed that presenting the claim before the reasons significantly improves the recall of arguments.

3. *Feedback Types*: This component shows the feedback types defined in Section A.2. The feedback type is shown in green if its test criterion is fulfilled and red if violated. For visually linking a feedback type to the argument components, we used a *brushing and linking* technique. Each feedback type is associated with a set of argument components, which are highlighted in the essay component (1) if the user hovers over a particular feedback type.

4. *Feedback Details*: This component shows a detailed description of the currently selected feedback type for guiding students to improve their arguments.

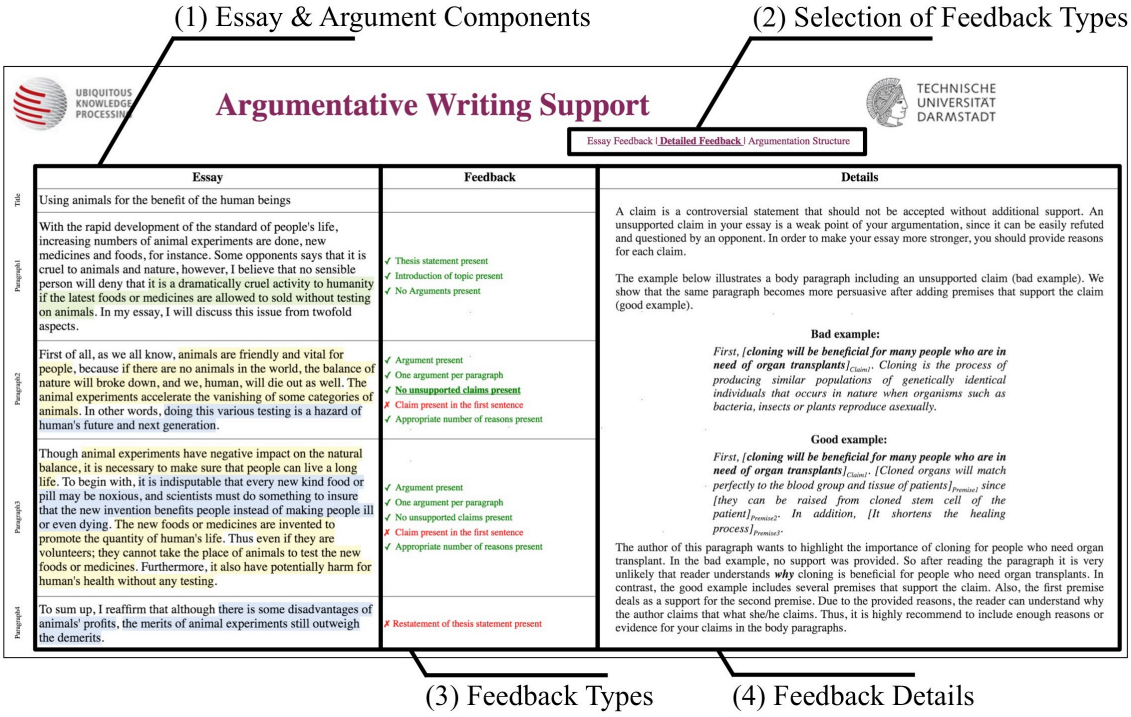


Figure A.2: User interface of the argumentative writing support system.

Figure A.2 shows a screenshot of the user interface. The screenshot in Figure A.3 shows the user interface visualizing the structure of a submitted essay. In the following sections, we provide the instructions[7] for each feedback type of the argumentative writing support system described above.

A.4 Document-Level Feedback

Title Present

A title introduces the topic in the first line of your essay. You should provide an appropriate title for your essay, so that the reader can get an overview of the essays' content. Examples are:

[7] These descriptions have been written collaboratively by the author of this thesis and Anshul Tak in the context of his term paper "*An Interactive System for Argumentative Writing Support*", Technische Universität Darmstadt, 2016.

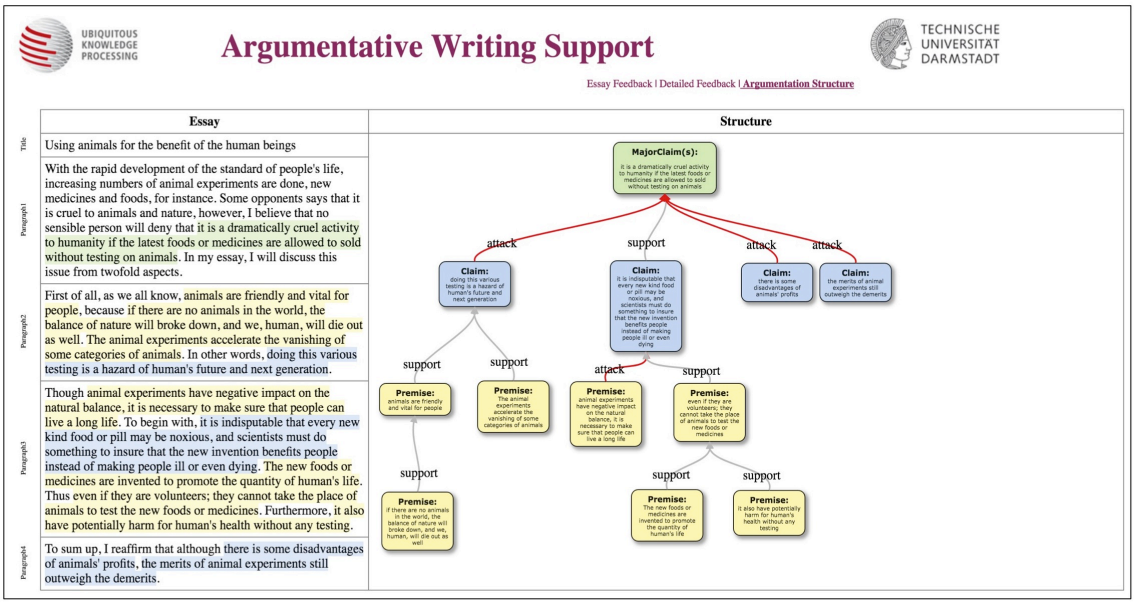


Figure A.3: Visualization of the argumentation structure.

- "*Should students be taught to compete or to cooperate?*"
- "*Living and studying overseas*"
- "*Leather and fur clothes should be banned*"

At Least Four Paragraphs

Each single idea should be separated into a single paragraph and meaningfully associated with your thesis statement. An essay should at least include the following paragraphs separated by line breaks: The first paragraph is called ***introduction***. It introduces the controversial topic of the essay and your stance about the topic in a thesis statement. All subsequent paragraphs except the last one are called ***body paragraphs***. Each of these individual paragraphs includes a single argument either supporting or attacking your opinion on the topic. Accordingly, each body paragraph should include a single claim, which is a central component of your argument, followed by one or more reasons for supporting your claim. Finally, the last paragraph is called ***conclusion***. It restates your thesis statement and provides a brief summary of your argumentation.

This common essay structure is illustrated in the following Figure:

Opposing Arguments Present

The myside bias is a tendency to ignore evidence against one's own position [1]. Consequently, people argue in a manner biased towards their own prior beliefs which frequently results in weak arguments [2]. It has been shown, that considering opposing positions when formulating arguments significantly improves the argumentation quality, the precision of claims and the elaboration of reasons [1]. Thus, in order to

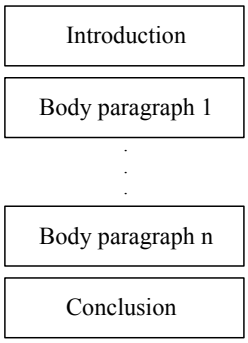


Figure A.4: Common structure of an argumentative essay.

increase the persuasiveness and quality of the essay, the author should include at least one opposing argument. The opposing argument is usually presented in the last paragraph before the conclusion. This paragraph is also called "rebuttal paragraph".

[1] Christopher R. Wolfe and M. Anne Britt. 2009. Argumentation schema and the myside bias in written argumentation. Written Communication, 26(2):183-209.

[2] Keith E. Stanovich, Richard F. West, and Maggie E. Toplak. 2013. Myside bias, rational thinking, and intelligence. Current Directions in Psychological Science, 22(4):259-264.

A.5 Paragraph-Level Feedback

Thesis Statement Present

The ***thesis statement*** is the most important sentence in your essay, since it represents the main idea of your essay and the entire essay is based on this statement. It also represents your opinion on the topic. It is not a fact nor a question but it represents ***your point of view*** on the topic. The thesis statement is also the answer to the research question of your essay and thus it should answer the question of the research question. Usually, the thesis statement is ***the last sentence of the introduction***. It should be understandable and easy to comprehend. The following examples illustrate two introductions taken from essays about cloning:

> *Since researchers at the Roslin Institute in Edinburgh cloned an adult sheep, there is an ongoing debate if cloning technology is morally and ethically right or not. Some people argue for and others against and there is still no agreement whether cloning technology should be permitted. However, as far as I'm concerned,* ***[cloning is an important technology for humankind]***$_{ThesisStatement}$ *since it would be very useful for developing novel cures.*

In the above example, the author has first introduced the topic. In the last sentence of the introduction paragraph a thesis statement is presented. By looking at the thesis statement, one can deduce that the author is in favor of cloning.

> *Cloning animals is possible since several years, and this has now opened up the possibility of cloning humans too. Although there are clear benefits to humankind of cloning to provide spare body parts. I believe* ***[it raises a number of ethical issues]***$_{ThesisStatement}$.

The last sentence contains the thesis statement of the author in another essay on the same topic, i.e. cloning. The author has illustrated that cloning definitely has its benefits but due to ethical issues she/he is not in favor of cloning. After reading the introduction, the reader can conclude that the essay will include arguments to support the thesis statement and present arguments against cloning.

Thus, it is clearly visible that the thesis statement is an important component of an essay since the whole essay depends on. Furthermore, developing a clear stance on the topic facilitates the development of well-reasoned arguments.

Topic Introduction Present

The introduction of an argumentative essay usually includes 2-4 sentences which introduce the topic. The introduction should immediately gain the attention of the reader and briefly introduce the main points of your essay. So it is important that the introduction includes a brief and interesting overview of the essay's topic. Note that this overview is ***non-argumentative*** but descriptive. Furthermore, the introduction should include a thesis statement that introduces your personal opinion about the topic. This thesis statement is usually included in the last sentence of the introduction.

No Argument in Introduction

The main purpose of the introduction is to provide a brief overview of the topic and to introduce your stance on the topic in a thesis statement (the thesis statement is usually present in the last sentence of the introduction). Your actual arguments should be stated in the following body paragraphs of your essay. Also note that including only one argument per body paragraph makes your essay easier to comprehend and clearly separates the different points of your argumentation.

Argument Present

Each body paragraph should include a single argument that either supports or attacks the stance of the author. An argument includes a claim that is supported by one or several premises. The claim is the central component of the argument. It introduces a new idea and connects the idea with the opinion of the author. Premises constitute the reasons for believing the claim to be true or false. Below in Figure A.5, we present a common structure of an argument:

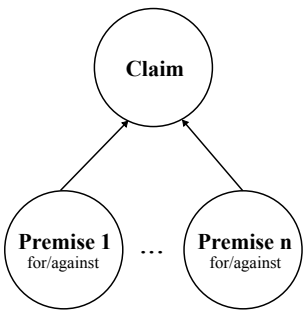


Figure A.5: Structure of an argument.

The following body paragraphs illustrates an argument about cloning:

> *First, [cloning will be beneficial for many people who are in need of organ transplants]$_{Claim}$. [Cloned organs will match perfectly to the blood group and tissue of patients]$_{Premise1}$ since [they can be raised from cloned stem cell of the patient]$_{Premise2}$. In addition, [It shortens the healing process]$_{Premise3}$.*

In the above body paragraph, the author presents a claim which states the benefits of cloning for people requiring organ transplants followed by reasons supporting claim.

No Unsupported Claim Present

A claim is a controversial statement that should not be accepted without additional support. An unsupported claim in your essay is a weak point of your argumentation, since it can be easily refuted and questioned by an opponent. In order to make your essay stronger, you should provide reasons for each claim.

The example below illustrates a body paragraph including an unsupported claim (bad example). We show that the same paragraph becomes more persuasive after adding premises that support the claim (good example).

Bad example

> *First, [cloning will be beneficial for many people who are in need of organ transplants]$_{Claim}$. Cloning is the process of producing similar populations of genetically identical individuals that occurs in nature when organisms such as bacteria, insects or plants reproduce asexually.*

Good example

> *First, [cloning will be beneficial for many people who are in need of organ transplants]$_{Claim}$. [Cloned organs will match perfectly to the blood group and tissue of patients]$_{Premise1}$ since [they can be raised from cloned stem cell of the patient]$_{Premise2}$. In addition, [it shortens the healing process]$_{Premise3}$.*

The author of this paragraph highlights the importance of cloning for people who need organ transplants. In the bad example, no support was provided. So after reading the paragraph, it is very unlikely that the reader understands why cloning is beneficial for people who need organ transplants. In contrast, the good example includes several premises that support the claim.

Claim is First Component

A claim is the central component of an argument. It introduces the main idea of the paragraph and makes your point about this idea. In addition, it relates your idea to the thesis statement. It has been shown that arguments are easier to recall and to understand if the claim is present in the first sentence of a body paragraph [1].

Bad example

[Variations in genetic model help species to sustain and evolve]$_{Premise}$. [With a similar genetic model, the entire population is subjected to genetic diseases because of similar genetic composition]$_{Premise}$. [Cloned individuals will be exposed to a higher risk of genetic diseases]$_{Claim}$. [Nettie Stevens and Edmund Wilson (1905) illustrated that when a XX genes in females and XY genes in males are cloned, these identical genes have greater risk for acquiring diseases from both parents]$_{Premise}$.

Good example

[Cloned individuals are exposed to a higher risk of genetic diseases]$_{Claim}$. [With a similar genetic model, the entire population is subjected to genetic diseases because of similar genetic composition]$_{Premise}$. [Nettie Stevens and Edmund Wilson (1905) illustrated that when a XX genes in females and XY genes in males are cloned, these identical genes have greater risk for acquiring diseases from both parents]$_{Premise}$. [Variations in genetic model help species to sustain and evolve]$_{Premise}$.

[1] M. Anne Britt and Aaron A. Larson. (2003) Constructing representations of arguments. Journal of Memory and Language 48(4): 794-810.

Appropriate Number of Reasons

The more support you provide for each of your claims, the stronger your argument will be. In order to strengthen your arguments, we recommend to provide at least two reasons (premises) for each claim and in each body paragraph respectively. These will make your arguments stronger and more convincing.

Bad example

Second, [cloning animals enables novel developments in science]$_{Claim}$ because [scientists use animals as models in order to learn about human diseases]$_{Premise}$.

Good example

Second, [cloning animals enables novel developments in science]$_{Claim}$ because [scientists use animals as models in order to learn about human diseases]$_{Premise1}$. Also, [animals are biologically similar to humans with much smaller lifespan and susceptible to the same health problems]$_{Premise2}$.

In first example, the argument sounds very unconvincing to the reader since various questions were not answered like e.g. "*Why animals?*", "*Are human diseases similar to that of animals?*" or "*Is animal life subjugate to human life?*". However, in the second example some of these questions were answered, which can convince the reader of the author's opinion. Hence, it is advised to put at least two premises for a claim in an argument.

Sufficiently Supported

The premises of a well-reasoned argument should provide enough evidence for accepting or rejecting its claim. This criterion is also known as ***sufficiency criterion***. An argument complies with the sufficiency criterion if its premises provide enough evidence for accepting or rejecting the claim. The following example argument illustrates a violation of the sufficiency criterion:

Bad example

> *It is an undeniable fact that tourism harms the natural habitats of the destination countries. As Australia's Great Barrier Reef has shown, the visitors cause immense destruction by breaking corals as souvenirs, throwing boat anchors or dropping fuel and other sorts of pollution.*

The premise of this argument represents a particular example (second sentence) that supports a general claim in the first sentence. The argument is a generalization from one sample to the general case. However, a single sample is not enough to support the general case. Therefore, the argument does not comply with the sufficiency criterion.

Good example

> *Cloning will be beneficial for people who are in need of organ transplants. Cloned organs will match perfectly to the blood group and tissue of patients since they can be raised from cloned stem cells of the patient. In addition, it shortens the healing process.*

Example 2 illustrates a sufficiently supported argument. It is reasonable to accept that transplantation patients will benefit from cloning if it enables a better match and an accelerated healing process.

Restatement of Thesis Statement

The thesis statement is the most important sentence in your essay since the entire essay is based on this statement. It represents your opinion on the topic. It is not a fact nor a question but it represents your point of view on the topic. The thesis statement is also the answer to the research question of your essay and thus it should answer the question of the research question. Usually, the thesis statement is the last sentence of the introduction. It should be also restated in the conclusion to remember the reader of your stance on the topic and to conclude your essay properly.

> *To sum up, although permitting cloning might bear some risks like misuse for military purposes, I strongly believe that* ***[this technology is beneficial for humanity]***$_{ThesisStatement}$*. It is likely that this technology bears some important cures which will significantly improve life conditions*

The above example illustrates a conclusion of an essay about cloning. In the first sentence of the conclusion the author starts with a short rebuttal to anticipate opposing positions followed by a restatement of the thesis statement that cloning is beneficial for humanity. Furthermore, the author summarizes the most important points in the last sentence of the essay.

A.6 Summary

In this part, we introduced an argumentative writing support system that incorporates the analysis models developed in this thesis. We introduced the architecture and the test criteria used to derive human understandable feedback from the argument analysis results. Moreover, we presented the interaction design of the user interface and provided the detailed feedback descriptions shown in the user interface for each feedback type. The presented argumentative writing support system forms the foundation for conducting extrinsic evaluations and for evaluating its effectiveness for supporting authors.

B Overview of Annotated Corpora

Corpus	*Domain*	*Lang*	*Task*	*ArgGran*	*CompGran*	*SiMuDo*	*#Docs*	*#Comps*	*Reliability*
Anand et al. (2011)	online discussions	en	AT	macro	-	-	4,873	-	$\kappa = .27$
Biran and Rambow (2011a)	blog threads	en	CI+CC	micro	multi	single	309	1,377	$\kappa = .69$
Biran and Rambow (2011b)	Wikipedia talk pages	en	CI+CC	micro	multi	single	118	2,404	$\kappa = .75$
Boltužić and Šnajder (2014)	online discussions	en	SI	macro	-	-	pairs 2,436	-	$\kappa = .49$
Cabrio and Villata (2012a)	online discussions	en	AT+SI	macro	-	-	pairs 200	-	-
Cabrio and Villata (2014)	several	en	SI	macro	-	-	pairs 792	-	$\kappa = .71$
Eckle-Kohler et al. (2015)	news	de	CI+CC	micro	multi	single	88	1,708	$\alpha_U = .40$
Florou et al. (2013)	focused web crawl	gr	-	macro	-	-	677	-	-
Goudas et al. (2014)	social media	gr	CI+CC	micro	clause	single	204	760	-
Habernal and Gurevych (2016a)	web content	en	CI+CC	micro	multi	single	340	1,319	$\alpha_U = .48$
Kirschner et al. (2015)	scientific articles	de	CI+SI	micro	sentence	single	24	~2,700	$\kappa = .43$
Kwon et al. (2007)	online comments	en	CI+CC	micro	sentence	single	119	240	$.62 \leq \kappa \leq .80$
Mochales-Palau and Moens (2009)	court cases	en	CI+CC	micro	sentence	single	47	1,067	$\kappa = .75$
Peldszus and Stede (2016)	micro texts	de/en	CC+SI	micro	clause	single	112	576	$\kappa = .83$
Reed et al. (2008)	various	en	CI+CC+SI	micro	clause	single	~700	~2,000	-
Rinott et al. (2015)	Wikipedia articles	en	CI+CC	micro	clause	multi	547	7,254	$\kappa = .39$
Rosenthal and McKeown (2012)	blogs and discussions	en	CI	micro	sentence	single	336	2,479	$\kappa = .53$
Sardianos et al. (2015)	news	gr	CI+CC	micro	clause	single	300	1,191	F1 = .76
Somasundaran and Wiebe (2010)	online discussions	en	AT	macro	-	-	3,921	-	-
Stab and Gurevych (2014b)	student essays	en	CI+CC+SI	micro	clause	single	90	1,552	$\alpha_U = .71$
Stab and Gurevych (2016)	student essays	en	CI+CC+SI	micro	clause	single	402	6,089	$\alpha_U = .77$
Walker et al. (2012)	online discussions	en	AT	macro	-	-	390k	-	$.22 \leq \kappa \leq .62$

Table B.3: Overview of annotated corpora (abbreviations are explained in Section 3.1).

C List of Lexical Indicators

Table C.4 shows all of the lexical indicators we extracted from 30 persuasive essays. The lists include 24 forward indicators, 33 backward indicators, 48 thesis indicators and 10 rebuttal indicators.

Category	***Indicators***
Forward (24)	"As a result", "As the consequence", "Because", "Clearly", "Consequently", "Considering this subject", "Furthermore", "Hence", "leading to the consequence", "so", "So", "taking account on this fact", "That is the reason why", "The reason is that", "Therefore", "therefore", "This means that", "This shows that", "This will result", "Thus", "thus", "Thus, it is clearly seen that", "Thus, it is seen", "Thus, the example shows"
Backward (33)	"Additionally", "As a matter of fact", "because", "Besides", "due to", "Finally", "First of all", "Firstly", "for example", "For example", "For instance", "for instance", "Furthermore", "has proved it", "In addition", "In addition to this", "In the first place", "is due to the fact that", "It should also be noted", "Moreover", "On one hand", "On the one hand", "On the other hand", "One of the main reasons", "Secondly", "Similarly", "since", "Since", "So", "The reason", "To begin with", "To offer an instance", "What is more"
Thesis (48)	"All in all", "All things considered", "As far as I am concerned", "Based on some reasons", "by analyzing both the views", "considering both the previous fact", "Finally", "For the reasons mentioned above", "From explanation above", "From this point of view", "I agree that", "I agree with", "I agree with the statement that", "I believe", "I believe that", "I do not agree with this statement", "I firmly believe that", "I highly advocate that", "I highly recommend", "I strongly believe that", "I think that", "I think the view is", "I totally agree", "I totally agree to this opinion", "I would have to argue that", "I would reaffirm my position that", "In conclusion", "in conclusion", "in my opinion", "In my opinion", "In my personal point of view", "in my point of view", "In my point of view", "In summary", "In the light of the facts outlined above", "it can be said that", "it is clear that", "it seems to me that", "my deep conviction", "My sentiments", "Overall", "Personally", "the above explanations and example shows that", "This, however", "To conclude", "To my way of thinking", "To sum up", "Ultimately"
Rebuttal (10)	"Admittedly", "although", "Although", "besides these advantages", "but", "But", "Even though", "even though", "However", "Otherwise"

Table C.4: List of lexical indicators.

D Guidelines for Annotating Argumentation Structures

D.1 Introduction

Argumentation Mining is an interdisciplinary research area that incorporates philosophy, psychology linguistics and computer science for establishing argumentation models and automated methods for identifying arguments in written texts. These tools will not only provide novel possibilities for educational applications like intelligent writing assistance, information retrieval platforms or automated assessment tools but will also open new opportunities for improving current legal information

retrieval applications or policy modeling platforms. However, a major prerequisite for developing novel Natural Language Processing (NLP) methods that are able to identify argument components and argumentative relations in written texts is the availability of annotated corpora. Due to this requirement and the complex structure of argumentative discourse, "*the automatic detection of arguments has been left nearly unstudied*" till 2008 (Reed et al., 2008).

The goal of this study is to create a language resource for argumentation mining by manually annotating the structure of arguments in persuasive essays. Since the annotation an assessment of arguments is a complex task, this document first provides a brief introduction to argumentation theory including the definitions of argument components, argumentation structures and argumentative relations before describing the steps of the annotation in detail.

Arguments in a Nutshell

An argument consists of several statements. In its simplest form, it includes one claim that is supported by at least one premise (Peldszus and Stede, 2013a; Britt and Larson, 2003; Toulmin, 1958). The *claim* (or also called conclusion Mochales-Palau and Moens (2009)) represents a controversial statement which the author tries to persuade the reader of. It is usually a proposition or assumption and should not be accepted by the reader without additional support. This characteristic distinguishes arguments from explanations where the conclusion is a true statement that is not arguable (e.g. an event that happened in the past). The second component of an argument, the *premise* (or sometimes called support (Besnard and Hunter, 2008) or reason (Britt and Larson, 2003)), underpins the plausibility of the claim. It is usually added by the proponent (writer) for persuading the reader of the claim. Considering the simplest form of an argument, a premise can be seen as a justification for the claim, whereas more complex argumentation structures can also include premises that aim at refuting a claim. These more complex structures are basically graphs that connect premises and claims by means of different relations. Figure D.1 illustrates the simplest form of an argument consisting of only one claim that is supported by one premise.

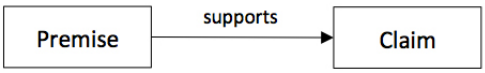


Figure D.1: Simple form of an argument.

Even such a simple form of an argument can be expressed in many different ways in written text. Some example patterns that can be found in written argumentation are the following:

<*claim*> because <*premise*>.
Since <*premise*> it is feasible that <*claim*>.
In view of the fact that <*premise*> it follows that <*claim*>.
<*premise*>. Therefore, <*claim*>.

However, there are many ways to express arguments in written texts and frequently the cue phrases (e.g. discourse markers like "*therefore*" or "*because*") are not present or misleadingly used in real texts. For instance, the following argument includes exactly one claim-premises-pair without any indicator:

> "*By wearing school uniforms, pupils are not able to develop their own style of fashion. Wearing school uniforms will have negative influence on the development of their characters.*"

In this example the second sentence is the claim which is supported by the first sentence but there is no indicator present which signals the argumentative relation between the two statements/sentences. The structures we will discover in persuasive essays are usually more complex and consist of several premises that either support or attack a certain claim. For instance, lets consider the following example:

> "*Although wearing school uniforms might foster the team spirit, it restricts the right of self-determination. Therefore, we should not force pupils to wear school uniforms.*"

This example includes three argument components (two in the first sentence and another one in the second sentence). The second sentence includes the claim "*we should not force pupils to wear school uniforms*" which is supported by the second premise in the first sentence "*it restricts the right of self-determination*". The first statement "*wearing school uniforms might foster the team spirit*" is a counter reason which attacks the claim. It is added by the author to make the argument stronger against any potential contra argument by an opponent. In this study we will annotate counter reasons as another premise which is connected to its target statement (in this example the claim) with an attack relation. The argumentation structure includes therefore three statements and two argumentative relation. The structure of this argument is illustrated in Figure D.2: In the next section we will introduce

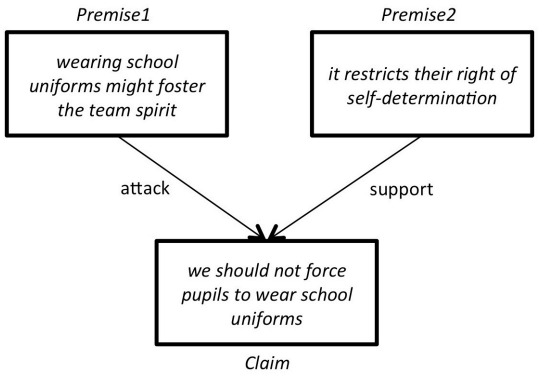


Figure D.2: Example structure of an argument.

the type of documents in which we will annotate argumentation structures and also describe some structural properties of these documents which might facilitate the annotation process.

Persuasive Essays

This annotation study is conducted on a set of persuasive essays. These essays usually exhibit a certain structure which will be briefly explained in this section. Usually, an essay starts with an introduction which includes a short description of the topic. The introduction describes the controversial topic of the essay and rarely includes arguments. However, the introduction frequently include a *thesis statement* which expresses the stance of the author about the topic. We refer to this statement as *major claim* (cf. Section D.3).

The actual arguments which either support or attack the major claim are given in the paragraphs following the introduction. Usually there are about two or four paragraphs before the essay concludes with a concluding paragraph. The last paragraph frequently includes a re-statement of the major claim which we will also annotate. It might also include a summary of the reasons supporting the major claim which we will annotate as claims either for or against the major claim. In rare cases the last paragraph also includes complete arguments which should also be annotated. Frequently, the very last sentences include some *recommendations* for future actions which can be considered as a result of the authors' discussion. These recommendations are not argumentative and should be neither annotated as major claim nor as arguments.

Figure D.3 illustrates the common structure of persuasive essays including the introduction, the body paragraphs and the concluding paragraph. The actually

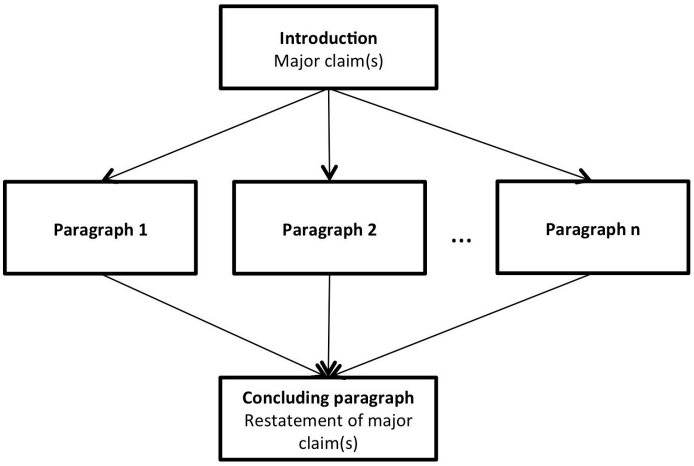


Figure D.3: Common structure of persuasive essays.

argumentation structure which we aim to annotate in this study will be a tree structure. The root node of this tree is the major claim (if the major claim is restated several times in the introduction or in the conclusion, this node will include several but semantically very similar statements/annotations). The following nodes in the tree structures are the claims of the arguments and the premises are the reasons given for underpinning the claims. To illustrate this in more detail figure D.4 illustrates an example argumentation structure. The relations from the claims of the arguments to the major claim are dotted since we will not explicitly annotated them. The relation of each argument to the major claim is indicated by a *stance*

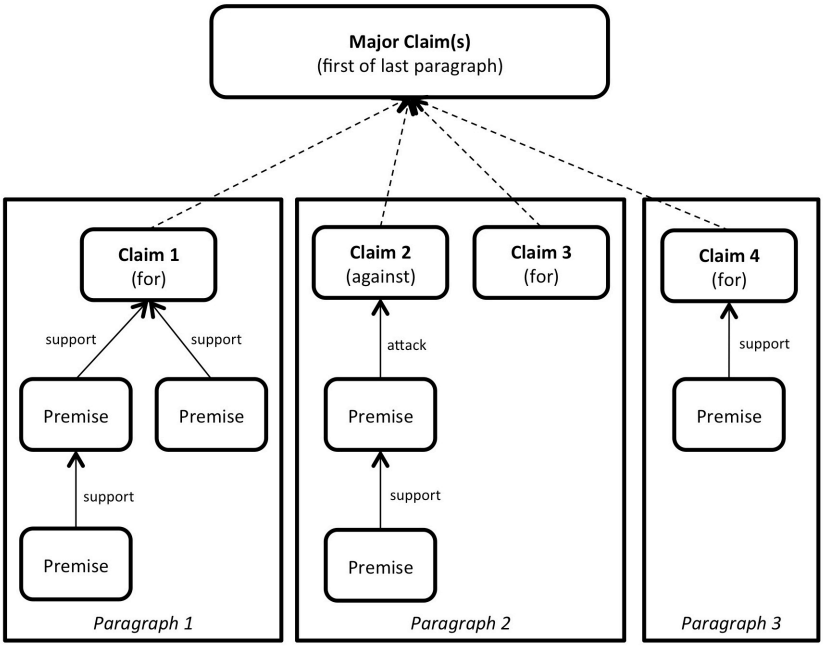


Figure D.4: An example of the argumentation structure of a persuasive essay.

attribute of each claim. This attribute can either be *for* or *against* as illustrated in figure D.4. Having described the argumentation structure that we aim to annotate in this study, the next section provides a brief overview of the annotation process before a more detailed description of each individual annotation step is given in the following chapters.

Overview of the Annotation Process

Previous sections briefly described the argumentation structures that we aim to annotate in this study. For annotating these structures, we split the annotation process into two steps: (1) Annotation of argument components and (2) annotation of argumentative relations. Each of these steps is further divided into different smaller sub steps

Annotation of argument components

The first step of the annotation process focuses on the annotation of argument components. It is further divided into 4 steps which should be followed in the described order:

1. Annotation of the Major Claims: In this step we annotate the major claim(s) which are either located in the introduction or conclusion of the essay. In seldom cases it might also be possible that several reformulations are present. In these cases each should be annotated. The details for annotating major claims are given in section D.3.

2. Annotation of Claims: Each claim is the central component of an argument

which either supports or attacks the major claim. So, it can be considered as a reason for the major claim. Claims are usually included in the paragraphs between the introduction and conclusion. In some cases there are also claims present in the introduction or conclusion, e.g. as a summarization as the key reasons for the major claim. These should also be annotated. Each claim has a stance attribute which incites if the argument is either for or against the major claim of the author. A detailed description how to annotate claims is provided in section D.4.

3. Annotation of Premises: In the final step of the argument component annotation, the premises for the claims are annotated. These are the reasons given by the for supporting or attacking the claims. Usually, the premises are included in the body paragraphs located closely to the claims. The detailed description of the annotation of premises is given in section D.5.

Annotation of argumentative relations
The second step focuses on the annotation of argumentative relations. In particular, we will annotate support and attack relations holding between the argument components to identify the structure of the arguments. For instance, it might be possible that some premises are sequentially connected with support relations or conversantly support a claim. The details of this step are described in chapter D.6.

D.2 Annotation of Argument Components

This chapter describes the annotation of argument components in detail. Section D.2 describes some general rules for annotating argument components. These rules focus on the aspect of the boundaries of argument components which should be followed for each of the three argument components (major claims, claims and premises). Section D.3 focuses on the annotation of major claims whereas the following two section focus on the the annotation of claims and premises. In particular, the annotation of claims and premises is closely related. So, it is important to read all sections and to gain a good understanding before starting with the actual annotation process. Each section includes several examples explaining the details of the argument component annotations in detail[8].

Argument Component Boundaries

The argument examples in section D.1 already showed that argument components do not necessarily cover a whole sentence. Frequently, a sentence might include several argument components which should be annotated separately (e.g. a claim and a premise in one sentence). In addition, so called "*shell language*", e.g. phrases like "*I am strongly convinced*", "*Another reason is that*" or "*From all these reasons follows that*" are not relevant for the content of the argument and should not be annotated. In this section, we will first provide a list of general rules and some examples which should help to identify the boundaries of each argument component.

[8] Some of the examples are taken from real essays either from `http://www.buowl.boun.edu.tr` or `http://www.essayforum.com`

Completeness Rule: An argument component should always cover a statement, which can stand in isolation as a complete sentence. A simple test to verify if an annotated component is a complete statement is to prepend the clause "*It is true that, <claim>*". If the resulting sentence is grammatically correct, the annotation is valid according to the completeness rule.

Relevance Rule: Include all words which are relevant for the argument component. This means that all content relevant subordinate clauses should be included in the component annotation. This also includes temporal information like "*In ancient times*", "*Recently*" or "*These days*" at the beginning of a sentence since it might be not possible to understand the whole argument without these times specifications.

Shell Language Rule: As mentioned before, shell language has no bearing on the context and thus, it is not relevant for the arguments. Indeed such phrases might include indicators which facilitate the identification of argument components however, for the content of the argument they are not important. Besides the above mentioned expressions, shell language might include terms and phases like "*For example*", "*According to the previous fact*", "*As can be seen*", "*Another important point which contributes to my argument is that*", "*I agree to this view that*", "*In this context*", etc. There is only one *exception* in which shell language should be included in the annotation, namely if it contains a negation that is important for the content of the argument. For example, the phrases "I do not agree that", "I disagree with the view that", etc. In these cases the shell expression should be included since it changes the meaning of the argument component.

Splitting Rule: A sentence should only be annotated completely if and only if it does not include an inference step between several statements and if it does not include shell language. That means, even if a sentence includes several complete statements, the sentence should only be splitted into two (or several) argument components if one statement is a reason for the other. In particular, it is important that sentence which include several complete statements connected with conjunctions like "*and*" or "*or*" usually do not include an inference step. This might for example happen if a sentence contains several reasons for a claim expressed in another sentence. In this case, the two premises should be annotated as one argument component. This also holds for conditional sentence (*if...then-statements*) since those do not include an inference step between a claim and a premise.

Punctuation Rule: Punctuations at the end of an argument component should not be included in the annotation.

The following examples illustrate correct argument component annotations according to these rules (the text in squared brackets illustrates the correct boundary for the argument component annotation):

> **Example 2.1.1**: "[Because of convenience many people drive with their own car]."

In this example the whole sentence should be annotated because otherwise the rel-

evance rule is violated. The term "*convenience*" is important for the statement. its also important to note that the shell expression "*Because of*" has also to be included because otherwise the component would not be a complete statement.

> **Example 2.1.2**: "This is due to the fact that [school uniforms are quite expensive and not every student can afford them]."

In this example the shell language at the beginning of the sentence is not included in the annotation due to the Shell Language Rule. The following two statements could be indeed annotated as independent statements, however, this would violate the splitting rule, since both are reasons given against wearing school uniforms.

> **Example 2.1.3**: "[More advance sport lessons should be provided during primary education], since [the health of students will benefit from regular exercises]."

This sentence includes a claim and a premise. Therefore, it should be splitted into two argument components.

> **Example 2.1.4**: "[I do not agree with the opponents of nuclear power]."

In this example, the shell language should be included included because otherwise the major claim would not be a complete sentence and thus the annotation would violate the Completeness Rule.

> **Example 2.1.5**: "[I disagree with the viewpoint that school uniforms have positive effects]."

In this example the shell expression at the beginning of the sentence includes a negation which is important for the content of the argument. Therefore, it should be included in the annotation.

D.3 Annotation of Major Claims

In persuasive essays the *major claim* represents the stance of the author about the essay topic. It is also called *thesis statement* and frequently indicated by opinion expressions like "*From my point of view...*", "*In my opinion...*", "*I strongly believe that...*", etc. Usually, the major claim is present in the introduction or conclusion of an essay or in both. In the introduction it has the characteristics of a general assertion or an opinion with respect to the topic, whereas in the conclusion the major claim summarizes the argumentation according to the author's stance.
For annotating major claims you should follow the rules for argument component boundaries described in the previous section. Please also note that all occurrences of the major claim should be annotated and that in some cases even the introduction or conclusion might include several reformulations which should be individually annotated.
For getting familiar with this particular type of argument component, we will investigate some examples of introductions and conclusions from real essays (the major claim is in squared brackets and bold-face):

> **Example 2.2.1** (Introduction): "Cloning is creating a genetic copy or replica of cells, tissues, embryos, and genes of an already existing organism. Thanks to advances in new technology, cloning of animals has succeeded, but a human has not been cloned so far because of the lack of technology and the prohibition by governments. I think that [**a human cannot be cloned**] because human cloning involves many risks."

In this introduction (2.2.1) the major claim is clearly indicated by the opinion expression "*I think that...*". Following the rules of argument component boundaries not the whole sentence should be annotated since the opinion expression can be considered as shell language and the statement following the discourse connective "*because*" is a reason given in support for the major claim. Therefore, the sentence includes an inference step between several statements. Since the second statement in the sentence is directly supporting the major claim it should be annotated as a claim (cf. Section D.4).

> **Example 2.2.2** (Conclusion): "As a result, [**human cloning may have advantages but no disadvantages**]. Yet there is no experiment of cloning a human so we should try it first by cloning every human for possible organ transplantation. Thus, we can make sure that cloning is used for good intentions."

In this conclusion (2.2.2) the major claim is not indicated by an opinion expression. However, it should be obvious that the first sentence includes the summarizing statement which represents the stance of the author. The shell expression "*As a result,*" is not included in the annotation.

> **Example 2.2.3** (Conclusion): "To sum up, for most people it might be the biggest happiness to have children. However, I firmly believe that [**having children is not everything in life**]. People can also live full and accomplished lives without children."

In example 2.2.3 the major claim is again signaled by an opinion expression which is not included since it is not important for the content of the major claim.

> **Example 2.2.4** (Conclusion): "In sum, I think human cloning may cause a lot of big and important problems if humans are cloned. However, I think cloning of animals and organs is beneficial. So [**'yes' to cloning of animals and organs, but 'no' to human cloning**]."

Example 2.2.4 includes several opinion expressions which represent conditional statements. However, the major claim which represents the stance of the author is included in the last sentence.

> **Example 2.2.5** (Introduction): "There are many ways for people to be happy. Some people are happy maintaining successful business affairs, some of them are happy having big amounts of money and many of them are happy bringing up their children. In my opinion, [**children are the**

> **ultimate bliss in our lives and if I reach the suitable age for marriage; I really want to have at least two children**]. Not only me but also many people plan to have children of their own as they add beauty to our lives."

In example 2.2.5 the major claim covers several statements. Since there is no inference included in this sentence, all the statements except the shell phrase at the beginning are annotated as one single major claim.

> **Example 2.2.6** (Introduction): The idea of school uniforms seems like an antiquated concept for many people. Unless a child attends private school, it is not normally practiced by children and families. Students studying in schools requiring school uniforms generally perform very well academically and seem happy wearing the same outfit every day. [**There are many benefits to wearing school uniforms that schools all around the world should incorporate into their public schools**].

Example 2.2.6 does not include any indicator like opinion expressions. In this case it is more complicated to identify the major claim. However, from the context ist should be clear that the previous sentence are some reasons supporting the stance of the author, namely the last sentence which is the major claim.

> **Example 2.2.7** (Introduction): "As the way to cloning has been found, there has been a debate about if it is right or wrong to clone a human. Some people think that cloning is beneficial for humankind while other people argue that [**cloning has too many disadvantages**]. I agree with the latter view and will give several reasons in the following paragraphs."

In example 2.2.7 the last sentence includes a stance expression which might signal the presence of a major claim. However, since this sentence does not include any content relevant information, the referenced statement should be annotated.

> **Example 2.2.8** (Introduction): "And finally medicine with the help of technology has developed its most extreme product, the human being! Some say that it is a big step towards immortality. However, I strongly disagree with this view because of the following reasons."

The introduction in 2.2.8 illustrates a complicated example. The author disagrees with a given statement. Just annotating the statement would be wrong because the author's stance is actually the opposite. The last sentence includes nothing content relevant. Therefore, we are not able to annotate a major claim. If the last to sentence would be incorporated in one single sentence we could annotate the whole sentence. However, these annotation guidelines does not allow to annotate several sentence as an argument component. So, we should search the major claim in the conclusion and if it is also not present, we will not annotate a major claim.

Note that in some essays the introduction or conclusion also includes reasons which support the major claim. In these cases it might be complicated to distinguish between major claims and the supporting statements which should be annotated as claims (cf. Section D.4). Example 2.2.9 illustrates such a case:

> **Example 2.2.9** (Introduction): "... . I believe that [**we should invest in cloning technology**]. Cloning helps to develop new cures for lethal diseases. Cloning of animals and organs can be beneficial for humankind. It could also foster research in biomedicine which would be an important step towards new technologies."

In this example only the first sentence contains the major claim, since it represents the stance of the author towards the topic (in this case "*Cloning*"). The second, third and fourth sentences provide several reasons to support the author's stance. Therefore these sentences (2-4) do not include major claims but reasons which should be annotated as claims.

D.4 Annotation of Claims

A claim is a direct support (or refutation) of the author's stance. So, it is a direct reason given in support (or attack in the case of a contra argument) of the major claim. In body paragraphs[9], a claim is usually supported with one or several reasons/premises whereas in the introduction or conclusion, a claim appears as a direct reason of the major claim. Commonly, the claim is an assumption that should not be accepted without additional support. Since the characteristic of claims in body paragraphs might differ from claims in introductions or conclusions, we distinguish these two cases in our guidelines. However, in both cases, each claim has a *stance attribute* which denotes if the claim is "*for*" or "*against*" the major claim. We will illustrate this attribute in the examples given in the following sections.

Claims in Body Paragraphs

A claim in an body paragraph is the central component of an argument. It appears frequently as an *initial assumption* located at the beginning of a paragraph or as a *conclusion* near the end. In few cases, the claim might also be located somewhere between the statements of a paragraph. Most frequently, one paragraph includes a single unique claim and there are only few cases where several claims/arguments are included in a single body paragraph. In this case, a paragraph includes several arguments covering different topics or aspects related to the topic.
Commonly, a claim does not appear without reasons (premises) in an body paragraph. So, for annotating claims in body paragraphs, it might help to identify which statements are reasons for others. If there is one statement which is not a reason for another statement in the paragraph, it is likely the claim of an argument. During the annotation process it might also help to be aware of the major claim, since the claims are direct reasons for the stance of the author. So, it is likely that claims in body paragraphs share some entities with the major claim (e.g. locations, persons or general noun phrases).
For getting familiar with the annotation of claims in body paragraphs, we will investigate some examples. In each example, the major claim of the essay is included

[9] paragraphs between introductions and conclusion

since it might help to identify the claims in body paragraphs. In each example the claim/s of the paragraph is/are in square brackets and underlined.

> **Example 2.3.1.1**:
> Major Claim: "Cloning is an important technology for humankind"
> Paragraph: "[The technology of cloning can be helpful for developing new cures]. For example, by reproducing organs like kidneys or levers, many people with serious diseases can be healed. In addition, it might be helpful for understanding other important processes in the area of gene technology which can help to invent new kinds of treatments."

In this example, the claim is given as the first statement in the paragraph. It is an initial assumption and a direct support of the major claim. The two following sentences are reasons given for the claim. These reasons are the premises which support the claim. Since the first statement does not serve as a reason for another statement (there are no outgoing support relations, cf. Section D.6), it is likely to be the claim. Also note that the rules for argument component boundaries hold for claims. The stance attribute of the claim in this example should be set to "*for*" since it supports the major claim.

> **Example 2.3.1.2**:
> Major Claim: "It is dangerous to clone humans"
> Paragraph: "The consequences of cloning humans are incalculable since no scientist has ever cloned a human being. It might cause terrible consequences and uncontrollable changes in the human gene pool. Therefore, [human cloning should be prohibited]."

In this example, there are two candidates which could serve as a premise. The first part of the first sentence seems to be a direct support for the major claim. However, this statement is a support for the last sentence, so it cannot be the claim of this paragraph. The last sentence however starts with the indicator "*therefore*" which is a strong signal for the presence of a claim. In addition, the first statement is a reason given why human cloning should be prohibited. The second sentence also seems to be a good reason for this statement. So the last sentence does not serve as a reason for another statement in this paragraph; it seems to be the conclusion of all the reasons given and therefore it should be annotated as the claim. The stance of this claim is again "*for*" because it supports the major claim.

> **Example 2.3.1.3**:
> Major Claim: "Some museums will not disappear"
> Paragraph: "Admittedly, [it is more convenient to learn about historical or art items online]. With Internet, people do not need to travel long distances to have a real look at a painting or a sculpture, which probably takes a lot of time and travel fees."

The third example includes a contra argument against the major claim. So the stance of the claim in the first sentence is set to "*against*". The second statement supports this contra claim; therefore, it is a premise given and the first statement is the claim of this argument.

Example 2.3.1.4:
Major Claim: "Technology negatively influences the way how people communicate"
Paragraph: "[Some people use their cellphone everywhere and do not even notice their environment]. Furthermore, [the language changed due to this new technology]."

Example 2.3.1.4 is a very infrequent case. There are two reasons given for the major claim. The first is about the overuse of cell phones and the second about the influence of technology on language. These two topics are not related and neither does the first statement support the second nor the second the first. However, since both are reasons for the major claim, the two statements are annotated as claims. The stance attribute of both claims should be set to "*for*" since both support the major claim. Note again that this case, occurs relatively infrequently in persuasive essays.

Example 2.3.1.5:
Major Claim: "School uniforms should be mandatory in each school."
Paragraph: "First, by wearing the same clothes, students learn to judge people without looking at their appearance and expensive brands of clothing. [School uniforms will decrease bullying which is very common in todays' schools]. If someone looks richer, most people feel like they have a higher social status or more power."

In this example the claim is located in the middle of the paragraph. It is a direct reason given for the major claim and the first and last sentence include some reasons which support the claim. Also note, that in this example no indicators are present. For identifying the claim in this case, it is necessary to recognize how the statements support each other. The stance is again "*for*".

Example 2.3.1.6:
Major Claim: "Cloning is a new technology that is necessary for our world"
Paragraph: "First, [cloning organs is useful for the treatment of lethal diseases]. Thanks to organ transplantation by cloning, people may be healthier and happier. Furthermore, [cloning animals enables developments in science] because more animals can be used for experiments. This is the useful side of cloning."

This example includes two different arguments. This is denoted by two different aspects of the topic. The first argument is about healing diseases by using cloning and the second aspect is about novel (more general) developments in science. So this example includes two different arguments which both support the major claim. Therefore, the stance attribute of each claim is "*for*".

Example 2.3.1.7:
Major Claim: "Cloning is a threat for our society."
Paragraph: "By cloning humans there would be a split in the society.

> Clones would have extraordinary abilities which are a result of improving their human genes. For instance, genes could be manipulated in such a way that this novel human generation is immune against common diseases. Consequently, [it would be difficult for clones to integrate well in today's society]."

In this example, the first and the last sentence seem to be good candidates for the claim. Both are similar, the given reasons support both statements and both seem to be good reasons for the major claim. However, on closer inspection the statement in the first sentence emerges as a reason for the last statement. In addition, the last statement includes the indicator "*consequently*" which is a strong indicator for a conclusion. Therefore, the last statement should be annotated as a claim. It's stance is "*for*" since it supports the major claim.

Example 2.3.1.2 and example 2.3.1.7 showed that indicators like discourse connectives can facilitate the identification of claims. Appendix D.7 includes a list of claim indicators which frequently signal the presence of claims. However, these indicators are not a warrantor for the presence of a claim. There are some cases in which such an indicator is used for a preliminary 'result' e.g. in a reasoning chain. So, even if one or several of these indicators is/are present in a paragraph, it is necessary to understand the content and to recognize which statements support (or attack) each other. As mentioned above, the major claim should be also considered when searching for a claim in paragraph.

Claims in Introductions and Conclusions

In the introduction or conclusion, a claim appears as a direct reason (or refutation) of the major claim. In many cases, the claims are located adjacent to the major claim. In contrast to body paragraphs, the introduction or conclusion infrequently includes complete arguments including premises and often the claims are reformulations of the arguments included in body paragraphs (for instance it is likely that a conclusion contains a condensed version of the key claims given previously). The following examples illustrate claims in introductions and conclusions. In these examples the major claim is in bold face (if present) and the claims are underlined. Both are also put in square brackets to illustrate the boundaries more precisely.

> **Example 2.3.2.1** (Introduction): "Do you want a twin that is cloned from you? Do you think it is necessary? Or do you think it is unethical and should be banned? I strongly believe that [**cloning is a new technology that is necessary for our world**]. I have various reasons for this: [it is necessary for the treatment of some illnesses such as leukemia and it provides our children better lives]."

This example starts with three rhetorical questions which introduce and clarify the debatable character of the topic. The following sentence includes a very precise standpoint of the author followed by reasons given as support for it. These reasons directly support the major claim. Therefore, they are annotated as claims with the stance attribute set to "*for*". Since the last sentence includes an enumeration

of several direct reasons for the major claim it is not split into several argument components.

> **Example 2.3.2.2** (Conclusion): "To sum up, although [cloning humans might bear some risks], I strongly believe that [**this technology is beneficial for humanity**]. [It is likely that this technology bears some important cures which will significantly improve the life conditions]."

This example of a conclusion includes two claims surrounding the major claim. The first claim is a rebuttal which illustrates some risks of cloning. It is a statement which directly attacks the major claim or the author's stance respectively, so the stance attribute for this claim should be set to "*against*". The last sentence includes a reason which supports the major claim. Therefore, it is annotated as a claim and the stance attribute of it is set to "*for*".

> **Example 2.3.2.3** (Introduction): "As the way to cloning has been found, there has been an argument about if it is right or wrong to clone a human. Some people argue for and others against cloning, but we cannot reach an agreement because there is no evidence supporting either side. However, as far as I am concerned, [**the disadvantages of cloning outweigh the advantages**] because [the consequence of these experiments are not foreseeable]."

The third example illustrates an introduction where the last sentence includes the major claim. In the same sentence there is a reason given which supports the major claim. Since the reason can be separated as a statement (cf. rules for argumentation boundaries in Section D.2), it is annotate as a claim. Its stance attribute is set to "*for*".

> **Example 2.3.2.4** (Introduction): "And finally medicine with the help of technology has developed its most extreme product, the human being! Some say that it is a big step towards immortality, while some claim that it is something unnatural. Since [it us unnatural and unethical], [**I definitely disagree with the idea of human cloning**]."

In this example, there a reason precedes the major claim. Since it can be separated using the guidelines for argument component boundaries, it is annotated as a supporting claim for the major claim. It's stance attribute is set to "*for*".

> **Example 2.3.2.5** (Conclusion): "To conclude, although [school uniforms are expensive], I think [**they should me mandatory in each school**]. [They will prevent bullying] because by wearing the same clothes pupils do not judge their classmates by their appearance and brand-name cloths."

The last example illustrates an argument following the major claim in a conclusion. The last sentences includes two statements. The first statement is a direct reason given for the major claim. Therefore, it is annotated as a claim (stance attribute is "*for*") followed by a premise in the same sentence. In addition, a contra claim precedes the major claim. It is a contra reason and therefore its stance attribute is set to "*against*".

D.5 Annotation of Premises

In this step, we focus on the annotation of the second argument component: the premise. A premise is a reason given for supporting or attacking an argument component. So it can be considered as a justification or refutation for convincing the reader of the truth or falsity of a claim. This also means that a premise is always connected to another argument component which could either be a claim or another premise if there are reasoning chains included in a paragraph. Note that in contrast to the claim annotation, the stance of each premise is not encoded in an attribute. It will be annotated by support or attack relations which link the premises to claims or other premises. For example the statement "*Children bring happiness and meaning to your life*" is a supporting premise whereas "*It is a heavy psychological burden to have children*" is an attacking premise for the claim "*Having children is the ultimate bliss in our lives*". In this case both should be annotated as premises, the distinction between supporting and attacking premises by means of argumentative relations is described in Chapter D.6.

Since the context and the identified claims from the previous step are important for annotating premises, the annotator should search for each claim in a paragraph and find the reasons given for it. It is possible that a claim and a premise are included in a single sentence or that a premise is only a part of a sentence. So the annotation of premises is also conducted at the clause level and the rules for argument components should be followed.

Usually, there are several premises given for a single claim in a paragraph. Sometimes there are also reasoning chains where several premises are linked together for supporting a claim. The following examples illustrate the annotation of premises in detail (the premises are wavy underlined).

> **Example 2.4.1** (Body paragraph): "First, [cloning can help human families to gain children]. For instance, [parents with no eggs or sperms can create children that are genetically related]. [Even same sex couples can have children without the use of donor sperm or donor eggs]."

In this example, the first statement is a claim which is supported by two premises. Both of the premises are indented to convince the reader that cloning is a positive development and that there are particular families which will benefit from this technology. Note that this example includes two argumentative relations connecting the two premises to the claim. Both of these relations indicate that the source argument components (the two premises) are reasons given for the target component (the claim). Having this structure in mind also helps to identify the claim. As mentioned in the introduction, the argumentation structure is always a tree and the root node of a tree is always a claim. So, for annotating argument components, it is helpful to imaging the argumentation structure and to recognize which statement supports or attacks another one and vice versa. In this example, it is only possible to connect the argument component two and three to the first one. So the first component, which is the root of the argumentation structure is likely to be the claim of the paragraph.

> **Example 2.4.2** (Body paragraph): "[There would be a lack of uniqueness and violate convictions regarding human individuality and freedom]. So,

> [clones could be seen as less than human compared with non-clones]. Therefore, [human cloning would divide our society into two different groups]."

This example illustrates a reasoning chain including one claim at the end of the paragraph and two preceding premises. The first premise is a reason given for the second premises which is a reason given for the claim. The chain of reasoning is indicated by indicators. Both indicators, "*so*" and "*therefore*", signal a conclusion based on preceding statement(s). However, since all statements form a chain, only the last statement (or the root of the tree) is annotated as a claim.

> **Example 2.4.3** (Body paragraph): "[Having children is the ultimate bliss in our lives]. There are many reasons which support my viewpoint. For example, [a close friend of mine is so happy since her child was born]. [She says that raising a child is like having an important goal in life] since [bringing up a human being is more satisfactory than anything else]."

This example illustrates a paragraph where a sentence does not include argumentative content. The second sentence only states that there are reasons without stating them. So, it is not annotated as a premise. The third sentence includes an example referring to a close friend. It is some kind of evidence indented to support the claim in the first sentence. Therefore, it is annotated as premise. The fourth sentence includes two premises which are connected by means of the discourse connective "*since*". Both are annotated as premises because the sentence includes an inference (cf. Section D.2).

> **Example 2.4.4** (Conclusion): "To conclude, although [having children could be considered as a financial burden], I think that [**it is the ultimate goal**]. [Nobody will regret to have children] since [having children brings happiness and meaning to our lives]."

Example 2.4.4 illustrates a conclusion including a premise. Generally, this case is quite infrequent. The conclusion begins with a claim against the major claim followed by another claim supporting the major claim in the last sentence. Since the last sentence includes two statements and the second statement supports the preceding one, the second statement is annotated as a premise for the preceding claim. Note again, that direct reasons given for the major claim should be annotated as claims and not as premises. Only if there are additional reasons given for a claim in a conclusion, like in example 2.4.4, they should be annotated as a premise.

> **Example 2.4.5**
> MajorClaim: "Cloning is an important technology for humankind"
> Body paragraph: "Some people believe that [human cloning would divide our society into two groups]. They argue that [clones will be seen as less compared to naturally born humans] because [clones will lack uniqueness and individuality]. However, [a human clone will not differ that much from other humans]. [They could develop exactly the same abilities as naturally born humans]."

This example shows a more complicated argumentation structure. The paragraph starts with a contra-claim stating that human cloning has some negative effects on our society. This claim is followed by a reasoning chain of supporting premises in the second sentence. The third sentence is an attacking premise of the claim which is supported by the last statement in the last sentence. Note again that the determination between supporting and attacking premises is conducted by means of argumentative relations. In this example the premise "*a human clone will not differ that much from other humans*" will be linked with an attack relation to the claim. The author also adds an additional support to this premise in the last sentence which will be linked using a support relation.

> **Example 2.4.6** (Body paragraph): "If human cloning became possible what would be the outcome of it? Basically I think [human cloning is against the laws of nature]. [There is a well organized balance of nature and human cloning may damage this wonderful balance] since [the number of people would increase due to cloning]."

Example 2.4.6 begins with a rhetorical question. Generally, questions are non-argumentative since the answer to the question is not known. So it is not annotated as an argument component. The second sentence includes the claim of this paragraph. It is supported by the first statement in sentence 3. Since the last sentence includes an inference and two statements, both are annotated as a premise.

As for claims, premises are sometimes signaled by indicators which facilitate the recognition. For instance, indicators like "*because*", "*the reason for this is*", "*in addition*" may signal the presence of a premise. A list of indicators for premises is provided in appendix D.8. Note again, that these indicators are not a guarantor for the presence of premises. It is necessary to understand the complete argumentation structure and to recognize which component supports or attack another one. For instance there might be cases where in a reasoning chain a claim indicator is used by the author for emphasizing an "intermediate result". Such an example is shown in 2.4.2. The indicator "*So*" at the beginning of the second sentence might indicate a claim. However, since the argument component following this indicator is only an intermediate result in a reasoning chain this statement is not annotated as the claim of this paragraph.

A particular characteristics of indicators is that they signal the direction of reasoning. For instance, indicators like "*because*", "*since*", "*in addition*", "*for example*" or "*first*"[10] signal that the following statement refers to a preceding statement whereas indicators like "*therefore*", "*as a result*", "*hence*" or "*thus*" signal that the following statement is a "result" of preceding statements. So, if an indicator is present in a paragraph which signals that the following argument component i is a "result" of preceding statements, the statement i is a good claim candidate. The same holds for indicators which signal that a statement refers to a preceding component. In this

[10] Indicators which signal enumerations like "*first*", "*second*", "*in addition*", "*furthermore*", etc. indeed indicate that the following statement refers to a preceding one. However, frequently these indicators are also used to enumerate the arguments in an essay and therefore the claims. In this case, a paragraph might start which an enumeration indicator followed by a direct reason for the major claim which is frequently the claim of a paragraph. If enumeration indicators are present inside of an paragraph it is likely that they enumerate the premises for a claim.

case the signaled component is likely to be a premise and not a claim. Recognizing the direction of the reasoning in a paragraph by means of indicators will strongly facilitate the annotation of argument components. So, before annotating any argument component in a paragraph, it might help recognize the direction of reasoning by means of the indicators (e.g. discourse connectives or shell language).

D.6 Argumentative Relations

Having annotated the argument components, we link the argument components with argumentative relations in order to build the tree structure of each argument. An argumentative relation is a directed link between two argument components with a particular source and target component. Such a relation either indicates that the source component is a justification (support relation) or a refutation (attack relation) for the target component. Since argumentative relations between claims and major claims are implicitly encoded in the stance attribute of the claim and we assume that those relations are the only ones which cross paragraph boundaries, we focus on the annotation of argumentative relations in paragraphs only. Therefore, each source component of an argumentative relation is a premise. However, the target is not restricted to be a claim, since there might be deeper tree structures including serial support. In other words, by annotating argumentative relations we identify for each premise the target it belongs to and recognize if the premise supports or attacks the target. Note that the target can either be a claim or another premise.

The following process illustrates the annotation process. For each paragraph including claims and premises the argumentation structure is build using the following steps:

1. Select a claim c
2. Link each premises in the paragraph if it obviously supports or attacks the claim.
3. For all not connected premises in the paragraph, test if it could be connected to an already connected premise. If that is not possible reformulate the premise and connect it to a matching claim or premise in the same paragraph

Following this process, ensures that the argumentation structure is a tree and that each premise is linked to exactly one argument component (either claim or premise). Also note, that in some cases, the support (or attack) of a single premise might be weak and not obvious. However, when combined with another premise the reason might become stronger. This happens for example if the author uses a particular event or an example in order to justify or refute a standpoint. So sometimes it might be necessary to consider premises in combination in order to identify the correct target of each individual premise. In the next sections, we will illustrate the annotation of support an attack relations in detail by providing several examples.

Annotating Support Relations

A support relation between two argument components indicates that the source component is a reason or a justification of the target relation. In the following

examples, the claim is underlined, the premise is wavy underlined and the argument components are enumerated using superscripts. Let's consider the following simple example:

> **Example 3.1.1** (Simple Example): "[An advanced gun background check should become routine in all gun sales]$_1$ because [it will prevent gun rampages]$_2$."

In this example there are only two argument components of which the first one is the claim and the second one is the premise. The indicator "*because*" signals that the second component is a justification for the first one. Therefore, the second argument component should be linked to the first one using a support relation.

> **Example 3.1.2** (Body paragraph): "First of all, [people cannot predict their own future or know what will happen tomorrow]$_1$. [The world is full of disasters such as wars, pollution, famine, drought, starvation, natural disasters and diseases]$_2$. So [it is just a big mistake to have children]$_3$."

Example 3.1.2 includes three argument components. The claim of this paragraph is present in the last sentence. Following our procedure, we first check if the first premise is a reason for the claim. Since the statement "no one can predict the own future" might be a reason for not having children, the first premise is linked with a support relation to the claim. The second premise seems also to be reason for the claim. Therefore, it is also linked using a support relation.

> **Example 3.1.3** (Body paragraph): "Furthermore, [it is a very heavy psychological and physical burden to have children]$_1$. [A mother carries her baby in her womb for nine months and 10 days and then the baby torments her during and after the birth]$_2$. [There is no peace, no silence or no sleep at home]$_3$. On the other hand, [the father has to work hard and earn more money]$_4$ because [the baby comes with his expenses]$_5$."

Example 3.1.3 includes five argument components. The first components is the claim. Argument component 2 which is a premise seems to be a reason, thus it is linked to the claim with a support relation. Also the argument component 3 is a reason which supports the claim. It is also linked to the claim using a support relation. Identifying the target for the argument component 4 and 5 is more complex. Argument component 4 can be considered as a reason for the claim. Although argument component 5 seems to be a reason, it should be linked to argument component 4, since both 4 and 5 are included in the same sentence and the author indicates that argument component 5 is a reason for argument component 4 by using the indicator "because". The complete structure is illustrated in Figure D.5.

> **Example 3.1.4** (Body paragraph): "[Having children is the ultimate bliss in our lives]$_1$. There are many reasons which support my viewpoint. For example, [a close friend of mine is so happy since her child was born]$_2$. [She says that raising a child is like having an important goal in life]$_3$ since [bringing up a human being is more satisfactory than anything else]$_4$."

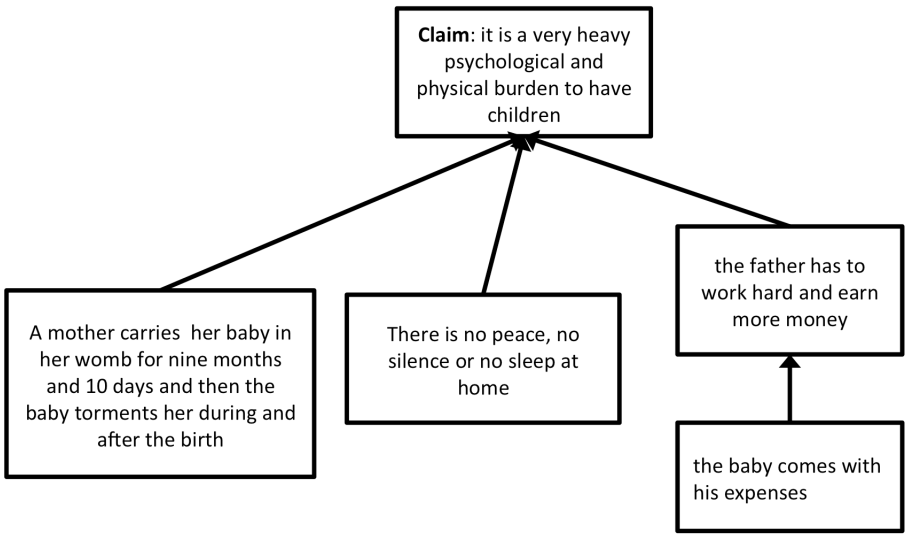


Figure D.5: Argumentation Structure of Example 3.1.3. All edges are support relations.

In example 3.1.4 the first argument component is a claim. Following our procedure, we first check if the first premise is a support for our claim and link it though it is only a weak support. However, the author uses the example of a close friend as evidence it is fine to consider it as a support. Argument component also seems to be a reason for the claim. Therefore it is also linked to the claim. Since argument component 4 is signaled by an indicator to be a support for argument component 3, it is linked to 3 not to the claim though it seems to be a good reason. However, by using the indicator the author explicitly linked it to argument component 3 and not the claim.

> **Example 3.1.5** (Body paragraph): "[Humankind will benefit from cloning]$_1$. For example, [cloning technology can be used to clone organs]$_2$. [It allows to raise new kidneys, levers and other vital organs]$_3$. Thus, [patients in need of organ transplantation will definitely benefit from cloning technology]$_4$. In addition, [cloning can help human families to gain children]$_5$. [It will allow parents with no eggs or sperms to create children that are genetically related]$_6$."

Example 3.1.5 is a more complex examples including six argument components and one claim. The first argument component is the claim in this paragraph. It is a general statement about cloning. The remaining argument components are premises and there are two different aspects included which are represented in two different branches of the argumentation tree. The first, brach is about cloning for raising organs whereas the second is about cloning for create children. In this example, it is difficult to follow the process described above since it includes a relatively deep argumentation structure. So we will first separate the two aspects and branches of the tree respectively. The first aspect includes premises 2, 3, and 4. All are about cloning in order to raise organs. And this aspect support the claim of this paragraph

in argument component 1. If we consider those three argument components in isolation, it seems that the component 4 is a subclaim since it is signaled with "thus" which indicates a partial result of the reasoning. Therefore, we consider it as the root node of the first branch although 2 and 4 seem to be reasons for the claim in 1. However, since there are two separated aspects which are given by the author for the author, we will model both as different branches in our tree. Argument components 5 and 6 represent the second branch and aspect respectively. Argument component 6 is given as a reason for argument component 5. So we link argument component 5 to the claim and component 6 to argument component 5. The resulting structure is illustrated in Figure D.6

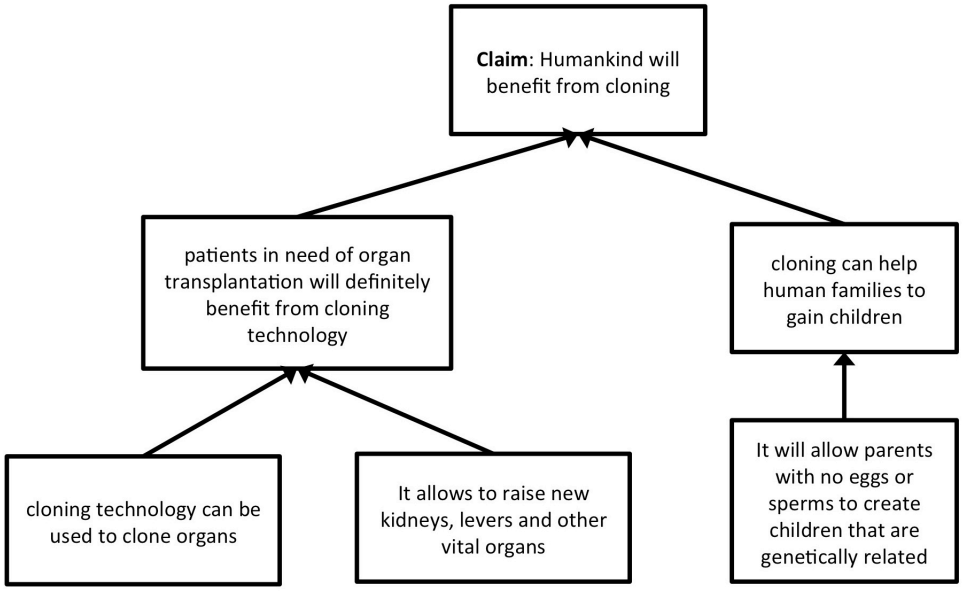


Figure D.6: Argumentation Structure of Example 3.1.5. All edges are support relations.

> **Example 3.1.6** (Body paragraph): "First, [cloning organs is useful for the treatment of lethal diseases]$_1$. [Thanks to organ transplantation by cloning, people may be healthier and happier]$_2$. Furthermore, [cloning animals enables developments in science]$_3$ because [more animals can be used for experiments]$_4$."

In example 3.1.6 two claims are present. The first one (argument component 1) is about cloning organs for developing novel cures and the second one (argument component 3) is about cloning animals. Following our procedure for argumentative relation identification, we select one of the two claims. Let's start with argument component 1. Obviously, the premise in argument component 2 is a good support and thus we link it to the first claim. Since the second premise in argument component 4 does not support the first claim, we continue with the second claim. The only premise which is not linked (argument component 4) seems to be a good support. Therefore, we link 4 to the claim in argument component 3.

Annotating Attack Relations

An attack relation between two argument components indicates that the source component is a refutation or a rebuttal of the target relation. Analog to the previous section, the claim is underlined, the premise is wavy underlined and the argument components are enumerated using superscripts. Let's consider the following simple example:

> **Example 3.2.1** (Simple Example): "[Having children is an incredible experience which everybody should do]$_1$. However, [it also comes along with a lot of responsibilities]$_2$."

This example illustrates a simple case of a attack relations. The claim in argument component is refuted by the premise present in the second component. Therefore, the second component should be linked to the claim using an attack relation. In this example the attack relation is signaled by the indicator "*However*" preceding the premise in the second argument component.

> **Example 3.2.2** (Body paragraph): "[Raising your own child is like having an important goal in your live]$_1$. Admittedly, [you will have great responsibilities and you also will have sleepiness nights]$_2$ but [these drawbacks will turn into a valuable experience when your kids become older]$_3$. Therefore, [Having children is the ultimate bliss in our lives]$_4$."

Example 3.2.2 includes 4 argument components of which the last one is the claim stating that having children is the ultimate bliss in our lives. The first argument component includes a reason for supporting this standpoint followed by a refutation or a doubt in argument component 2 which states that having children is related to having great responsibilities which is a negative point of having children. Therefore,

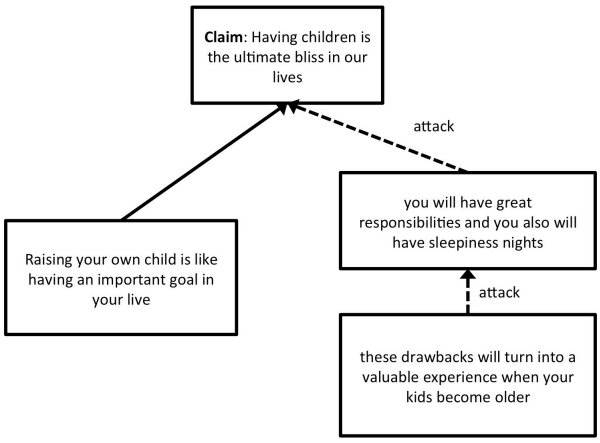


Figure D.7: Argumentation Structure of Example 3.2.2. Dashed arrows indicate attack relations and solid lines support relations.

argument component 2 is linked with an attack relation to the claim. In the same sentence, the author refutes this premise again which is signaled by the indicator

"*but*". Therefore, argument component 3 is linked with an attack relation to argument component 2. Note that this constellation is a common practice to prevent any potential criticism. By including in your argument a potential rebuttal (argument component 2) and stating why it is not relevant (argument component 3), the overall argument becomes stronger. Also note, that in this example, there are two good candidates for the claim, the first and the last component. However, since the last component is more general, signaled with the indicator "*therefore*", and the first argument component is a reason given for the last component, the last argument component is annotated as the claim. The structure of this argument is illustrated in Figure D.7.

Example 3.2.3
MajorClaim: "Cloning is an important technology for humankind"
Body paragraph: "Some people believe that [human cloning would divide our society into two groups]$_1$. They argue that [clones will be seen as less compared to naturally born humans]$_2$ because [clones will lack uniqueness and individuality]$_3$. However, [a human clone will not differ that much from other humans]$_4$. [They could develop exactly the same abilities as naturally born humans]$_5$."

Example 3.2.3 illustrates a common example of an opposing paragraph which might be the last body paragraph of an essay. It starts with a claim against the standpoint of the author followed by a supporting reason (argument component 2) of this contra position. This reason is again supported in the same sentence by the next argument

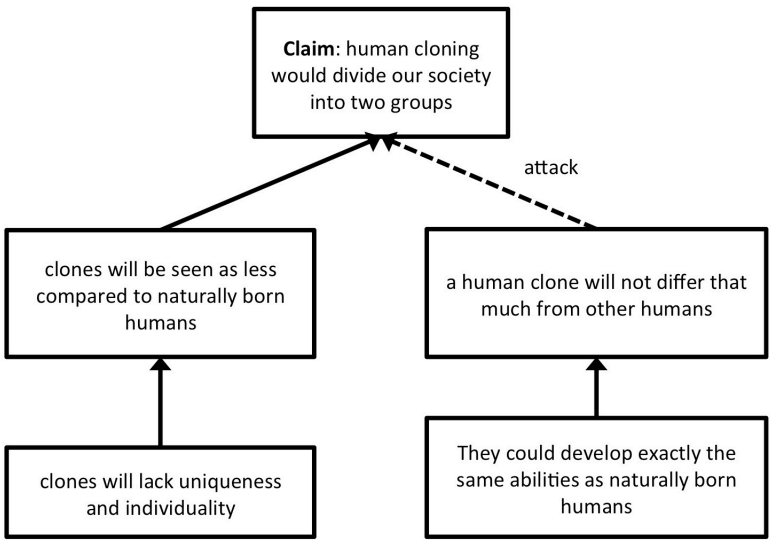


Figure D.8: Argumentation Structure of Example 3.2.3. Dashed arrows indicate attack relations and solid lines support relations.

component (argument component 3). In argument component 4 the author starts to attack the contra claim by putting forward a contra reason in argument component 4. Since this contra reason seems to be a refutation of the contra claim in argument

component 1, we link it using an attack relation. In argument component 5, the author provides another reason why his reason is true. So argument component 5 is linked to argument component 4 with an support relation. The whole structure of this argument is illustrated in Figure D.8

> **Example 3.2.4** (Conclusion): "To sum up, although [human cloning bears some risks for our society]$_1$, [the benefits of human cloning still outweigh its drawbacks]$_2$. Therefore, I strongly believe, that [**human cloning should be allowed in order to improve our medical system**]$_3$."

Example 3.2.4 is a conclusion including a major claim in argument component 3, a contra claim in argument component 1 and a contra premise in component 2. Since the first argument component directly attacks the major claim in argument component 3, it is annotated as a claim with its stance attribute set to "*against*". The following argument component is a reason given that the preceding contra claim is not true. Therefore, it is linked to to argument component 1 using an attack relation.

D.7 Claim Indicators

accordingly
as a result
consequently
conclude that
clearly
demonstrates that
entails
follows that
hence
implies
in short
in conclusion
indicates that
it follows that
it is highly probable that
it should be clear that
it should be clear
points to the conclusions
proves that
shows that
so
suggests that
the point I'm trying to make
therefore
thus
to sum up

we may deduce

D.8 Premise Indicators

assuming that
as
as indicated by
as shown
besides
because
deduced
derived from
due to
for
for example
for instance
for the reason that
furthermore
given that
in addition
in light of
in that
in view of
indicated by
is supported by
may be inferred
moreover
researchers found that
this can be seen from
since
since the evidence is
what's more
whereas

E Guidelines for Annotating Argumentation Flaws

E.1 Introduction

In this study we will annotate argumentation flaws in student essays. In particular, we will apply the argument evaluation framework proposed by Johnson and Blair (1994) which includes the following three criteria:

- *Relevance*: The relevance criterion addresses the connection between premises and the claim. In particular, a premise is only relevant if it counts in favor of the truth (or falsity) of the claim and thus provides some evidence for believing the claim to be true (or false).
- *Acceptability*: The acceptability criterion addresses the credibility of each statement in the argument. An argument violates the acceptability criterium if it includes unwarranted assumptions or factual flaws.
- *Sufficiency*: The sufficiency criterion addresses all reasons together. The premises are considered to be sufficient if they provide good grounds (considered together) for accepting the claim.

According to Govier (2010, p.87) an argument is *cogent* if it fulfills all of those criteria. An argument is fallacious if it violates one or more of these criteria.

In the following sections, we will first provide a brief introduction to argumentation (Section E.1) followed by the guidelines for identifying the standard form of arguments (Section E.1) and the annotation process (Section E.1). Chapters E.2-E.4 introduce the three criteria in detail and provide real examples of flawed arguments.

Argumentation in a nutshell

Argumentative practices are omnipresent in our daily verbal communication and thinking. We engage argumentation in order to infer certainty, to obtain widely accepted conclusions or to persuade a particular audience. In general, argumentation is a verbal activity of reason which aims at increasing or decreasing the acceptability of a controversial standpoint (van Eemeren et al., 1996, p. 5). Each *argument* involved in this process consists of several components. It includes a claim and one or more premises. The *claim* is a controversial statement and the central component of an argument. The *premises* constitute the reasons for believing the claim to be true or false (Damer, 2009, p. 14). An example of a basic argument including one claim and one premise is:

> Premise: "*Scientists demonstrated that cloning could be used to raise organs.*"
>
> ---
>
> Claim: "*Humankind will benefit from modern cloning technology.*"

The claim in this argument states what the author wants to convince the reader of whereas the premise is a reason given for supporting the claim. A general pattern for such a simple argument is "*<Claim> because <Premise>*". The next example illustrates a more complex argument including several premises and a particular structure:

Premise 1: "*Cloned organs match to the blood group and tissue of patients.*"
— SubPremise 1.1: "*Cloned stem cells of the patient can be used to raise organs.*"
Premise 2: "*Using cloned organs instead of donor organs shortens the healing process.*"
— SubPremise 2.1: "*Finding an appropriate donor is time consuming.*"

Claim: "*Cloning benefits the medical area of organ transplantation.*"

In contrast to the first example, the second example also includes sub premises intended to support the premises given for the claim. The general structure of each argument can be considered as a tree in which the claim (the central component of an argument) is the root node.

Both examples are given in its *standard form* in which all premises are present and the claims are clearly stated. However, an argument can be represented in text in many different ways and in real instances of arguments, some of the components may be missing, highly paraphrased, spread over several sentence or implicit. For instance, the previous argument can be represented in the following way:

> *"Cloning will be beneficial for people who are in need of organ transplants. Cloned organs will match perfectly to the blood group and tissue of patients since they can be raised from cloned stem cells of the patient. In addition, it shortens the healing process. Usually, it is very rare to find an appropriate organ donor and by using cloning in order to raise required organs the waiting time can be shortened tremendously."*

In this representation of the argument, the claim is clearly stated in the first sentence and it is clearly visible which premises are given for supporting the claim. Another lexical representation could be as follows:

> *"Many patients suffer from weak organs like livers, hearts or kidneys. Using stem cells would improve the quality and duration of the healing process tremendously. Finding an appropriate donor can take several years in particular cases."*

In this lexical representation the actual claim that cloning is beneficial for the medical field of organ transplantation is not present. Indeed the term cloning is not present at all. So the representation requires to know that stem cells could be cloned from the patient, that stem cells can be used to raise organs, etc. In comparison with the first lexical representation it is also hardly possible to recognize how the statements support each other and what the reasoning structure of the argument is. However, in order to evaluate the quality of an argument it is necessary to be aware of its standard form and to know what the author wants to convince us of. Therefore, we will first investigate several examples in order to recognize the claim and the standard form of an argument.

Identifying the Standard Form of Arguments

The standard form includes the premises (reasons) given for a claim in an ordered structure. Usually, the standard form lists all premises (and sub premises) before the claim (conclusions) analogue to the examples provided above. One way to identify the standard form of an argument is to first recognize what the author wants to convince the reader of (claim) before identifying the reasons given for supporting the claim. Fortunately, student essays exhibit a common structure which facilitates the identification of the standard form of the individual arguments.

A persuasive essay is usually written according to a *prompt* which outlines the *topic* and asks the author to develop a stance on the topic. The stance on the topic of the author is encoded in the major claim (thesis statement) in the introduction and restated in the conclusion. The actual arguments including claims and premises are present in the body paragraphs of the essay and either support or attack (contra arguments) the major claim. Being aware of the topic and the stance of the author on the topic facilitates the identification of the claim and the standard form respectively. So the first step in the annotation process is to read the prompt and the major claim(s) before identifying the claim and the premises of the actual argument. The following examples illustrate the identification of the standard form.

Example 1.2.1 (essay007_2):

Prompt: "*With the rise in popularity of the internet, newspapers will soon become a thing of the past. To what extent do you agree or disagree?*"

MajorClaim(s): "*newspapers have lost their competitive advantage to sustain their prolonged existence*"

Paragraph: "*Another point is that, from the economic aspect, buying newspapers appears to be a waste of money when the internet becomes available for every one. It is clear to recognize that the internet service is being provided at a low cost or even free in many countries. The question arises as to whether or not a person spends an extra money buying newspapers to receive the same, even usually less information than those he can have with the internet? The answer, perhaps, is that hardly would rational people do so. For this reason, the number of people reading newspapers may continue falling sharply, possibly leading to the close-downs of many in the coming time.*"

The prompt includes the topic of the essay which is in example 1.2.1: "*Will newspapers disappear because of the availability of the Internet?*". The major claim reveals the supporting stance of the author. In particular, the authors' stance/major claim is "*newspapers will disappear*". For supporting this stance, the author claims that "*the number of people reading newspapers may continue falling sharply*". As reasons/premises for supporting this claim she/he states that "*Spending money for newspapers is a waste of money*" because "*The Internet provides the same information as newspapers*" and since "*Internet is cheaper and even provided for free in many countries*". Note that this argument also implicitly assumes that "*newspaper providers cannot run their business if no one buys newspapers*". So the standard form for the argument is as follows:

Example 1.2.1 (standard form):

Premise1: "*Spending money for newspapers is a waste of money*"
Premise2: "*The internet provides the same information as newspapers*"
Premise3: "*Internet is cheaper and even provided for free in many countries.*"
Premise4: "*Newspaper providers cannot run their business if they do not sell newspapers*" (implicit)
Premise5: "*No one will spend money buying newspapers*"

Claim: "*The number of people reading newspapers may continue falling sharply possibly leading to a close-down of many newspaper providers.*"

Note that persuasive essay guidelines recommend to include the claim either in the first or last sentence of a paragraph. In addition, the claim is usually a direct reason

given for the major claim. So a good strategy to identify the claim is to check if the first or last sentence includes a direct reason for the major claim and if the remaining sentences provide reasons/premises for it.

The topic of the next example is "*cooperation or competition in primary education*". The major claims of the essay indicate the stance of the author which supports cooperation during primary education.

Example 1.2.2 (essay001_1):

Prompt: "*Some people think that children should be taught to compete, but others think that children should be taught to cooperate. What do you think?*"

MajorClaim(s):
(1) "*we should attach more importance to cooperation during primary education*"
(2) "*a more cooperative attitudes towards life is more profitable in one's success*"

Paragraph: "*First of all, through cooperation, children can learn about interpersonal skills which are significant in the future life of all students. What we acquired from team work is not only how to achieve the same goal with others but more importantly, how to get along with others. During the process of cooperation, children can learn about how to listen to opinions of others, how to communicate with others, how to think comprehensively, and even how to compromise with other team members when conflicts occurred. All of these skills help them to get on well with other people and will benefit them for the whole life.*"

In order to identify the claim of the argument, we check the first and last sentence since those are likely to include the claim. In this example, the first sentence includes the claim of the argument. In order to verify if we have found the correct claim, we can formulate the major claim as a question and check if our claim candidate answers the question. In our example, the question is "*Why should we attach more importance to cooperation during primary education?*" and the answer is "*Cooperation is important for the future life of children*". For identifying the reasons/premises of the claim we can again formulate the claim as a question like "*Why is cooperation important for the future life of children*" and the answer is that "*through cooperation children can learn interpersonal skills*". Consequently the standard form of the argument is as follows:

Example 1.2.2 (standard form):

Premise1: "*Cooperation fosters interpersonal skills (Children learn how to achieve the same goal with others, how to get along with others, listen to opinions, communicate with others, think comprehensively, compromise with other team members, etc.)*"
Premise2: "*Interpersonal skills will benefit them in their future life*"

Claim: "*Cooperation is important for children's future life*"

The topic of example 1.2.3 is again "*Will newspapers disappear because of the availability of the Internet?*" and the authors' stance is that "*newspapers will disappear.*

Example 1.2.3 (essay007_3):

Prompt: "*With the rise in popularity of the internet, newspapers will soon become a thing of the past. To what extent do you agree or disagree?*"

MajorClaim(s): "*newspapers have lost their competitive advantage to sustain their prolonged existence*"

Paragraph: "*Last, but not least, when taking environment into consideration, people must conceive that the more newspapers are published, the more trees are cut down. This is simply the contributor to the deforestation which is happening all over the world today. At this point, newspapers' production will have to face environmentalists on its way to be alive.*"

In this example neither the first sentence nor the last sentence include a clear claim. However, the whole paragraph covers a particular *aspect* namely "*environmental damage*". So the actual claim which is not explicitly mentioned in the paragraph is that "*newspaper production harms the environment*". The reasons/premises given by the author are that "*trees need to be cut down for producing newspapers*" which is indeed a particular kind of environmental damage. The argument also implicitly assumes that producing paper requires wood and consequently felling trees. The standard form of this argument is:

Example 1.2.3 (standard form):

Premise1: "*The more newspapers are published the more trees are cut down*"
Premise2: "*Deforestation (felling trees) harms the environment*"
Premise3: "*Producing paper requires wood*" (implicit)

Claim: "*Newspaper production harms the environment*"

This argument indicates that recognizing the claim might require to interpret the paragraph and to identify the general aspect of the argument. In order to test the identified aspect and claim respectively, we can again formulate the major claim as a question and check if the claim answers the question. In this example the question is "*Why will newspapers disappear?*" and the answer "*newspaper production harms the environment*" (claim).

Example 1.2.4 illustrates a more complicated example. The topic of the essay is: "*Is higher education or primary education more important for the development of a country?*". The authors' stance is that both play an important role for the development of a country (see major claim).

Example 1.2.4 (essay008_1):

Prompt: "*For successful development of a country, should a government focus its budget more on very young children education rather than universities?*"

MajorClaim(s):
(1) "*a government is supposed to offer sufficient financial support for both*
(2) "*a government should spare effort on young children education as well as universities*"

Paragraph: "*Concerning that elementary education, like the base of a architecture, is the fundamental requirement to be a qualified citizen in today's society, government should guarantee that all people have equal and convenient access to it. So a lack of well-established primary education goes hand in hand with a high rate of illiteracy, and this interplay seriously compromises a country's future development. In other words, if countries, especially the developing countries, are determined to take off, one of the key points governments should set on agenda is to educate more qualified future citizens through elementary education.*"

In the present argument the author argues about primary education only. The terms "*take-off*" and "*future development of a country*" suggest that the aspect of

the argument is economic development. So we can infer the claim that "*primary education is essential for economy*". As a reason the author states that a lack in primary education of citizens correlates with a high rate of illiteracy which in turn seriously harms the future development (economy) of a country. Following this interpretation the standard form of this argument is as follows:

Example 1.2.4 (standard form):

Premise1: "*A lack of primary education correlates with high rate of illiteracy.*"
Premise2: "*A high rate of illiteracy seriously compromises a country's future development.*"

Claim: "*Primary education is essential for the economical development of a country.*"

The topic of example 1.2.5 is "*studying abroad*". In particular, the prompt asks the student to present specific reasons why students study abroad. The author supports studying abroad and states in the major claim(s) that "*studying abroad has many advantages*".

Example 1.2.5 (essay006_2):

Prompt: "*Many students choose to attend schools or universities outside their home countries. Why do some students study abroad? Use specific reasons and details to explain your answer.*"

MajorClaim(s):
(1) "*studying abroad has many advantages*
(2) "*studying abroad does not only have advantages, but also can change us in a very positive way*"

Paragraph: "*One other important factor is the new academic experience that the students can obtain at the institution where they are pursuing their studies. For example, they will get exposed to a different educational system. They will meet new professors and new classmates which makes the academic experience different from that in their home country.*"

The claim of the argument is present in the first sentence of the paragraph. The author states that "*students gain novel academic experiences by studying abroad*". The author justifies the claim with two reasons. First, she/he states that "*students will be exposed to a different educational system*" which can be considered as a novel academic experience. The second reason is present in the last sentence of the paragraph and states that "*students will meet new professors and classmates*". Finally, the author states that these academic experiences are different from those in their home country. So the standard form of the argument is as follows:

Example 1.2.5 (standard form):

Premise1: "*Students will be exposed to a different educational system*"
Premise2: "*Students will meet new professors and classmates.*"
Premise3: "*These academic experiences are different from those in their home country*"

Claim: "*Students gain novel academic experiences by studying abroad*"

Previous examples illustrated the identification of the standard form of an argument. The procedure includes the following steps: (1) identification of the topic (prompt), (2) identification of the authors' stance (major claim), (3) Identification of the claim (frequently in the first or last sentence of the paragraph) and (4) Identification of the reasons. Remember that formulating the major claim or the claim as a question can facilitate the identification of the claim and/or premises.

Note that the identification of the standard form is an important prerequisite for assessing an argument since the standard form includes the argument in a structured form and clearly represents the connection between the given reasons and the claim. Therefore, we will first identify the standard form before assessing the argument. If necessary, please write down the standard form before assigning the evaluation criteria. So far, we have investigated arguments which fulfill all three criteria. The next chapters will introduce the particular criteria in detail.

Annotation Process

The process for analyzing the arguments in persuasive essays includes the following steps:

1. Identify the topic of the essay by carefully reading the prompt of the essay.
2. Identify the authors' stance on the topic by reading the major claim
3. Identify the claim in the paragraph (frequently in the first or last sentence)
4. Recognize the standard form by identifying the premises (Chapter E.1)
5. Assess the relevance (Chapter E.2)
6. Assess the acceptability (Chapter E.3)
7. Assess the sufficiency (Chapter E.4)

E.2 Relevance

In a good argument the premises need to count in favor of the truth (or falsity) of the claim. In other words, the given premises need to be relevant for the claim. The following example illustrates a relevance issue:

> Premise: "*Students earn a lot of experience during their stay in another university.*"
>
> ---
>
> Claim: "*Students who studied abroad will contribute more in their future jobs*"

The claim asserts that students who studied abroad contribute more than other students who didn't study in a foreign country. As reason for supporting this claim, the author states that students who studied in another university gained a lot of experience. However, more experience is not a guarantor for more commitment. So the premise does not counter in favor of the claim and consequently the claim does not follow from the premise. Therefore, the premise is not relevant to the truth of the claim. This particular type of relevance flaw is also called *non-sequitur* which translates to it-does-not-follow.

A premise also violates the relevance criterion (1) if it assumes the truth of the claim or (2) if it just paraphrases/reformulates the claim. This case is illustrated in the following example:

> Premise: "*Having children brings happiness to our life.*"
>
> ---
>
> Claim: "*Parents are so serendipitous with their kids.*"

The claim states that having children brings happiness to our life. The premise just reformulates the claim which cannot be considered as a reason for the claim. A better premise would be for example that raising children is like having an important goal in your life and that someone is more happy if she has a goal. This type of fallacy is called *begging the question* which is generally defined as an argument in which the premises presuppose the truth of the claim. It is also called *arguing in a circle*.

Having illustrated the basic notion of relevance issues, we will investigate real examples from student essays.

> **Example 2.1 (essay005_2)**:
>
> **Prompt**: "*The idea of going overseas for university study is an exciting prospect for many people. But while it may offer some advantages, it is probably better to stay home because of the difficulties a student inevitably encounters living and studying in a different culture. To what extent do you agree or disagree this statement? Give reasons for your answer.*"
>
> **MajorClaim(s)**:
> (1) "*one who studies overseas will gain many skills throughout this experience*
> (2) "*living and studying overseas gives the individual a new perspective on the subject that is studied or in general life*"
>
> **Paragraph**: "*Second, living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet. One who is living overseas will of course struggle with loneliness, living away from family and friends but those difficulties will turn into valuable experiences in the following steps of life. Moreover, the one will learn living without depending on anyone else.*"

The topic of example 2.1 is "*studying abroad*" and the authors' stance is "*for*" (major claim: students studying abroad gain many skills/experience). The claim of the argument is present in the first sentence and the aspect of the argument is "*standing on your own feet*". By reformulating the major claim as a question we can verify that the claim of the argument is "*By studying overseas students learn to stand on their own feet.*". The second sentence includes a *rebuttal* to make the argument stronger against any potential criticism. Note that we will not consider rebuttals in our assessment since those are usually only meant to prevent criticisms and not to support the actual claim. The last sentence includes a reason for the claim (this is also indicated by the term "*moreover*" which indicates an additional reason/premise). So the actual argument of the author is "$[$*By studying overseas students learn to stand on their own feet*$]_{claim}$ *because* $[$*the one will learn living without depending on anyone else*$]_{premise}$". Since "*standing on your own feet*" is just a reformulation of "*living without depending on anyone else*", the reason given is not relevant to the truth of the claim. So we will annotate the argument with a relevance flaw.

Example 2.2 is taken from an essay about "*influence of computer on children*" (topic). The author agrees that computers have a bad influence on children.

Example 2.2 (essay024_1):

Prompt: "*Using computer every day can have more negative than positive effects on your children? do you agree or disagree?*"

MajorClaim(s): "*It still has its bad side, especially for children*"

Paragraph: "*First, using computer constantly has bad influence on children's eyes. When they concentrate on computer for too long, their eyes will get tired, which is the main reason for some eyes problems, typically shortsighted.*"

The claim of the argument is present in the first sentence. It states that that "*Using computers has bad influence on children's eyes*". The next sentence elaborates that claim and does not provide any reason why the usage of computers harms the eyes of children. So it does not answer the question and thus cannot be relevant to the truth of the claim.

Example 2.3 illustrates that the identification of the standard form and the correct claim respectively is crucial for evaluating arguments. The topic of the essay is "*Will newspapers disappear because of the availability of the Internet?*" and the authors' stance is that "*newspapers will disapear*".

Example 2.3 (essay007_1):

Prompt: "*With the rise in popularity of the internet, newspapers will soon become a thing of the past. To what extent do you agree or disagree?*"

MajorClaim(s): "*newspapers have lost their competitive advantage to sustain their prolonged existence*"

Paragraph: "*First of all, to obtain information, using the internet is quicker and more convenient than reading newspapers. Contrary to the past when people had to wait long hours to take a daily newspaper, nowadays, they can acquire latest news updated every second through their mobile phones or computers connected to the internet, everywhere and at anytime. As can be seen, these devices and machines are very common in all parts of the world, making it easier for people to read a number of things that newspapers cannot provide in only some pages. Hence, the print media has failed to keep its important role in the provision of information.*"

Both, the first and last sentence include a potential claim which both provide a reason for the major claim. If we would select the first sentence as our claim, all premises given in the paragraph would be relevant. However, if we select the last sentence as our claim, the given reasons do not answer the question. So this example illustrates that carefully reading the argument and recognizing the correct standard form is crucial to obtain a good agreement among the annotators. This also includes a careful investigation of discourse markers like "*therefore*", "*for this reason*", "*consequently*" etc. In this particular example the last sentence is marked with the discourse marker "*hence*" which indicates that the author concludes the following statement from the previous ones. So in all cases in which the identification of the claim is ambiguous, we select the claim which includes a strong lexical indicator because it is likely that the indicator is included intentionally by the author. So in this argument, we select the last sentence as the claim and check if the given reasons are relevant to it. Consequently, the claim in this argument is "*print media*

(newspapers) has failed to keep its important role in the provision of information". As reason the author states, that the internet is more convenient, quicker and almost available everywhere. However, these reason do not justify that newspaper failed to provide information. Newspaper still provide information although there is a novel medium which is quicker, more convenient etc. Therefore, we annotate the argument with a relevance issue.

Example 2.4 illustrates a contra argument which attacks the stance of the author and the major claim respectively. The topic is "*Should students do physical exercises at school or focus on academic studies only*". The authors' stance is that "*physical education should not be outweighed*". However, the paragraph includes an argument supporting physical exercises. In these cases we just negate the major claim and evaluate the entire argument as described in previous examples.

Example 2.4 (essay011_1):

Prompt: "*Some people say that physical exercise should be a required part of every school day. Other people believe that students should spend the whole school day on academic studies. Which opinion do you agree with? Use specific reasons and details to support your answer.*"

MajorClaim(s):
(1) "*the physical education part should not be outweighed*"
(2) "*the more attention is withdrawn to a problem of physical education, the more influence we can get on academic success of students*"

Paragraph: "*First of all, the ancient Latin proverb says: "Anima sana in corpore sano", which means that healthy body is basis for healthy soul. There is a direct linkage between people's health and the frequency of doing physical exercises. The more we practice, the better we feel. Additionally, there is a fact that one can improve health conditions with everyday physical activity: gradually, step by step we change physical condition to better. So, if the entire nation, from their school ages begin going for sports and try doing that everyday, eventually it may help to build healthy nation.*"

The claim of this argument is visible in the entire paragraph. Each sentence actually states the same connection between physical exercises and health. So the claim is that "*Doing sport improves health conditions*". However, none of the sentences answers the question why regular sport activities improve health conditions although the paragraph includes some discourse connectors like "*additionally*" and "*so*" which denote a reasoning structure between the statements. Therefore, we annotate the argument with the relevance flaw[11].

Example 2.5 illustrates an example in which the entire paragraph is not relevant to the major claim. The topic of the essay is that "*email and text messaging is a threat for written language*" and the author agrees on it by stating that "*they (email and text messaging) decrease the position of written language*".

Example 2.5 (essay015_1):

Prompt: "*Email and text messaging have transformed communication but they are seriously threatening the status of written language. How far do you agree or disagree with this statement?*"

[11] Note that in contrast to an unwarranted assumption, this paragraph includes indeed a reasoning structure which is indicated by discourse connectors whereas an unwarranted assumption is a claim without any evidence.

MajorClaim(s):
(1) "*they may be factors decreasing the position of written language*"
(2) "*we cannot deny the threat from email and text messaging to the status of written language because their obvious popularity*"

Paragraph: "*First of all, it is easy to point out many benefits users can get from email and text messaging, which answers to the question of their great popularity. Before email and mobile phone, human beings communicated by meeting directly, sending letters or later, calling from home phones. Such ways usually made people have troubles for the expensiveness, difficulties in far communication or emergency and the loss of information. However, email and cell phones have improved the obstacles above. People can send or receive electronic letters anywhere and anytime they want. Especially, thanks to the function provided by email and text messaging (SMS), it is cheaper and faster to transfer a lot of information, even to many people at the same time. They are so convenient that the majority of population prefer to use them instead of some traditional ways.*"

The author mainly emphasizes the benefits of emails and text messaging. However, none of the statements in the paragraph refers to the resulting threat to written language. So the entire paragraph does not answer the question why emails and text messaging threatens written language and therefore the 'argument' cannot be considered to be relevant to the topic.

Example 2.6 is about the influence of television to the communication among friends and family members. The author agrees that television as a negative effect on the communication.

Example 2.6 (essay017_3):

Prompt: "*Do you agree or disagree with the following statement? Television has destroyed communication among friends and family. Use specific reasons and examples to support your opinions.*"

MajorClaim(s): "*television devastate families ties*"

Paragraph: "*Last but not least, some television programs have a negative effect on viewers. Some thriller or action movies advertise negative activities such as doing criminal activity. These genre of programs has more negative effects on our teenagers. They usually try to mimic what they see on movies in their daily behaviors.*"

However, the paragraph includes another aspect namely influence of television programs on teenagers which is not relevant to the communication.

Example 2.7 is about gender equality in universities. The author agrees that universities should allow equal opportunities for all genders.

Example 2.7 (essay018_2):

Prompt: "*Universities should accept equal numbers of male and female students in every subject. To what extent do you agree or disagree?*"

MajorClaim(s): "*universities must open choice for everyone to select his subject*"

Paragraph: "*However, the current job market affects the decision of universities. Every university designs its courses according to the need of employers. The job selection criteria are pre decided. For instance employers need male candidates for the heavy work such as repairing and installation of heavy machine. They prefer male staff for night shifts. They want female candidates for soft natured work like counseling, teaching, designing etc. Therefore, universities follow the requirement of job providers and decide subject suitable for particular gender.*"

The paragraph includes a contra argument with the claim that "*the job market affects the decision of universities*". As reasons the author states that universities design their courses according to the needs of employees which assumes the truth of the claim. Another reason employees prefer a particular gender for specific tasks which does not answer the question why universities prefer a particular gender for a specific subject. So the first reason is not relevant to the actual question (claim) and the other is just a reformulation. No given reason actually answers the question; thus the argument includes a relevance flaw.

E.3 Acceptability

The reason of a good argument should be based on indisputable common knowledge or facts because if the reasons given for supporting a claim are not acceptable, it is very unlikely that a particular audience will believe the claim to be true. So, we will assign the acceptability flaw to all arguments which reasons are false, highly questionable, unwarranted or if there is a good reason for not accepting the reasons. Note, that we will only assess the premises or claims which are unsupported (argument components which do not have any support). The following examples illustrate acceptability flaws.

The topic of example 3.1 is "*television destroys communication among friends and family members*" and the author agrees that "*television devastates family ties*".

> **Example 3.1 (essay017_1)**:
>
> **Prompt**: "*Do you agree or disagree with the following statement? Television has destroyed communication among friends and family. Use specific reasons and examples to support your opinions.*"
>
> **MajorClaim(s)**:
> (1) "*television devastate families ties*"
> (2) "*it is a global problem to solve this problem and it is need global contribution to find a solution*"
>
> **Paragraph**: "*In spite of enjoying watching television shows, it is really time consuming task. Emergence of new television channels or shows are not surprising. We kill time by tracking these channels so at the end of the day we understand it is late for doing any other activities. This is a problem that suffers all of us everyday. At the end, there is no time for call to our parents or friends.*"

The claim of the argument is present in the first and last sentence. It states that "*watching television is so time consuming that there is no time to communicate with parents and friends*". As reasons the author states that "*novel television channels and shows emerge frequently*" and that "*by tracking these programs there is no time for doing other activities (like communicating with friends and family members)*". So the author implicitly assumes that we (maybe everyone) assign a higher priority to watching television than to communicating with our friends and family members. This assumption is not warranted and relatively unlikely. So the argument is not acceptable.

Example 3.2 includes several aspects and also a rebuttal. The prompt asks the student if funding primary education or higher education is more important for the

development of a country. The author is of the opinion that both are important for the development of a country.

Example 3.2 (essay008_2):

Prompt: "*For successful development of a country, should a government focus its budget more on very young children education rather than universities?*"

MajorClaim(s):
(1) "*a government is supposed to offer sufficient financial support for both*"
(2) "*a government should spare effort on young children education as well as universities*"

Paragraph: "*That's not enough, for higher education, which represent the cutting edge of a country's academic achievements, act as the vanguard in economic and social advance. The high technology and new ideas applied into practice may not only lead a country to flourish but also elevate its status in the international community. Some may argue that universities can support themselves well by donation or invention, but this argument is invalid. Researches into humanities and art still need large amount of money, what's more, government's big budget on universities may attract more excellent intellectuals and researchers into the country and enjoy a higher reputation worldwide.*"

The paragraph includes several aspects and thus several arguments. In such a case we will evaluate each aspect/argument individually. The first two sentences are about the economic aspect. The claim is that "*higher education fosters economy*". A reason is indicated in the first half of the second sentence which states that "*technologies (which are a result of research in higher education) can be applied in to practice*" and that "*these technologies have an economic value*". Note that the second half of the sentence is related to another aspect/argument explained later. Since there is at least some kind of justification for the claim which seems to be reasonable, this aspect/argument is acceptable. The third and first half of the fourth sentence includes a rebuttal which we will not evaluate in this study. The second half of the last sentence states that "*government's big budget will attract researchers into the country and results into a higher reputation in the international community*". So the two aspects of this paragraph are economy and international reputation. The claim of the economy has a justification (admittedly it is a vague one). However, the second claim that "*funding universities will increase the reputation of country*" is not justified. There is no premise which supports the claim. Therefore, the paragraph includes an unwarranted assumption. A claim (direct reason for the major claim) which has no justification. We consider these unwarranted claims as not acceptable since students are asked to include evidence. Thus we annotate the entire paragraph as not acceptable.

Example 3.3 is about the question if governments should improve roads or public transportation. The authors stance is that governments should support public transportation.

Example 3.3 (essay010_1):

Prompt: "*Should governments spend more money on improving roads and highways or should governments spend more money on improving public transportation (buses, trains, subways) . Why? Use specific reasons and details to develop your essay*"

MajorClaim(s):
(1) "*governments spend more money on buses, trains, and subways investment*"
(2) "*public transportation systems even have more important advantages*"

Paragraph: "*First and foremost, carbon emission cut is significantly essential for protecting the atmosphere. The fact is that the more cars and motorbikes are on roads, the more seriously the ozone layer is damaged. If governments use more money to improve roads, there is a strong likelihood that more people drive their private cars work. This is sure to lead to more carbon emitted into the atmosphere, which can cause skin cancer and destroy the natural environment. Whereas, if there are more good buses, trains, or subways, people are inclined to use less private vehicles, which decreases the amount of carbon released. Obviously, the policy that concentrates money on developing public transportation brings an advantageous impact on earth.*"

The claim of this argument is that "*improving public transportation reduces carbon emission*". As reasons the author states that by improving roads, it will be likely that more people use their private cars which leads to more emission. The author also justifies her/his claim by stating that it is likely that more people will use public transportation when those are improved. So the claim is well justified. However, the paragraph includes a false statement namely that carbon emission causes skin cancer which is not true. Although carbon emission might damage the ozone layer which in turn leads to more dangerous sun rays whig indeed could cause skin cancer, the author states that the the carbon in the atmosphere can cause skin cancer. Since this causal linkage is not directly true, we cannot accept the argument and annotate it with the acceptability flaw. So besides unwarranted claims/aspects, we annotate an argument as not acceptable if it includes false statements.

The topic of example 3.4. is "*animal experiments*" and the author states that animal experiments are required to benefit humans. Note that the first major claim emphasizes the need of animal experiments with respect to particular aspect (testing food and medicine) and that the second major claim generally supports animal experiments. In such a case we will evaluate the relevance of the arguments according to the more general major claim and thus in this example according to the general supporting stance of the author about animal experiments.

Example 3.4 (essay016_1):

Prompt: "*Some people think it is acceptable to use animals for the benefit of the human beings. Some people think it is wrong to exploit animals for the human purposes. What is your opinion?*"

MajorClaim(s):
(1) "*it is a dramatically cruel activity to humanity if the latest foods or medicines are allowed to sold without testing on animals*"
(2) "*the merits of animal experiments still outweigh the demerits*"

Paragraph: "*First of all, as we all know, animals are friendly and vital for people, because if there are no animals in the world, the balance of nature will broke down, and we, human, will die out as well. The animal experiments accelerate the vanishing of some categories of animals. In other words, doing this various testing is a hazard of human's future and next generation.*"

The paragraph includes a contra argument against the stance of the author and thus attacking animal experiments. As in previous examples we will evaluate the argument by negating the major claim. The claim of this argument is present in the last sentence. It states that "*animal experiments is a hazard of human's future and next generations*". As reason the author states that "*animals are vital for humans because of the natural balance*" and that "*animal experiments accelerate the vanishing*

of particular animals". This argument has at least two acceptability problems. First, it assumes that humans will be in danger of extinction if animals disappear which is a debatable statement and not justified by the author. Second, the argument explicitly states that animals experiments accelerate the vanishing of some animals which is highly questionable since animal experiments are usually conducted with particular species. Therefore, the argument is not acceptable.

Example 3.5 is about the question if students should stay at home or live independently from their parents. The authors' stance is that young adults should live with their parents.

> **Example 3.5 (essay019_1)**:
>
> **Prompt**: "*Some young adults want independence from their parents as soon as possible. Other young adults prefer to live with their families for a longer time. Which of these situations do you think is better?*"
>
> **MajorClaim(s)**: "*for the young adults it will be better to live with their parents*"
>
> **Paragraph**: "*Living in their own houses will save the young adults a lot of money. If they live separate from their parents they will have to pay for a loan, electricity, water or even for a meal. If the young adults still study they will have to combine studying with working, because they will not have money to pay for everything. There will not be such worries when young adults live in their own home, because parents will take care for them. Moreover, parents will give their children money in order they to focus only on studying.*"

The claim of the argument is that staying at home is cheaper than living in an own accommodation. As reason the author states that living alone requires to pay electricity, water, etc. and that young adults need to find a job account for all these costs. Although these reason seem to be acceptable the argument assumes that parents will provide money which also explicitly states in the last sentence. Since this is an unwarranted assumption and might not be true in many cases, we will annotate the argument as not acceptable.

E.4 Sufficiency

An argument is a good one, if the reasons are sufficient for believing the claim to be true. However, there are arguments which infer a very general claim from one example only. The classical example is the following argument "*My neighbor has an academic degree and is the mayor of our town. Therefore, all mayors have an academic degree*". The very general claim is inferred from only one sample/instance. So the reason given as evidence is not sufficient for arriving at the very general claim that all mayors have an academic degree. The following examples illustrate this flaw in real arguments:

Example 4.1 illustrates an argument about the topic "*international tourism*". The author supports tourism by stating that "*tourism contributes to economy and preservation of culture and environment*" in the major claim(s).

> **Example 4.1 (essay003_1)**:
>
> **Prompt**: "*International tourism is now more common than ever before. Some feel that this is a positive trend, while others do not. What are your opinions on this?*"

MajorClaim(s): (1) "*it has contributed to the economic development as well as preserved the culture and environment of the tourist destinations*"
(2) "*international tourism has both triggered economic development and maintained cultural and environment values of the tourist countries*"

Paragraph: "*Firstly, international tourism promotes many aspects of the destination country's economy in order to serve various demands of tourists. Take Cambodia for example, a large number of visitors coming to visit the Angkowat ancient temple need services like restaurants, hotels, souvenir shops and other stores. These demands trigger related business in the surrounding settings which in turn create many jobs for local people improve infrastructure and living standard. Therefore tourism has clearly improved lives in the tourist country.*"

The claim of the argument is present in the last sentence of the paragraph. It states that "*tourism has improved life in the tourist country (economically)*". As reason the author cites one example which shows this improvement. However, only one example is not sufficient for inferring the general claim of the argument. Therefore, the reason given are not sufficient and we annotate the argument with the sufficiency flaw.

The prompt of example 4.2 asks the author to choose an important technological innovation and to provide reasons for the choice. The stance of the author is that "*technology (in general) helped us to have a more comfortable life*". Note that we will not evaluate if the claim is relevant to the topic.

Example 4.2 (essay012_2):

Prompt: "*Advance in transportation and communication like the airplane and the telephone have changed the way that nations interact with each other in a global society. Choose another technological innovation that you think is important. Give specific reason for your choice.*"

MajorClaim(s): "*technology has helped us to have more comfortable life*"

Paragraph: "*Another important aspect on technology is transferring money. Today, students can apply for foreign universities much easier than before. Not only with the help of sending email, but also using credit cards to pay all necessary fees online. Therefore, with the advent of internet and online paying systems, you can do many thing at your home easily.*"

The claim of the argument is that "*With the advent of the internet and online paying system many tasks became easier*". The author supports this claim by stating that students can apply easier for foreign universities and it is easier to pay the necessary fees with the help of these technologies. However, this is only one example for justifying the very general claim. So the given support is not sufficient for justifying the claim.

Example 4.3 is about the question if students should be taught to compete or cooperate. The author supports that students should be taught to cooperate.

Example 4.3 (essay001_2):

Prompt: "*Some people think that children should be taught to compete, but others think that children should be taught to co-operate that become more useful adults. What do you think?*"

MajorClaim(s):
(1) "*we should attach more importance to cooperation during primary education*"
(2) "*a more cooperative attitudes towards life is more profitable in one's success*"

Paragraph: "*On the other hand, the significance of competition is that how to become more excellence to gain the victory. Hence it is always said that competition makes the society more effective. However, when we consider about the question that how to win the game, we always find that we need the cooperation. The greater our goal is, the more competition we need. Take Olympic games which is a form of competition for instance, it is hard to imagine how an athlete could win the game without the training of his or her coach, and the help of other professional staffs such as the people who take care of his diet, and those who are in charge of the medical care. The winner is the athlete but the success belongs to the whole team. Therefore without the cooperation, there would be no victory of competition.*"

The claim of the argument is present in the last sentence (indicated by the term "*therefore*"). It is: "*cooperation is a requirement for being successful in competition*". The first two sentences include a rebuttal to make the argument stronger. Note again that we will not consider rebuttals in our assessment. As reason for supporting the claim, the author cites olympic games and that athletes require cooperation with team member to win in the competition. However, this reason represent only one example which is not enough for arriving at the very general claim. So we annotate the argument with the sufficiency flaw.

Example 4.4 (essay012_3):

Prompt: "*Advance in transportation and communication like the airplane and the telephone have changed the way that nations interact with each other in a global society. Choose another technological innovation that you think is important. Give specific reason for your choice.*"

MajorClaim(s): "*technology has helped us to have more comfortable life*"

Paragraph: "*Another technological innovations which help people around the world is related to medical equipments. Biomedical engineers could make a significant effect on increasing life expectancy the world. For example, one of their inventions was related to artificial heart valves which can be count as a turning point in heart surgeries. In the past time doctors used pig heart's valve to implant, but the patient could not be alive more than 3 years after the replacement. But now, biomedical engineers can make artificial heart valves which works well and doctors can implant them easily.*"

The next example (4.5) is again about tourism and the author supports tourism by stating that "*tourism contributes to economy and preservation of culture and environment*" in the major claim.

Example 4.5 (essay003_2):

Prompt: "*International tourism is now more common than ever before. Some feel that this is a positive trend, while others do not. What are your opinions on this?*"

MajorClaim(s):
(1) "*it has contributed to the economic development as well as preserved the culture and environment of the tourist destinations*"
(2) "*international tourism has both triggered economic development and maintained cultural and environment values of the tourist countries*"

Paragraph: "*Secondly, through tourism industry, many cultural values have been preserved and natural environments have been protected. For instance, in Vietnam, many cultural costumes and natural scenes, namely 'Trong Dong' drum performance and 'Ha Long' bay, are being encouraged to preserve and funded by the tourism ministry. Without this support and profit from tourism, many traditional cultures would disappear due to its low income works. Thus, tourism has survived many non-tangible cultural values and beauty scenes.*"

For supporting her/his stance the author claims that "*tourism has survived many non-tangible cultural values and beauty scenes*". As reason she/he cites a particular example from Vietnam which is not sufficient to infer this very general claim → sufficiency flaw.

E.5 Arguments with Several Flaws

This chapter includes examples of arguments with several flaws. For each argument, we provide the violated quality criteria.

Example 5.1 (essay012_1):

Prompt: "*Advance in transportation and communication like the airplane and the telephone have changed the way that nations interact with each other in a global society. Choose another technological innovation that you think is important. Give specific reason for your choice.*"

MajorClaim(s): "*technology has helped us to have more comfortable life*"

Paragraph: "*First and foremost, email can be count as one of the most beneficial results of modern technology. Many years ago, peoples had to pay a great deal of money to post their letters, and their payments were related to the weight of their letters or boxes, and many accidents may cause problem that the post could not be delivered. But nowadays, all people can take advantage of internet to have their own email free, and send their emails to everyone in no time, besides they can be sure if their emails have been delivered or not.*"

Relevance
Acceptability

Example 5.2 (essay013_1):

Prompt: "*Agree or disagree:Technology has made children less creative than they were in the past.*"

MajorClaim(s):
(1) "*technology makes children even more creative*"
(2) "*They kept researching new technology and became successful at a very young age*"

Paragraph: "*First, technology inspires children to create new things. Children are curious about everything around them, so when they come across a high-tech product like a cellphone, they will be obsessed with its mysterious functions and eager to know how it works. For example, Bill Gates was attracted by the original huge computer, then he did everything he could to understand how it worked. After he had figured out all the stuff, he then began to promote the computer. In the end, he successfully invented a computer that was easy to use and lightweight. Therefore, instead of interfering children's creativity, technology actually encourages children to learn and to create.*"

Sufficiency
Relevance

Example 5.3 (essay005_3):

Prompt: "*The idea of going overseas for university study is an exciting prospect for many people. But while it may offer some advantages, it is probably better to stay home because of the difficulties a student inevitably encounters living and studying in a different culture. To what extent do you agree or disagree this statement? Give reasons for your answer.*"

MajorClaim(s):
(1) "*one who studies overseas will gain many skills throughout this experience*"
(2) "*living and studying overseas gives the individual a new perspective on the subject that is studied or in general life*"

Paragraph: "*Also, employers are mostly looking for people who have international and language skills. Becoming successful in this study will give the student an edge in job market. Therefore, one who has studied and lived overseas will become more eligible for the job than his/her peers.*"

Relevance
Acceptability

Example 5.4 (essay013_1):

Prompt: "*Agree or disagree:Technology has made children less creative than they were in the past.*"

MajorClaim(s):
(1) "*technology makes children even more creative*" (2) "*They kept researching new technology and became successful at a very young age*"

Paragraph: "*First, technology inspires children to create new things. Children are curious about everything around them, so when they come across a high-tech product like a cellphone, they will be obsessed with its mysterious functions and eager to know how it works. For example, Bill Gates was attracted by the original huge computer, then he did everything he could to understand how it worked. After he had figured out all the stuff, he then began to promote the computer. In the end, he successfully invented a computer that was easy to use and lightweight. Therefore, instead of interfering children's creativity, technology actually encourages children to learn and to create.*"

Sufficiency
Relevance

Example 5.5 (essay013_2):

Prompt: "*Agree or disagree:Technology has made children less creative than they were in the past.*"

MajorClaim(s):
(1) "*technology makes children even more creative*"
(2) "*They kept researching new technology and became successful at a very young age*"

Paragraph: "*Second, technology widen children's knowledge. In the past, children were only able to see things from one perspective. However, with highly advanced technology, children are able to get information from foreign countries and even communicate with foreign friends. Therefore, they will be able to learn about different cultures and different ways of thinking. Knowledge is the base of creativity. The diverse knowledge that children gained from different parts of the world, inspires children and encourage them to create new things.*"

Relevance
Acceptability

Example 5.6 (essay014_2):

Prompt: "*Students at schools and universities learn far more from lessons with teachers than from others sources (such as the internet, television).To what extent do you agree or disagree?*"

MajorClaim(s): "*students learn far more from their teachers than from other source*"

Paragraph: "*Those who feel that students learn far more from other sources, such as the Internet and television, firmly believe that within this sources students learn lots of things which they can't learn in classes. They can only input some key words and google it, and then there are numberless articles and websites related to it. In this case, students learn things easily. Moreover, they contend that good television programs do teach students. For instance, Discovery Channel has many instructive episodes. Students have knowledge of others cultures, outer space etc.*"

Sufficiency
Acceptability

Example 5.7 (essay017_2):

Prompt: "*Do you agree or disagree with the following statement? Television has destroyed communication among friends and family. Use specific reasons and examples to support your opinions.*"

MajorClaim(s): "*television devastate families ties*"

Paragraph: "*Second, I think watching television programs makes us lazy. We usually spend our times in front of television for at least three ours a day when we come back home after eight-hour working in the office. Hence, we do not have time to spend for healthy activity such as going to a gym or doing other sports. This makes us a lazy person who prefers to do sedentary activities like watching television shows rather than be active and sociable.*"

Relevance
Acceptability

List of Figures

List of Tables

Bibliography

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*, pp. 64–68, Baltimore, MD, USA, 2014.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, 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*, pp. 1395–1404, San Diego, California, 2016.

Leila Amgoud, Nicolas Maudet, and Simon Parsons: 'Modelling Dialogues using Argumentation', in: *Proceedings of the 4th International Conference on Multi-Agent Systems*, pp. 31–38, Boston, MA, USA, 2000.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor: 'Cats Rule and Dogs Drool!: Classifying Stance in Online Debate', in: *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pp. 1–9, Portland, OR, USA, 2011.

Ron Artstein and Massimo Poesio: 'Inter-coder agreement for computational linguistics', *Computational Linguistics* 34 (4): 555–596, 2008.

Yigal Attali, Will Lewis, and Michael Steier: 'Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring', *Language Testing* 30 (1): 125–141, 2013.

Monroe C. Beardsley: *Practical Logic*, Prentice-Hall, 1950.

Beata Beigman Klebanov and Michael Flor: 'Argumentation-Relevant Metaphors in Test-Taker Essays', in: *Proceedings of the First Workshop on Metaphor in NLP*, pp. 11–20, Atlanta, GA, USA, 2013.

Beata Beigman Klebanov and Derrick Higgins: 'Measuring the use of factual information in test-taker essays', in: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pp. 63–72, Montreal, Quebec, Canada, 2012.

Bo Bennett: *Logically Fallacious*, eBookIt.com, 2012.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger: 'A taxonomy of argumentation models used for knowledge representation', *Artificial Intelligence Review* 33 (3): 211–259, 2010.

Jonathan Berant and Percy Liang: 'Imitation Learning of Agenda-based Semantic Parsers', *Transactions of the Association for Computational Linguistics* 3: 545–558, 2015.

Philippe Besnard and Anthony Hunter: *Elements of Argumentation*, MIT Press, 2008.

Or Biran and Owen Rambow: 'Identifying Justifications in Written Dialogs', in: *Fifth IEEE International Conference on Semantic Computing (ICSC)*, pp. 162–168, 2011a.

Or Biran and Owen Rambow: 'Identifying justifications in written dialogs by classifying text as argumentative', *International Journal of Semantic Computing* 05 (04): 363–381, 2011b.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow: 'TOEFL11: A corpus of non-native English', *ETS Research Report Series* 2013 (2): i–15, 2013.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič: 'Joint Morphological and Syntactic Analysis for Richly Inflected Languages', *Transactions of the Association for Computational Linguistics* 1: 415–428, 2013.

Filip Boltužić and Jan Šnajder: 'Back up your Stance: Recognizing Arguments in Online Discussions', in: *Proceedings of the First Workshop on Argumentation Mining*, pp. 49–58, Baltimore, MA, USA, 2014.

Simon Philip Botley: 'Argument structure in learner writing: a corpus-based analysis using argument mapping', *Kajian Malaysia* 32 (1): 45–77, 2014.

Chloé Braud and Pascal Denis: 'Combining Natural and Artificial Examples to Improve Implicit Discourse Relation Identification', in: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1694–1705, Dublin, Ireland, 2014.

Leo Breiman: 'Random Forests', *Machine Learning* 45 (1): 5–32, 2001.

M. Anne Britt and Aaron A. Larson: 'Constructing representations of arguments', *Journal of Memory and Language* 48 (4): 794 – 810, 2003.

Jill Burstein, Martin Chodorow, and Claudia Leacock: 'Automated Essay Evaluation: The Criterion Online Writing Service', *AI Magazine* 25 (3): 27–36, 2004.

Jill Burstein and Magdalena Wolska: 'Toward evaluation of writing style: finding overly repetitive word use in student essays', in: *Proceedings of the tenth conference of European chapter of the Association for Computational Linguistics*, EACL '03, pp. 35–42, Budapest, Hungary, 2003.

Jodie A. Butler and M. Anne Britt Britt: 'Investigating Instruction for Improving Revision of Argumentative Essays', *Written Communication* 28 (1): 70–96, 2011.

Elena Cabrio, Sara Tonelli, and Serena Villata: 'From Discourse Analysis to Argumentation Schemes and Back: Relations and Differences', in: *Computational Logic in Multi-Agent Systems*, Lecture Notes in Computer Science Vol. 8143, pp. 1–17, Springer Berlin Heidelberg, 2013.

Elena Cabrio and Serena Villata: 'Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions', in: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 208–212, Jeju Island, Korea, 2012b.

Elena Cabrio and Serena Villata: 'Natural language arguments: A combined approach', in: *Proceedings of the 20th European Conference on Artificial Intelligence*, ECAI '12, pp. 205–210, Montpellier, France, 2012a.

Elena Cabrio and Serena Villata: 'NoDE: A Benchmark of Natural Language Arguments', in: *Proceedings of COMMA*, pp. 449–450, 2014.

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*, pp. 1405–1413, San Diego, California, 2016.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski: 'Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory', in: *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, pp. 1–10, Aalborg, Denmark, 2001.

Richard Eckart de Castilho and Iryna Gurevych: 'A broad-coverage collection of portable NLP components for building shareable analysis pipelines', in Nancy Ide and Jens Grivolla (Eds.): *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pp. 1–11, Dublin, Ireland, 2014.

S. le Cessie and J.C. van Houwelingen: 'Ridge Estimators in Logistic Regression', *Applied Statistics* 41 (1): 191–201, 1992.

Kenneth Ward Church and Patrick Hanks: 'Word Association Norms, Mutual Information, and Lexicography', *Computational Linguistics* 16 (1): 22–29, 1990.

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 '12, pp. 840–850, Avignon, France, 2012.

Jacob Cohen: 'A Coefficient of Agreement for Nominal Scales', *Educational and Psychological Measurement* 20 (1): 37–46, 1960.

Robin Cohen: 'Analyzing the Structure of Argumentative Discourse', *Computational Linguistics* 13 (1-2): 11–24, 1987.

Michael Collins: 'Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms', in: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pp. 1–8, Pennsylvania, PA, USA, 2002.

Michael Collins: 'Head-Driven Statistical Models for Natural Language Parsing', *Computational Linguistics* 29 (4): 589–637, 2003.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa: 'Natural Language Processing (Almost) from Scratch', *Journal of Machine Learning Research* 12: 2493–2537, 2011.

David A. Conway: 'On the Distinction between Convergent and Linked Arguments', *Informal Logic* 13: 145–158, 1991.

Irving M. Copi and Carl Cohen: *Introduction To Logic*, Macmillan Publishing Company, 8th edition, 1990.

Corinna Cortes and Vladimir Vapnik: 'Support-Vector Networks', *Machine Learning* 20 (3): 273–297, 1995.

James Curran, Stephen Clark, and Johan Bos: 'Linguistically Motivated Large-Scale NLP with C&C and Boxer', in: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, ACL '07, pp. 33–36, Prague, Czech Republic, 2007.

T. Edward Damer: *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Reasoning*, Wadsworth Cengage Learning, 6th edition, 2009.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith: 'Frame-Semantic Parsing', *Computational Linguistics* 40:1: 9–56, 2014.

Peter Davies: 'Improving the quality of students' arguments through 'assessment for learning'', *Journal of Social Science Education (JSSE)* 8 (2): 94–104, 2009.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch: 'DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data', in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pp. 61–66, Baltimore, MD, USA, 2014.

Phan Minh Dung: 'On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games', *Artificial Intelligence* 77 (2): 321–357, 1995.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych: 'On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse', in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pp. 2236–2242, Lisbon, Portugal, 2015.

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

Frans H. van Eemeren and Rob Grootendorst: *A Systematic Theory of Argumentation: The pragma-dialectical approach*, Cambridge University Press, 2004.

Frans H. van Eemeren, Rob Grootendorst, and Francisca Snoeck Henkemans: *Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Developments*, Routledge, Taylor & Francis Group, 1996.

Mohammad Hassan Falakmasir, Kevin D. Ashley, Christian D. Schunn, and Diane J. Litman: *Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings*, chapter Identifying Thesis and Conclusion Statements in Student Essays to Scaffold Peer Review, pp. 254–259, Springer International Publishing, 2014.

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 - Volume 1*, HLT '11, pp. 987–996, Portland, OR, USA, 2011.

Vanessa Wei Feng and Graeme Hirst: 'A Linear-Time Bottom-Up Discourse Parser with Constraints and Post-Editing', in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 511–521, Baltimore, MA, USA, 2014.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning: 'Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling', in: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, ACL '05, pp. 363–370, Ann Arbor, Michigan, 2005.

Joseph L. Fleiss: 'Measuring nominal scale agreement among many raters.', *Psychological Bulletin* 76 (5): 378–382, 1971.

Eirini Florou, Stasinos Konstantopoulos, Antonis Koukourikos, and Pythagoras Karampiperis: 'Argument extraction for supporting public policy formulation', in: *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 49–54, Sofia, Bulgaria, 2013.

George Forman and Martin Scholz: 'Apples-to-apples in Cross-validation Studies: Pitfalls in Classifier Performance Measurement', *SIGKDD Explor. Newsl.* 12 (1): 49–57, 2010.

Austin J. Freeley and David L. Steinberg: *Argumentation and Debate: Critical Thinking for Reasoned Decision Making*, Wadsworth, Cengage Learning, 12th edition, 2009.

James B. Freeman: *Thinking logically: Basic concepts for reasoning*, Prentice-Hall, 1988.

James B. Freeman: *Argument Structure: Representation and Theory*, Argumentation Library Vol. 18, Springer, 2011.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui: 'Analyzing Argumentative Discourse Units in Online Interactions', in: *Proceedings of the First Workshop on Argumentation Mining*, pp. 39–48, Baltimore, MA, USA, 2014.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis: 'Argument Extraction from News, Blogs, and Social Media', in: *Artificial Intelligence: Methods and Applications*, Lecture Notes in Computer Science Vol. 8445, pp. 287–299, Springer International Publishing, 2014.

Trudy Govier: *A Practical Study of Argument*, Wadsworth, Cengage Learning, 7th edition, 2010.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot: *International Corpus of Learner English (Version 2)*, Presses universitaires de Louvain, 2009.

Wayne Grennan: *Informal Logic*, Montreal: McGill-Queen's University Press, 1997.

Leo Groarke: 'Informal Logic', in Edward N. Zalta (Ed.): *The Stanford Encyclopedia of Philosophy*, summer 2015 edition, 2015.

Leo Groarke and Christopher W. Tindale: *Good Reasoning Matters! A Constructive Approach to Critical Thinking*, Oxford University Press, 5th edition, 2012.

Ivan Habernal and Iryna Gurevych: 'Argumentation Mining in User-Generated Web Discourse', *Computational Linguistics* p. (in press), 2016a.

Ivan Habernal and Iryna Gurevych: 'Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM', in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1589–1599, Berlin, Germany, 2016b.

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

Charles Leonard Hamblin: *Fallacies*, Methuen & Co Ltd, 1970.

Hans Hansen: 'Fallacies', in Edward N. Zalta (Ed.): *The Stanford Encyclopedia of Philosophy*, summer 2015 edition, 2015.

Kazi Saidul Hasan and Vincent Ng: 'Predicting Stance in Ideological Debate with Rich Linguistic Knowledge', in: *Proceedings of COLING 2012: Posters*, pp. 451–460, Mumbai, India, 2012.

Kazi Saidul Hasan and Vincent Ng: 'Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates', in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 751–762, Doha, Qatar, 2014.

Arthur C. Hastings: *A Reformulation of the Modes of Reasoning in Argumentation*, Ph.D. thesis, Evanston, Illinois, 1963.

A. Francisca Snoeck Henkemans: 'State-of-the-Art: The Structure of Argumentation', *Argumentation* 14 (4): 447–473, 2000.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka: 'HILDA: A Discourse Parser Using Support Vector Machine Classification', *Dialogue and Discourse* 1 (3): 1–33, 2010.

David Hitchcock: *Anyone Who Has a View: Theoretical Contributions to the Study of Argumentation*, chapter Toulmin's Warrants, pp. 69–82, Springer Netherlands, Dordrecht, 2003.

Patrick J. Hurley: *A Concise Introduction to Logic*, Wadsworth, Cengage Learning, 2012.

Nani T. Indrajani and Anggie Angeline: 'The Types of Argument Structure Used by Hillary Clinton in the CNN Democratic Presidential Debate', *k@ta* 11 (2): 184–200, 2010.

Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu: 'Cutting-plane Training of Structural SVMs', *Machine Learning* 77 (1): 27–59, 2009.

George H. John and Pat Langley: 'Estimating Continuous Distributions in Bayesian Classifiers', in: *Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338–345, Montreal, Quebec, Canada, 1995.

Ralph H. Johnson: *Manifest rationality*, Lawrence Erlbaum, 2000.

Ralph H. Johnson and Anthony Blair: *Logical Self-Defense*, MacGraw Hill, 1994.

Ralph H. Johnson and Anthony J. Blair: *Logical Self-Defense*, McGraw-Hill Ryerson, 1977.

Ralph H. Johnson and Anthony J. Blair: *Logical Self-Defense*, International Debate Education Association, 2006.

Dave Kemper and Pat Sebranek: *Inside Writing: Persuasive Essays*, Great Source Education Group, 2004.

Manfred Kienpointner: *Alltagslogik: Struktur und Funktion von Argumentationsmustern*, Stuttgart: Fromman-Holzboog, 1992.

Yoon Kim: 'Convolutional Neural Networks for Sentence Classification', in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pp. 1746–1751, Doha, Qatar, 2014.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych: 'Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications', in: *Proceedings of the 2nd Workshop on Argumentation Mining*, pp. 1–11, Denver, CO, USA, 2015.

Dan Klein and Christopher D. Manning: 'Accurate Unlexicalized Parsing', in: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pp. 423–430, Sapporo, Japan, 2003.

Klaus Krippendorff: *Content Analysis: An Introduction to its Methodology*, Sage, 1980.

Klaus Krippendorff: 'Measuring the Reliability of Qualitative Text Analysis Data', *Quality & Quantity* 38 (6): 787–800, 2004.

Damjan Krstajic, Ljubomir J. Buturovic, David E. Leahy, and Simon Thomas: 'Cross-validation pitfalls when selecting and assessing regression and classification models', *Journal of Cheminformatics* 6 (10), 2014.

Sandra Kübler, Ryan McDonald, Joakim Nivre, and Graeme Hirst: *Dependency Parsing*, Morgan and Claypool Publishers, 2008.

Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W. Shulman: 'Identifying and Classifying Subjective Claims', in: *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, pp. 76–81, Philadelphia, PA, USA, 2007.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira: 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data', in: *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 282–289, San Francisco, CA, USA, 2001.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim: 'Context Dependent Claim Detection', in: *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pp. 1489–1500, Dublin, Ireland, 2014.

Hyojung Lim and Jimin Kahng: 'Review of Criterion®', *Language Learning & Technology* 16 (2): 38–45, 2012.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng: 'Recognizing Implicit Discourse Relations in the Penn Discourse Treebank', in: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, EMNLP '09, pp. 343–351, Suntec, Singapore, 2009.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan: 'A PDTB-styled end-to-end discourse parser', *Natural Language Engineering* 20 (2): 151–184, 2014.

Marco Lippi and Paolo Torroni: 'Context-Independent Claim Detection for Argument Mining', in: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pp. 185–191, Buenos Aires, Argentina, 2015.

Marco Lippi and Paolo Torroni: 'Argumentation Mining: State of the Art and Emerging Trends', *ACM Transactions on Internet Technology* p. (to appear), 2016.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova: 'Using Entity Features to Classify Implicit Discourse Relations', in: *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pp. 59–62, Stroudsburg, PA, USA, 2010.

Jim D. MacKenzie: 'The Dialectics of Logic', *Logique et Analyse* 94: 159–177, 1981.

Nitin Madnani, Michael Heilman, Joel Tetrault, and Martin Chodorow: 'Identifying High-Level Organizational Elements in Argumentative Discourse', in: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pp. 20–28, Montreal, Quebec, Canada, 2012.

William C. Mann and Sandra A. Thompson: 'Rhetorical Structure Theory: A Theory of Text Organization', *Technical Report ISI/RS-87-190*, Information Sciences Institute, 1987.

Daniel Marcu and Abdessamad Echihabi: 'An Unsupervised Approach to Recognizing Discourse Relations', in: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 368–375, 2002.

Christian M. Meyer, Margot Mieskes, Christian Stab, and Iryna Gurevych: 'DKPro Agreement: An Open-Source Java Library for Measuring Inter-Rater Agreement', in: *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations (COLING)*, pp. 105–109, Dublin, Ireland, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean: 'Distributed Representations of Words and Phrases and their Compositionality', in: *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, Curran Associates, Inc., 2013.

Makoto Miwa and Mohit Bansal: 'End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures', in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1105–1116, Berlin, Germany, 2016.

Raquel Mochales-Palau and Marie-Francine Moens: 'Study on Sentence Relations in the Automatic Detection of Argumentation in Legal Cases', in: *Proceedings of the 2007 Conference on Legal Knowledge and Information Systems: JURIX 2007: The Twentieth Annual Conference*, pp. 89–98, Leiden, Netherlands, 2007.

Raquel Mochales-Palau and Marie-Francine Moens: 'Study on the Structure of Argumentation in Case Law', in: *JURIX the twenty-first annual conference on legal knowledge and information systems*, pp. 11–20, Florence, Italy, 2008.

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 '09, pp. 98–107, Barcelona, Spain, 2009.

Raquel Mochales-Palau and Marie-Francine Moens: 'Argumentation mining', *Artificial Intelligence and Law* 19 (1): 1–22, 2011.

Marie-Francine Moens: 'Argumentation mining: Where are we now, where do we want to be and how do we get there?', in: *Post-proceedings of the Forum for Information Retrieval Evaluation (FIRE 2013)*, pp. 4–6, New Delhi, India, 2013.

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, pp. 225–230, Stanford, CA, USA, 2007.

Huy Nguyen and Diane Litman: 'Improving Argument Mining in Student Essays by Learning and Exploiting Argument Indicators versus Essay Topics', in: *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2016)*, pp. 485–490, 2016.

Naoaki Okazaki: 'CRFsuite: a fast implementation of Conditional Random Fields (CRFs)', http://www.chokkan.org/software/crfsuite/, 2007.

Daniel J. O'Keefe: 'Two Concepts of Argument', *Journal of the American Forensic Assiciation* 13 (3): 121–128, 1977.

Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker: 'And That's A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue', in: *Proceedings of the 2nd Workshop on Argumentation Mining*, pp. 116–126, Denver, CO, USA, 2015.

Joonsuk Park and Claire Cardie: 'Identifying Appropriate Support for Propositions in Online User Comments', in: *Proceedings of the First Workshop on Argumentation Mining*, pp. 29–38, Baltimore, MA, USA, 2014.

Andreas Peldszus: 'Towards segment-based recognition of argumentation structure in short texts', in: *Proceedings of the First Workshop on Argumentation Mining*, pp. 88–97, Baltimore, MA, USA, 2014.

Andreas Peldszus and Manfred Stede: 'From Argument Diagrams to Argumentation Mining in Texts: A Survey', *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7 (1): 1–31, 2013a.

Andreas Peldszus and Manfred Stede: 'Ranking the annotators: An agreement study on argumentation structure', in: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 196–204, Sofia, Bulgaria, 2013b.

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*, pp. 938–948, Lisbon, Portugal, 2015.

Andreas Peldszus and Manfred Stede: 'An annotated corpus of argumentative microtexts', in: *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation*, pp. 801–815, Portugal, Lisbon, 2016.

Chaim Perelman and Lucie Olbrechts-Tyteca: *The New Rhetoric: A Treatise on Argumentation*, University of Notre Dame Press, 1969.

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)*, pp. 260–269, Sofia, Bulgaria, 2013.

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)*, pp. 1534–1543, Baltimore, MA, USA, 2014.

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)*, pp. 543–552, Beijing, China, 2015.

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*, pp. 1384–1394, San Diego, California, 2016.

Vivien Perutz: *A Helpful Guide to Essay Writing!*, Student Services, Anglia Ruskin University, 2010.

Emily Pitler, Annie Louis, and Ani Nenkova: 'Automatic Sense Prediction for Implicit Discourse Relations in Text', in: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 683–691, Suntec, Singapore, 2009.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab: 'Automatic Committed Belief Tagging', in: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pp. 1014–1022, Beijing, China, 2010.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber: 'The Penn Discourse Treebank 2.0.', in: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

Minghui Qiu and Jing Jiang: 'A Latent Variable Model for Viewpoint Discovery from Threaded Forum Posts', in: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1031–1040, Atlanta, Georgia, 2013.

Ross Quinlan: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

Sarvapali D. Ramchurn, Lluís Godo Sierra, Carles, and Nicholas R. Jennings: 'Negotiating Using Rewards', *Artificial Intelligence* 171 (10–15): 805—837, 2007.

Lance A. Ramshaw and Mitchell P. Marcus: 'Text chunking using transformation-based learning', in: *Proceedings of the 3rd ACL Workshop on Very Large Corpora*, pp. 82–94, Cambridge, MA, USA, 1995.

Chris Reed, Raquel Mochales-Palau, Glenn Rowe, and Marie-Francine Moens: 'Language Resources for Studying Argument', in: *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC '08, pp. 2613–2618, Marrakech, Morocco, 2008.

Chris Reed and Glenn Rowe: 'Araucaria: Software for argument analysis, diagramming and representation', *International Journal on Artificial Intelligence Tools* 14 (4): 961–980, 2004.

Chris Reed and Glenn Rowe: 'Translating Toulmin Diagrams: Theory Neutrality in Argument Representation', *Argumentation* 19 (3): 267–286, 2006.

Chris Reed and Douglas Walton: 'Argumentation Schemes in Argument-as-Process and Argument-as-Product', in: *Proceedings of the Conference Celebrating Informal Logic @25*, Windsor, ON, USA, 2003.

Chris Reed, Douglas Walton, and Fabrizio Macagno: 'Argument Diagramming in Logic, Law and Artificial Intelligence', *Knowledge Engineering Review* 22 (1): 87–109, 2007.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. 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, pp. 440–450, Lisbon, Portugal, 2015.

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 Companion on World Wide Web*, WWW '16 Companion, pp. 991–996, Montréal, Canada, 2016.

Niall Rooney, Hui Wang, and Fiona Browne: 'Applying Kernel Methods to Argumentation Mining.', in: *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, FLAIRS '12, pp. 272–275, Marco Island, FL, USA, 2012.

Sara Rosenthal and Kathleen McKeown: 'Detecting Opinionated Claims in Online Discussions', in: *Sixth IEEE International Conference on Semantic Computing, ICSC 2012*, pp. 30–37, Palermo, Italy, 2012.

Victor D. Sampson and Douglas B. Clark: 'Assessment of Argument in Science Education: A Critical Review of the Literature', in: *Proceedings of the 7th International Conference on Learning Sciences*, ICLS '06, pp. 655–661, Bloomington, IN, USA, 2006.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis: 'Argument Extraction from News', in: *Proceedings of the 2nd Workshop on Argumentation Mining*, pp. 56–66, Denver, CO, 2015.

Robert E. Schapire and Yoram Singer: 'BoosTexter: A Boosting-based Systemfor Text Categorization', *Machine Learning* 39 (2-3): 135–168, 2000.

Edward Schiappa and John P. Nordin: *Keeping Faith with Reason: A Theory of Practical Reason*, Pearson Learning Solutions, 2013.

William A. Scott: 'Reliability of Content Analysis: The Case of Nominal Scale Coding', *Public Opinion Quarterly* 19 (3): 321–325, 1955.

Alan Sergeant: 'Automatic Argumentation Extraction', in: *Proceedings of the 10th European Semantic Web Conference*, ESWC '13, pp. 656–660, Montpellier, France, 2013.

Mark D. Shermis and Jill Burstein: *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, Routledge Chapman & Hall, 2013a.

Mark D. Shermis and Jill Burstein (Eds.): *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, chapter The e-rater automated essay scoring system, Routledge Chapman & Hall, 2013b.

Don Shiach: *How to write essays*, How To Books Ltd, 2nd edition, 2009.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts: 'Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank', in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, WA, USA, 2013.

Anders Søgaard: 'Estimating effect size across datasets', in: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 607–611, Atlanta, Georgia, 2013.

Marina Sokolova and Guy Lapalme: 'A Systematic Analysis of Performance Measures for Classification Tasks', *Information Processing & Management* 45 (4): 427–437, 2009.

Swapna Somasundaran and Janyce Wiebe: 'Recognizing Stances in Online Debates', in: *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, ACL '09, pp. 226–234, Suntec, Singapore, 2009.

Swapna Somasundaran and Janyce Wiebe: 'Recognizing Stances in Ideological Online Debates', in: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pp. 116–124, Los Angeles, CA, USA, 2010.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane: 'Applying Argumentation Schemes for Essay Scoring', in: *Proceedings of the First Workshop on Argumentation Mining*, pp. 69–78, Baltimore, MA, USA, 2014.

Radu Soricut and Daniel Marcu: 'Sentence Level Discourse Parsing Using Syntactic and Lexical Information', in: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pp. 149–156, Edmonton, Canada, 2003.

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)*, pp. 1501–1510, Dublin, Ireland, 2014b.

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)*, pp. 46–56, Doha, Qatar, 2014a.

Christian Stab and Iryna Gurevych: 'Parsing Argumentation Structures in Persuasive Essays', *arXiv preprint arXiv:1604.07370* 2016.

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych: 'Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective', in: *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pp. 40–49, Bertinoro, Italy, 2014.

Keith E. Stanovich, Richard F. West, and Maggie E. Toplak: 'Myside Bias, Rational Thinking, and Intelligence', *Current Directions in Psychological Science* 22 (4): 259–264, 2013.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii: 'BRAT: A Web-based Tool for NLP-assisted Text Annotation', in: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pp. 102–107, Avignon, France, 2012.

Stephen N. Thomas: *Practical reasoning in natural language*, Prentice-Hall, 1973.

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

Stephen E. Toulmin: *The Uses of Argument*, Cambridge University Press, Updated Edition, 2003.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer: 'Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network', in: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL '03, pp. 173–180, Edmonton, Canada, 2003.

Peter D. Turney: 'Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews', in: *Proceedings of the 40th Annual*

Meeting on Association for Computational Linguistics, ACL '02, pp. 417–424, Philadelphia, Pennsylvania, 2002.

Henning Wachsmuth and Stein Benno: 'Sentiment Flow - A Universal Model for Discourse-level Argumentation Analysis', *Special Section of the ACM Transactions on Internet Technology: Argumentation in Social Media* p. (to appear), 2016.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein: 'Sentiment Flow—A General Model of Web Review Argumentation', in Lluís Márquez, Chris Callison-Burch, and Jian Su (Eds.): *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 15)*, pp. 601–611, Lisbon, Portugal, 2015.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King: 'A Corpus for Research on Deliberation and Debate', in: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012.

Douglas Walton and David M. Godden: *Reason Reclaimed*, chapter Informal Logic and the Dialectical Approach to Argument, pp. 3–17, Newport News, Vale Press, Virginia, USA, 2007.

Douglas Walton, Chris Reed, and Fabrizio Macagno: *Argumentation Schemes*, Cambridge University Press, 2008.

Douglas N Walton: *Argumentation schemes for presumptive reasoning*, Routledge, 1996.

Anthony Weston: *Rulebook for Arguments*, Hackett Publishing Company, 3rd edition, 2000.

Anne Whitaker: *Academic Writing Guide 2010: A Step-by-Step Guide to Writing Academic Papers*, City University of Seattle, 2009.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann: 'Recognizing Contextual Polarity in Phrase-level Sentiment Analysis', in: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 347–354, Vancouver, British Columbia, Canada, 2005.

Christopher R. Wolfe and M. Anne Britt: 'Argumentation Schema and the Myside Bias in Written Argumentation', *Written Communication* 26 (2): 183–209, 2009.

John Woods and Douglas Walton: *Fallacies: Selected Papers 1972-1982*, College Publications, 2007.

Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor J. M. Bench-Capon: 'Semi-Automated Argumentative Analysis of Online Product Reviews.', in: *COMMA*, Frontiers in Artificial Intelligence and Applications Vol. 245, pp. 43–50, IOS Press, 2012.

Robert J. Yanal: 'Dependent and Independent Reasons', *Informal Logic* 13 (3): 137–144, 1991.

Matthew D. Zeiler: 'ADADELTA: an adaptive learning rate method', *arXiv 1212.5701* 2012.

Ehrenwörtliche Erklärung[†]

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades "Dr.-Ing." mit dem Titel "Argumentative Writing Support by means of Natural Language Processing" selbständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, den 13. Dezember 2016

Dipl.-Inform. Christian Matthias Edwin Stab

[†] Gemäß §9 Abs. 1 der Promotionsordnung der TU Darmstadt

Wissenschaftlicher Werdegang des Verfassers[†]

10/03 – 07/09	Studium der Informatik an der Technischen Universität Darmstadt
01/09 – 07/09	Diplomarbeit „*Interaktionsanalyse für adaptive Benutzerschnittstellen*“ am Fraunhofer Institut für graphische Datenverarbeitung (IGD)
08/09 – 09/13	Wissenschaftlicher Mitarbeiter am Fraunhofer Institut für graphische Datenverarbeitung (IGD)
10/13 – 02/17	Doktorand am Fachgebiet Ubiquitous Knowledge Processing (UKP-Lab) der Technischen Universität Darmstadt

[†] Gemäß §20 Abs. 3 der Promotionsordnung der TU Darmstadt

Publikationsverzeichnis des Verfassers

Christian Stab and Iryna Gurevych. 2017. Recognizing Insufficiently Supported Arguments in Argumentative Essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '17, pp. (to appear), Valencia, Spain

Christian Stab and Iryna Gurevych. 2016. Recognizing the Absence of Opposing Arguments in Persuasive Essays. In *Proceedings of the 3rd Workshop on Argument Mining*, pp. 113–118, Berlin, Germany

Beata Beigman Klebanov, **Christian Stab**, Jill Burstein, Yi Song, Binod Gyawali and Iryna Gurevych. 2016. Argumentation: Content, Structure, and Relationships with Essay Quality. In *Proceedings of the 3rd Workshop on Argument Mining*, pp. 70–75, Berlin, Germany

Christian Stab and Ivan Habernal. 2016. Existing Resources for Debating Technologies. *Report of Dagstuhl Seminar on Debating Technologies (15512)*, pp. 32, Wadern, Germany

Christian Stab and Ivan Habernal. 2016. Detecting Argument Components and Structures. *Report of Dagstuhl Seminar on Debating Technologies (15512)*, pp. 32-33, Wadern, Germany

Iryna Gurevych and **Christian Stab**. 2016. Argumentative Writing Support: Structure Identification and Quality Assessment of Arguments. In *Report of Dagstuhl Seminar on Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments (16161)*, pp. 87-88, Wadern, Germany

Christian Stab and Iryna Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pp. 46–56, Doha, Qatar

Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, pp. 1501–1510, Dublin, Ireland

Christian M. Meyer, Margot Mieskes, **Christian Stab**, and Iryna Gurevych. 2014. DKPro Agreement: An Open-Source Java library for Measuring Inter-Rater Agreement. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations* COLING '14, pp. 105–109, Dublin, Ireland

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective. In *Proceedings of the Workshop on*

Frontiers and Connections between Argumentation Theory and Natural Language Processing, pp. 40–49, Bertinoro, Italy