# Automatic recognition of argumentation structure in short monological texts

**Andreas Peldszus**


Doctoral Thesis

submitted to the Faculty of Human Sciences,
Department Linguistics at the University of Potsdam

in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy (Dr. phil.)
in Computational Linguistics

Supervisor:
Prof. Dr. Manfred Stede

Place and date of defense:
Potsdam, October 27, 2017

Submitted April 21, 2017
Accepted May 3, 2017
Defended October 27, 2017

Reviewers:
Prof. Dr. Manfred Stede
Prof. Chris Reed

Published online at the
Institutional Repository of the University of Potsdam:
URN urn:nbn:de:kobv:517-opus4-421441
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-421441

# Summary

The aim of this thesis is to develop approaches to automatically recognise the structure of argumentation in short monological texts. This amounts to identifying the central claim of the text, supporting premises, possible objections, and counter-objections to these objections, and connecting them correspondingly to a structure that adequately describes the argumentation presented in the text.

The first step towards such an automatic analysis of the structure of argumentation is to know how to represent it. We systematically review the literature on theories of discourse, as well as on theories of the structure of argumentation against a set of requirements and desiderata, and identify the theory of J. B. Freeman (1991, 2011) as a suitable candidate to represent argumentation structure. Based on this, a scheme is derived that is able to represent complex argumentative structures and can cope with various segmentation issues typically occurring in authentic text.

In order to empirically test our scheme for reliability of annotation, we conduct several annotation experiments, the most important of which assesses the agreement in reconstructing argumentation structure. The results show that expert annotators produce very reliable annotations, while the results of non-expert annotators highly depend on their training in and commitment to the task.

We then introduce the 'microtext' corpus, a collection of short argumentative texts. We report on the creation, translation, and annotation of it and provide a variety of statistics. It is the first parallel corpus (with a German and English version) annotated with argumentation structure, and – thanks to the work of our colleagues – also the first annotated according to multiple theories of (global) discourse structure.

The corpus is then used to develop and evaluate approaches to automatically predict argumentation structures in a series of six studies: The first two of them focus on learning local models for different aspects of argumentation structure. In the third study, we develop the main approach proposed in this thesis for predicting globally optimal argumentation structures: the 'evidence graph' model. This model is then systematically compared to other approaches in the fourth study, and achieves state-of-the-art results on the microtext corpus. The remaining two studies aim to demonstrate the versatility and elegance of the proposed approach by predicting argumentation structures of different granularity from text, and finally by using it to translate rhetorical structure representations into argumentation structures.

# Zusammenfassung

Ziel dieser Arbeit ist die Entwicklung von Methoden zur automatischen Erkennung der Argumentationsstruktur in kurzen, monologischen Texten. Dies umfasst einerseits, die zentrale These des Textes, stützende Prämissen, mögliche Einwände und Widersprüche gegen diese zu identifizieren. Andererseits gilt es, diese Elemente in einer Gesamtstruktur zu verbinden, die die im Text vorgebrachte Argumentation angemessen beschreibt.

Hierzu muss zuerst eine geeignete Darstellung der Argumentationsstruktur gefunden werden. Anhand einer Reihe von Anforderungen wird die Literatur zu Theorien der Diskurs- sowie der Argumentationsstruktur systematisch ausgewertet. Die Theorie von J. B. Freeman (1991, 2011) erweist sich hierbei als geeigneter Kandidat zur Repräsentation von Argumentationsstruktur. Darauf aufbauend wird ein Annotationsschema abgeleitet, welches auch komplexe Strukturen klar darstellen und mit verschiedenen, für authentischen Text typischen Segmentierungsproblemen umgehen kann.

Um das Schema hinsichtlich der Zuverlässigkeit der Annotation empirisch zu testen, werden mehrere Annotationsexperimente durchgeführt, von denen das wichtigste die Übereinstimmung bei der Rekonstruktion der Argumentationsstruktur erfasst. Die Ergebnisse zeigen, dass Fachexperten sehr verlässlich annotieren, während die Ergebnisse von Nicht-Experten in hohem Maße vom Training und ihrem Engagement für die Aufgabe abhängen.

Schließlich wird das ‚microtext'-Korpus vorgestellt, eine Sammlung kurzer argumentativer Texte. Die Erstellung, Übersetzung und Annotation wird beschrieben, die Strukturen statistisch ausgewertet. Es handelt sich um das erste mit Argumentationsstrukturen annotierte Parallelkorpus (in Deutsch und Englisch) und – dank der Arbeit unserer Kollegen – auch um das erste, das mit verschiedenartigen Diskursstrukturen annotiert wurde.

In einer Reihe von sechs Studien werden dann Methoden zur automatischen Erkennung von Argumentationsstrukturen entwickelt und am Korpus erprobt: Die ersten beiden konzentrieren sich auf das Lernen lokaler Modelle für einzelne Aspekte der Argumentationsstruktur. In der dritten Studie wird der in dieser Dissertation vorgeschlagene Ansatz entwickelt: das ‚Evidenzgraph'-Modell, mit dem global optimale Argumentationsstrukturen erkannt werden können. Dieses wird dann in der vierten Studie systematisch mit anderen Ansätzen verglichen und erzielt beste Ergebnisse auf dem microtext-Korpus. Die verbleibenden zwei Studien zielen darauf ab, die Vielseitigkeit und Eleganz des Ansatzes zu demonstrieren, z.B. bei der Ableitung von Argumentationsstrukturen unterschiedlicher Granularität oder bei der Übersetzung rhetorischer Strukturen in Argumentationsstrukturen.

# Acknowledgments

First and foremost, my deep appreciation and sincere thanks go to my supervisor, Manfred Stede. I have been studying and working for and with him for many years now, and can hardly express how much he taught me about language and computational linguistics, about science and academics, and about reading and writing. He played a central role in initially spurring and then nurturing my interest in the study of discourse and argumentation. As a supervisor, he was always approachable, only one door or one email away. Looking back, he knew very well when to encourage me to buckle down and when not to – which is not to say that I was always wise enough to follow his advice.

For his enthusiasm and interest, I thank Chris Reed, who generously agreed to talk through my work on various occasions. He helped me realise that exciting things could be built upon something that was initially of mostly theoretical interest to me.

I am grateful to have been able to visit our colleagues in Toulouse at IRIT. During our collaboration, I benefitted much from the feedback that Stergos Afantenos, Nicholas Asher, Jérémy Perret, and Patrick Saint-Dizier provided. Similarly, I would like to thank the Dundee-Warsaw-connection – members of which I met not only in Dundee and in Warsaw – especially Kasia Budzynska, John Lawrence, and of course Chris Reed. They always gave me a warm welcome, challenging questions, and sometimes even a round of arm-wrestling.

It was a privilege to attend several conferences and workshops to present my ideas, which evolved with every thoughtful comment I received from anonymous reviewers and participants.

Our research community at the Linguistics Department, including Alexander Koller, and my colleagues Tatjana Scheffler, Željko Agić and Christoph Teichmann, offered fruitful discussions in meeting rooms and hallways. For many stimulating exchanges, I thank my office mates – here in order of appearance – Okko Buß, Florian Kuhn, Jonathan Sonntag, Arne Neumann, Wladimir Sidorenko, and Yulia Grishina.

Without being able to list them all here, I thank the participants of the experiments conducted in the course of this work. My special thanks go to Markus, who agreed to serve as an early test subject in many pilot studies and considerably helped to refine the experiment setup. I am also thankful to Kirsten Brock and Regina Peldszus for translating the corpus.

For allowing me to pursue this thesis project, grants from the Collaborative Research Centre (SFB) 632 and Cusanuswerk are gratefully acknowledged. Furthermore, I would like to thank Cusanuswerk for all other forms of support I enjoyed. I am grateful to the

Potsdam Graduate School and FNK for enabling me to travel, and to Annett for helping me sort out the paperwork afterwards.

I am indebted to Regina, who tracked down a number of typographical errors in the all but final version of this work after they went undetected by the spell-checker, and adorned earlier hard-copy drafts with anthropomorphised illustrations of my linguistic aberrations.

Finally, I am endlessly thankful to my parents and to my wife Bettina. Without their patience, encouragement, humour, and love this thesis would not exist.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

The aim of this thesis is to develop approaches to automatically recognise the structure of argumentation in short monological texts. This amounts to identifying the central claim of the text, supporting premises, possible objections, and counter-objections to these objections, and connecting them correspondingly to a structure that adequately describes the argumentation presented in the text. In this work, we will thus address the following research questions:

1. **Representation:** How should we represent the structure of argumentation? Which representational devices are required?
2. **Annotation:** Can analysts annotate argumentation structures reliably? Which aspects are easy to agree on? Which constellations are likely to cause disagreements?
3. **Recognition:** To what degree can argumentation structures be recognised automatically? How can we model both the local relations between text segments, as well as global structural preferences?

Argumentation is such a fundamental linguistic activity in human life, however, that there is an abundance of different sorts of evidence of this activity. We would burst the scope of a thesis if we aimed at covering the vast plenitude of all these forms of argumentation. For this work, we hence narrow down the scope.

First, we focus on written monologue text. We do not consider transcripts of argumentative dialogues, conversations, debates, or mediations between multiple parties. Instead, the object of this research is a text, planned and written by an author, arguing for a claim.

Second, we work with short texts. Representing and recognising argumentative structures in long essays, chapters, or even books is beyond the scope of this work. Instead, we examine short texts as one would find them in a paragraph or in a short opinion piece. But the methods we will develop can also be applied to texts as long as a news commentary.

Third, the argumentative relations we aim to represent and recognise are those between arguments expressed in the document. The structures of interest are thus intra-document argumentation structures. We will not attempt to model the relations between different documents, such as e.g. between different posts in a forum thread or comment stream, or between articles and the responses they provoked.

Finally, our aim is to study textual argumentation, as we produce and consume it on an everyday basis, e.g. when we read newspaper commentaries, when we propose ideas, when

[1] German universities should on no account charge tuition fees. [2] This would simply mean that only those people with wealthy parents or a previous education and a part-time job while studying would be able to apply for a degree programme in the first place. [3] Even without tuition fees half of the student population has to work while studying and hence has less time for studying or recreation. [4] *One could argue that an increase in tuition fees would allow institutions to be better equipped.* [5] But what is the good of a wonderfully outfitted university if it doesn't actually allow the majority of clever people to broaden their horizons with all that great equipment?

Figure 1.1: An example of a short argumentative text (micro_k12).

we discuss what to do, or when we argue how to perceive the world. This means we are not focusing on forms of argumentation that are specific to a certain field or discipline, as e.g. in legal decisions, in scientific articles, or in technical documents. All of these come with their own forms of expressing, structuring and referencing arguments, and additional knowledge of the conventions and of the language used is required in order to understand the arguments therein.

## 1.1 Argumentation

Before we describe in more detail what we want to achieve in this thesis, we want to introduce important terminology, using an example text which is shown in Figure 1.1. Note that we will provide a more elaborate discussion of most of the terms introduced here in the corresponding theory Chapters 2 and 3.

The object of our study are argumentative texts. In such texts, the author develops a standpoint towards a controversial issue and tries to convince the reader to accept her standpoint. The position the author is alluring to is the *central claim* or the *thesis* of the text. In our example, the central claim is expressed in the first underlined sentence. The controversial issue is the question whether German universities should charge tuition fees or not, and it is the author's standpoint that this should not be the case.

In order to persuade the reader to accept the central claim, the author puts forward *arguments*. As an argument, we conceive the complex of one or more *premises* supporting one *conclusion*. In our example, the second sentence can be considered a premise lending support to the conclusion in the first sentence, the central claim: The author supports her standpoint to not charge tuition fees by warning about the negative consequence that only the affluent or already educated would be able to study if such fees were charged. We will say that an argumentative relation of support holds between the second and the first sentence, or that it is the argumentative *function* of the second sentence to support the first.

Yet, not all sentences in our example text serve the purpose of supporting the central claim. The fourth italicised sentence mentions an objection, a claim that is in opposition to the thesis of the text. The argumentative function of the fourth sentence is not that of support but of *attack*: Tuition fees could be used to better equip the university, and this speaks against the claim to do without them. Of course this objection was not mentioned to remain undisputed and effectively weaken the author's argument. It is mentioned for a reason, namely to be challenged by a corresponding counter-attack. In our example, the author first concedes that tuition fees could help to improve equipment, but then, in turn, questions the value of universities that are inaccessible to the majority of 'clever people'.

This interaction between claims pro and contra the main thesis of the text can be described on a dialectical level. Two argumentative *roles* are in conflict: the *proponent* who presents and defends the central claim, and the *opponent* who critically questions this reasoning. The author presents both sides, although from the point of view of the proponent.

When multiple arguments are related to each other and form larger complexes, we speak of an *argumentation*. The way the arguments are put together determines the structure of argumentation. These structures can be visualised as argument *diagrams*. We will investigate the different ways to create such structures in Chapters 2 and 3.

Finally, an important questions concerns what to employ as the basic units of argumentation. In our example we have for the sake of this introduction relied on units of sentential size to represent claims and premises. Frequently, though, the text can be divided into smaller *segments*, such as single clauses or even short phrases. What is constitutive of a corresponding unit is that the associated text span can be interpreted as expressing a proposition. In discourse theory, this is typically captured by the idea of an *elementary discourse unit* (EDU). We will argue that not every EDU is argumentatively relevant and should be integrated into the argumentation structure. Furthermore, oftentimes adjacent EDUs have to be combined and understood as one more complex proposition, in order to be of relevance for the argumentation structure. We will thus introduce the concept of *argumentative discourse units* (ADU), which span over one or more EDUs and serve as the fundamental units of argumentation structure.

### Utilised corpora

The example text shown in Figure 1.1 is taken from the 'microtext' corpus – a corpus that has been collected and studied in the course of this thesis and will also be used in our experiments in automating the recognition of argumentation structure. The texts of this corpus are authentic argumentations, but constrained to be relatively short, clear, and dense. We will provide motivation for using this corpus, as well as a detailed description of it in Chapter 5. Another resource we will use, though only in two smaller studies, is the corpus of Pro & Contra commentaries, taken from the Potsdam Commentary Corpus (PCC) [Stede,

2004, Stede and Neumann, 2014]. These texts have been extracted from a daily German newspaper, and explicitly discuss a controversial issue of regional political interest. In comparison to the microtexts, these commentaries are longer and are produced by professional journalists, which has a clear impact on the style of the writing. Most importantly, they are less restricted and may e.g. contain parts which are not relevant to the argument itself. A more detailed description and comparison will be given later on (in the Chapters 4.4 and 6.4). We consider the microtexts a good starting point for the study of automatic recognition of argumentation structure, whereas the Pro & Contra commentaries are more complex and therefore more challenging examples of argumentation. The methods proposed in this thesis are thus mainly developed and tested on the microtexts, but aim to be applicable also on the more complex Pro & Contra commentaries.

## 1.2 Argumentation Mining – The Problem

The problem that we address has been discussed under the term 'argumentation mining', which was first introduced by Mochales Palau and Moens [2009]. The problem of argumentation mining can be described as a process that seeks to automatically recognise the structure of argumentation in a text by identifying and connecting the central claim of the text, supporting premises, possible objections, and counter-objections to these objections. This overall goal can be split up into various different tasks, some of which will depend on others. In the following, we will introduce the tasks involved:[1]

1. **Argumentativeness detection**: Not every text is argumentative. Identifying texts which aim to actively persuade the reader to believe a certain proposition, to accept a characterisation or evaluation, or to adopt the commitment to act in a certain way, is not always a trivial task. It is thus the first step to select text that should be the subject of argumentative analysis.

2. **ADU identification**: Once we have selected a text that aims to persuade the reader through arguments, we have to identify which spans in the text form argumentatively relevant units. Even very dense argumentative texts exhibit parts irrelevant to the argumentation, because they merely set the scene, provide background information, or are digressions or deal a side-blow. Such spans will not participate in the argumentation structure and thus have to be excluded.

   Identifying argumentative units can be implemented in different ways. Most approaches rely on boundaries of linguistic units on the propositional level, such as

[1] Note that some of these tasks have been combined in related work, and not all tasks have to be tackled depending on the goals and purposes of the corresponding work. A more detailed and systematic review of these approaches will later be given in Chapter 6.

full sentences or clause-sized units, EDUs. If a text is segmented into such units, ADU identification amounts to discarding irrelevant sentences / EDUs, and eventually combining adjacent EDUs to one ADU. Other approaches rely on a custom boundary detection and directly learn to delimit argumentative unit as free token-spans.

3. **ADU classification**: The next step is to determining the type of for each ADU. This is an optional task, which is not necessary for deriving an argumentation structure, but it is often employed in order to describe the argumentative unit in more detail. Different schemes and typologies have been proposed, involving stance, evidence types, rhetorical status, and argumentative role and function. The proposed schemes have different foci, as they may serve different purposes, often depending on the domain and genre of the text and potential downstream tasks.

4. **Relation identification**: The ADUs are now connected by identifying whether an argumentative relation holds between them or not. This process results in a tree- or graph-structure. Ideally, it should yield a connected structure, which has one root (the ADU expressing the central claim of the text) and every other ADU is either directly or indirectly in relation with it.

5. **Relation classification**: Until now, the relations are unlabelled. In a second step, the type of argumentative relation is determined. A minimal distinction would involve the coarse-grained classes of supporting relations versus attacking relations, but much more fine-grained relation types have been proposed.

6. **Argument completion**: Finally, a last and very challenging step is the postulation of those 'implicit' ADUs, often referred to as 'suppressed' or 'missing premises' that are required to be accepted in order to make the argument felicitous: enthymeme reconstruction.

In this work, we will mainly address subtasks 3 to 5: Given a text segmented into relevant ADUs, identify the argumentation structure. We will also tackle aspects of the prior task of (2) ADU identification, but not the full task of ruling out argumentatively irrelevant material.

## 1.3 Motivation

Finding arguments in text automatically is a relatively new research area. It constitutes a very interesting intersection of different fields including natural language processing, logic, and artificial intelligence, but it is not an end in itself. In this section, we want to provide our motivation as to why, besides all theoretical reasons, it is worthwhile to engage in this scientific field. We argue that argumentation mining is potentially relevant to any kind of

text mining application that is directed at argumentative text. We will provide a handful of examples for different applications in several domains and how they could benefit from automatic argumentative analysis.

### Retrieval of argumentation

When we know that a certain genre of text is highly argumentative or has clearly argumentative parts, argumentation mining systems could help in identifying and retrieving these arguments from larger collections of text, thus facilitating the search for reasons and argumentative patterns.

In the legal domain, Mochales Palau and Moens [2009], amongst others, discuss the importance of finding argumentations and their structure in legal cases, as a subtask of the more general problem of finding *precedents* for a case that is currently under investigation.

Another example are scientific texts. While the work in all scientific disciplines is clearly related to argumentation, the texts in some disciplines are more amenable to automatic analysis than others. In domains such as the biomedical, where the scientific arguments are often presented in a very concise and conventionalised way, argumentative structures might be identified more easily and used for analysing and representing the progression of scientific debate [Teufel, 2010]. Furthermore, it might also facilitate discovery, such as when using argument mining techniques to support the detecting of drug-drug interactions in the field of biomedical text mining [Tari et al., 2010].

Finally, argument mining might be useful for the widely-popular task of opinion mining, which aims at detecting users' appreciations or disappointments with products or services in user-generated text. A natural extension is to also find automatically the *reasons* they provide for their evaluations. This is of interest in the domain of consumer feedback about commercial products as in product reviews or social media. Several researchers have explored this direction, including Wyner et al. [2012]. Another example is the e-policy domain, where politicians aim to understand the need and appreciations of new laws by analysing and searching in the feedback collected in designated networks for public political participation or in social media. First steps into this direction have been made for instance by Liebeck et al. [2016].

### Assessing the quality of argumentation

Argumentation structures also exhibits a lot of features relevant for assessing the quality of argumentation. It is for example possible to quantify the amount of argumentatively relevant text portions, to separate out the ordering and presentation of different argumentative chains, and to find uncountered objections or only weakly supported claims. Depending on the granularity of the scheme, it might even be possible to identify certain argumentative patterns which are considered fallacious.

One obvious application for this is the automatic scoring of student essays. Available systems are based on different heuristics, e.g. on orthography, vocabulary, length, document structure etc; but only recently the structure of argumentation has been used as a feature to also automatically assess the well-formedness of the expressed argumentation [Wachsmuth et al., 2016, Ghosh et al., 2016].

Another example, where understanding the quality of argumentation is useful, is ranking product reviews. Shopping platforms which receive a large amount of product reviews are faced with the challenge to place useful and important reviews in a prominent position. Argumentation mining techniques might help to sort out well-balanced reviews that give proper justification for their evaluations, cover different aspects, but also consider potential downsides.

A similar problem applies to the many news and blog platforms, which allow users to comment and discuss published articles. On the one hand, there is the need to select and promote comments of high quality from the often heterogeneous stream of incoming feedback, in order to give new users a starting point to engage. On the other hand, there is the challenge to detect use of abusive language, and posts which leave the grounds of a matter-of-fact debate.

**Supporting argumentation and public deliberation**

Argumentation mining techniques are also used as a tool for supporting the debate between users and to foster the public deliberations on online platforms. Automatic recognition of the structure of argumentation can here be used to help users to study the argument, to structure and classify it as belonging to one or the other category. Furthermore, this analysis can be a starting point for improving or countering the argument.

There exist different argumentation tools such as e.g. 'Carneades' [Gordon, 2010], which when given a structural representation of the argument, visualise the argument and support the user in further developing the argument, or develop strategies how to successfully attack it [Walton, 2011]. If at least a preliminary analysis of the structure of an argument could be derived automatically, this would greatly improve the reach and applicability of such tools. In a similar way, automatic analyses could help the users of e-policy platforms and debate portals to collaboratively organise and classify arguments of ongoing debates.

In the educational domain, similar approaches could be used to help students improve their essay writing skills [Stab, 2017] or to provide better feedback when reviewing other students' essays [Nguyen et al., 2017].

**Text representations for downstream NLP tasks**

Finally, it has been shown that certain NLP tasks thrive on being aware of the structure of the discourse they are applied to. Argumentation mining in this case can provide discourse structures specialised to represent the structures exhibited in argumentative text.

A first step into this direction was made by Teufel and Moens [2002] and later by Contractor et al. [2012], who showed how automatic document summarization systems can benefit from an argumentative representation in the domain of scientific papers. In a similar fashion, more general representations of discourse structure had been used to guide extractive [Marcu, 2000, Polanyi et al., 2004b] or abstractive summarization systems [Gerani et al., 2014].

Discourse structures have furthermore been exploited to improve sentiment analysis and opinion mining. There is work demonstrating the advantage of using local discourse relations [Somasundaran et al., 2009], but also higher level discourse structures [Bhatia et al., 2015]. The argumentation present in product reviews was for example focused by Wachsmuth et al. [2014] in this context.

Last but not least question-answering systems could profit from discourse and argumentation structures. One example is [Bosma, 2004], who uses extractive summaries based on rhetorical structures in order to generate answers to a query. Argumentative analyses might here serve as a basis for answering certain question types, as ultimately – whenever a question answering system is faced with 'why' questions – it is asked for justifications and explanations.

## 1.4 Thesis organisation

The thesis is organised in seven chapters. In the following, we will give an overview of these chapters.

**Chapter 2 – Theories of the structure of argumentation:** The first step towards an automatic analysis of the structure of argumentation is to know *how* to represent it. We thus define requirements as well as a few desiderata for suitable theories of the structure of argumentation. Then, we first review the literature on the theory of discourse, to determine whether the computational linguistics community already provided us with a solution that could be directly or indirectly used. As this is not the case, we then review the literature on theories of the structure of argumentation per se. We will identify the work of Freeman [1991, 2011] as a suitable candidate.

**Chapter 3 – A synthesised scheme of the structure of argumentation:** Based on Freeman's theory, we devise a synthesised scheme for representing the structure of argumentation in authentic text. The scheme represents complex and nested attack and counter-attack structures, as we think, in a more elegant way than in Freeman's original formulation. Fur-

thermore, we extend it with features required to cope with various segmentation issues typically occurring in authentic text.

**Chapter 4 – Agreeing on the structure of argumentation:** Since the theory we use in our scheme has not been empirically tested for reliability, we conducted annotation experiments in order to assess inter-annotator agreement. We report on three experiments, of which the first one is the most important. It assesses the agreement for three groups of annotators: naive student annotators, more experienced student annotators, and expert annotators. The results show that experts produce very reliable annotations with $\kappa$=0.83, but that the results of non-expert annotators highly depend on their training in and commitment to the task.

**Chapter 5 – A corpus of texts annotated with argumentation structure:** In this chapter, we present the 'microtext' corpus, a collection of short argumentative texts. It is the first parallel corpus of argumentation (with a German and English version). We report on the creation, translation, and annotation and provide a variety of statistics. Furthermore, we describe different transformations of the corpus with varying granularity that we will use in our experiments. Finally, it is the first corpus annotated according to multiple theories of global discourse structure, as it has also been annotated according to two other theories of discourse structure.

**Chapter 6 – Automatic Recognition of Argumentation Structure:** This chapter presents our efforts to automatically predict the argumentation structures of the microtext corpus. We first systematically review related work and then present six studies on the automatic recognition of argumentation structure. The first two studies focus on learning local models for different aspects of argumentation structure. In the third study (Ch. 6.5), we develop the approach proposed in this thesis for predicting globally optimal argumentation structures: the 'evidence graph' model. In the following study, we systematically compare this model to other approaches for recognising argumentation structures and derive state-of-the-art results using improved local models. The remaining two studies aim to demonstrate the versatility and elegance of the proposed approach by using it to predict argumentation structure of different granularity from text, and finally by using it to translate rhetorical structure representations into argumentation structures.

**Chapter 7 – Conclusion:** In the last chapter, we summarise what has been achieved in this thesis. We discuss the results and provide an outlook for future work.

### Previously published material

Several results of this thesis have already been published as journal papers, and in conference or workshop proceedings. We will highlight in the beginning of each chapter which of its contributions have already been previously published, and in case of joint work specify the contribution of the author of this thesis.

# 2 Theories of the structure of argumentation

When we analyse arguments expressed in a written piece of text or brought forward in a conversation, it becomes evident that argumentation is not merely unfolding in a single sentence. Although it is possible to express arguments in a single sentence with two clauses, they usually stretch over multiple sentences. A theory that aims to represent argumentation will consequently have to describe the role of, and the relations between, those sentences and clauses of a text.

One should hence take into consideration the range of theories on *discourse relations* and *discourse structure* that philosophy, linguistics and pragmatics have brought forward. The goal of these is to explain the coherence of a text in general: How does a text appear to the reader as a unified whole, and how do its parts relate to one another? Several such discourse structure theories have been proposed, and their design decisions often reflect the specific sub-discipline in which they originated: Some are extension of syntactic theories, others of semantic theories or even theories of logical inference. The central organising principle will be different across these theories: Some propose a sequential structure, others a hierarchical or a graph-structure. Also, the size of the object to be described varies: Some theories propose multiple local and disconnected structures, others one global and connected structure.

The aim of this section is to review the literature on discourse structure and investigate the use of specific theories for the representation of argumentation and its structure.

**Previously published material**

Substantial parts of this literature review (Sections 2.2.1, 2.2.5, and 2.3) have been published as an earlier version in [Peldszus and Stede, 2013b] and, partially, in a more compact form also in [Peldszus and Stede, 2016b].

## 2.1 Requirements and desiderata

In our review of the theories of discourse and argumentation structure, we will consider the following requirements which a perfect candidate would all fulfil. Recall, though, that especially the discourse theories have not been developed to represent argumentation structure in the first place. We will therefore also investigate whether the existing structures could be mapped in a way that suits our purpose of representing the structure of argumentation.

- **Inferentiality**: The theory should offer a set of relations between text segments suitable to represent the inferential step from premises to conclusion, as well as those from objections and rejections of objections.
- **Dialectics**: It should be able to represent the dialectical interchange inherent in argumentation, either by explicitly distinguishing dialectic roles or by offering a set of relations fine-grained enough to derive the roles from instances of relations.
- **Compositionality**: Expressions of argumentation can be found in different sizes, ranging from a short justified request to extensive scientific disputes. The theory should thus offer structure building operations that allow to compose larger elements from smaller elements.
- **Non-Linearity**: The connections between arguments can be entangled. For instance, an author could first present several possible objections and later rule them out. In such a parallel structure, the connections between each objection and its rejection would cross each other. A candidate theory should not rule out such instances of argumentation due to linearisation constraints.
- **Long-distance dependencies**: The connections between arguments can have a far reach. An argumentative relation could even hold between the very first and very last sentence of a text. The theory should be able to handle these long-distance dependencies.

Besides these requirements, a few desiderata are of interest. Those are not necessarily required for finding a suitable candidate theory, but they capture aspects that could make a theory even more promising in the context of the goal of this work. If a theory does not meet some of these desiderata, we will not directly dismiss it, but instead investigate whether and how it could be adjusted to our needs. Note that some desiderata cover issues of annotation and automatic recognition, topics which will be introduced more thoroughly in Chapter 4 and Chapter 6 respectively.

- **Text genre independence**: The theory should not be restricted only to one specific text genre. A theory that is for example geared towards describing the argumentation in fundraising letters or product reviews is not necessarily readily applicable to political commentaries. We would first have to verify which parts of the classes, relations, and structures offered by the theory are general enough, which ones could

be generalised to other genres, and which would remain obsolete when applied to a new text genre.

- **Domain independence**: Similarly, parts of the theoretical inventory might be specific to a certain topic, for instance when the theory is mainly applied to reviews *of films* or scientific arguments *in chemistry*. In this case, again, an investigation of the applicability of the offered concepts to other domains or in general would be required.
- **Reliability of annotation**: Apart from subjectively plausible theorising, empirical studies of the replicability and reliability of the theoretical concepts are considerably appreciated. Agreement studies demonstrate how reliable and stable annotators can apply the theory to new and unseen texts and indicate possible overlaps and confusions between concepts. Without a reasonable level of reliability, these concepts are very unlikely to be automatically recognisable.
- **Annotated corpora**: Finally, if reliability is proven, a text corpus annotated according to this scheme is indispensable. It is essential for two reasons: On the one hand, annotated corpora allow the empirical study of the phenomenon of interest, be it qualitatively by searching for specific examples, or quantitatively by analysing its frequencies, co-occurrences, and regularities. On the other hand, the methods for automatic recognition of structures applied here require annotated data to learn from. Not having an annotated corpus at hand implies that we have to create such a resource for our purposes.

## 2.2 Discourse structure

### 2.2.1 Text zoning

A very simple way to structure a discourse is to divide it into several non-overlapping zones, where all sentences or clauses in a specific zone can be associated with the same conceptual category. One such flat partitioning of the text is already given in its logical document structure, e.g. when the text consists of multiple chapters, sections, and / or paragraphs. Other conceptual categories have been proposed, of which we will discuss two: First we will look at functional zones [see also Swales, 1990], in particular a scheme proposed to divide a scientific text into so called *argumentative zones*, and then we will review more content oriented *topic zones*.

#### Argumentative zones

In the *argumentative zoning* approach, a scientific text is divided into different functional zones. The conceptual categories in this zoning are closely related to the idea of a knowledge claim [Teufel, 2010]: a researcher claiming a scientific finding as her own. It thus

captures a core aspect of scientific argumentation in addressing the debate about the central question 'Who found out what?' For example, in a scientific publication, the authors present their own work (their approach, method, resources, experiments, results) and relate it to other researchers' work. Some of the related work is presented as generally accepted knowledge, or as providing the basis for their own approach. Some previous work is presented neutrally, while other might be presented in contrast with or in comparison to the own work, whereby potential drawbacks and weaknesses of previous work are highlighted.

Argumentative zoning aims to assign one functional category to each sentence of a scientific text. It thereby divides the text into zones of multiple sentences with equal function. The scheme distinguishes between seven classes: *Own* for the claimed knowledge, *Background* for generally accepted knowledge, *Basis* for work that the claimed knowledge is based on or supported by, *Other* for neutral descriptions of previous work, and *Contrast* for comparisons and critical reproduction of other work. Aside of the knowledge claim related zones, scientific papers typically also contain a specification of the research goal, which is captured in the *Aim* class, and sections describing the logical document structure of the text, represented by the *Textual* class.

The argumentative zoning scheme has been defined in [Teufel et al., 1999]. Annotations experiments have shown that it can be annotated reliably [Teufel, 1999, 2010]. The scheme has originally been applied to a corpus of scientific articles in the domain of computational linguistics. A more fine-grained scheme has been applied to chemistry articles [Teufel et al., 2009]. Related schemes for functional zones in scientific text have been proposed and applied for various domains: for abstracts and introductions in computer science theses [Feltrim et al., 2006], for full articles in genetics [Mizuta and Collier, 2004], and in astrophysics [Merity et al., 2009]. A functional zoning in the legal domain has been presented by [Hachey and Grover, 2006]. Another approach to functional zones in scientific discourse has been presented in [Liakata et al., 2010]. Here, the zones are based on 'core scientific concepts'. The scheme puts more emphasis on the fine-grained distinctions of methodological concepts such as experiment, model, result, method.

Can the argumentation zoning approach describe the structure of argumentation of a text? Do the representations fulfil the requirements brought forward in Section 2.2? First of all, the structures proposed by the argumentative zoning are not expressive enough to represent the inferential relations between the different sentences. The only relations inherent to a flat-typed partitioning are the linear order and the relation of being in the same zone. The only way a relational structure could be derived is by assuming implicit relations to hold between the different classes of zones, but the specification of those and their usefulness for representing argumentation has not been worked out yet. We therefore conclude the inferentiality requirement is not met. The dialectical participants, on the other hand, are inherent in the strict own work / prior work distinction which is often further explicated by citations. Dialectical role switches might, furthermore, be represented through typical

zoning patterns, such as adjacent *Own* and *Contrast* sentences. Since the structure of a partitioning is non-relational, the last two requirements, non-linearity and long distance dependencies, can also not be met.

Besides our requirements, we see two other obstacles: The first is genre restriction. Even though the scheme has been applied to different scientific domains, it is still a scheme addressing the argumentative zones of *scientific* writing. With the knowledge claim as the central concept, the scheme is restricted to scientific discourse and cannot directly be applied to non-scientific argumentative discourse, such as those in news editorials or product reviews, for instance. We assume, however, that a more general specification of argumentative zones in general text is possible, and a very preliminary step towards this will be taken in Chapter 4.4. The second obstacle is the focus on sentences as the basic unit of investigation. Argumentatively relevant propositions are very often expressed at a sub-sentential level, for example in subordinate clauses. Many relevant argumentative moves might be missed when restricting to sentences.

Nevertheless, argumentative zones in the general sense as a clause- or sentence-wise labelling might serve as a useful feature for a more fine-grained analysis of the structure of argumentation. A very simple approach to enrich structure could be to use argumentative zoning labels as terminal symbols in a phrase-structure like tree representation of an argumentative structure of text. We will see a specification of such grammar of argumentative text later (see Chapter 6.1). Finally it is worth pointing out that the models of automatic recognition of argumentation structure that we will present in Chapter 6.3 are to some extent inspired by the methods used for the recognition of argumentative zones.

#### Topic zones

Another zoning of a text that might be of interest for the argumentation is topic zoning. The topic structure of a text describes what the text is about. The text may have a general topic that is described e.g. in the first two paragraphs, and then the text moves on to discuss other related topics. A topic zoning is thus a partitioning of the text into topics dominating the different parts of the text. Models of topic structure typically exploit vocabulary shifts in the text to identify topic switches, i.e. the move from one topic to a new topic and thus the boundaries of topic zones. What the actual topics of the zones are and how they relate to each other (despite of being different) is usually not captured by those models.

An influential early model of topic zones is the 'TextTiling' approach [Hearst, 1994, 1997]. Here, a topic switch is modelled as a critical change in the lexical similarity between adjacent parts of a text. Lexical similarity is defined in terms of lexical co-occurrence, one of the cohesive devices mentioned above. When two text parts are lexically similar, this is interpreted as a continuation of the topic; when they are dissimilar, a topic switch could be implied. The approach has been tested on relatively long expository magazine texts (with

a length of 21 paragraphs, 1,800 to 2,500 words). For these texts, two observations have been made: On the one hand, they are usually not divided into sections but only organised as a sequence of paragraphs; on the other, topics typically span over multiple paragraphs. One motivation for the TextTiling approach is then that it could identify an additional structuring aspect, the topic zones, and thus serve information retrieval and text indexing.

The topic zoning of the TextTiling approach is surely not meant to be a descriptive theory of discourse structure. However, it is prototypical in defining a technique of structuring text using only what the text consists of: words and their distribution across a text. It is thus in principle domain- and genre-independent and can be easily used to structure new and unseen text. Note, however, that it has been applied to different text genres with varying success [Stede, 2011, p. 37f], as not all genres produce topics as easy to separate as the genre of expository text. Furthermore, the results of the approach are highly dependent on the choice of technical parameters of the implementation (window-width etc.). Finally, one should be aware that the procedure is not easy to evaluate, given that annotators typically find it hard to code topic zones and often disagree on specific boundaries, which lead Hearst [1997] and others to define the true topic boundary by a simple majority vote of the annotators.

Since this early work by Hearst, a considerable amount of research has been conducted on improving the models of lexical cohesion and the boundary prediction exploiting it, today usually grouped under the term 'topic segmentation'. While early approaches modelled the similarity between text parts with simple measures such as word repetition or lexical chains, more elaborate models have been proposed recently including probabilistic models, vector-space models, sequential, and even hierarchical models. A comprehensive overview is given in [Purver, 2011]. Topic zoning approaches have not only been applied to different text genres of monologue text, but also to segment various dialogue transcripts. When evaluated on monologue text, artificial datasets are often used, where unstructured text of different topics (e.g. a set of different news-stories) are concatenated and treated as one text with a topic switch.

Can topic zoning serve as a model for the argumentative structure of text? The obvious answer to this question is 'no'. None of the requirements are met. With the exception of hierarchical topic models, the proposed structure is only a linear segmentation and compositionality is not granted. Furthermore, the predicted zones are typically too coarse-grained for both linear and hierarchical topic models, as they span over multiple sentences or even multiple paragraphs. It is thus very unlikely that it could capture the argumentative shifts and moves on a sub-sentential level, even when the parameters of the model are adjusted to work on smaller windows, and especially when human annotators find it difficult to decide even for the coarse zones. Most importantly: The only relation the model can express is the relation of two adjacent zones being of a different topic. Consequently, the relation

set is neither fine-grained enough to represent argumentative shifts, nor could it capture non-linear and long-distance dependencies.

Nevertheless, the topical structure of text might be indicative of the larger argumentative structure. It might for instance be useful in identifying larger argumentative threads in a longer argument and could thus serve as a feature for automatic recognition of argumentation structure. A first approach using topic structure in argumentation mining has been given in [Lawrence et al., 2014, Lawrence and Reed, 2015].

### 2.2.2 Discourse connectives and local discourse relations

Another line of research which has been very influential in computational linguistics is the study of local discourse relations. The most prominent example is probably the investigation of discourse relations marked the Penn Discourse Tree Bank (PDTB) [Prasad et al., 2008a].

A discourse relation holds between two arguments.[1] Each argument is a text span describing an abstract object [Asher, 1993] such as a proposition, a situation, event, or fact. The PDTB scheme distinguishes two instantiations of discourse relations: Relations that are lexically grounded in an 'explicit' discourse connective, and relations that can be inferred by the reader without being explicitly signalled by a discourse connective. These inferred relations are marked up by the annotators with an 'implicit' discourse connective that could express the inferred relation.

Explicit connectives can be of one of three syntactic classes [Prasad et al., 2008b]: Subordinating conjunctions (such as *because*, *when*, *since*, *although*), coordinating conjunctions (e.g. *and*, *or*), and adverbials (as *however*, *otherwise*) and prepositional phrases (like *as a result*, *for example*). Modified connectives (e.g. even *if*, partly *because*), parallel connectives (e.g. *either . . . or*, *if . . . then*) and conjoined connectives (e.g. *if and when*) are also allowed. Discourse markers that do not relate two abstract objects, such as clausal adverbials and other cue phrases, are not considered. In total, about 100 different explicit connectives are annotated in the PDTB.

The two arguments of a discourse relation are simply labelled 'Arg1' and 'Arg2'. For explicit connectives, Arg2 is the argument binding the connective syntactically, which is why it is also referred to as the 'internal' argument, while Arg1 is the 'free', 'external' argument, which is usually harder to identify. The arguments of implicit connectives simply reflect their linear order in the text. In terms of size, arguments are required to be minimal in the sense that they convey all the information required for the interpretation of the relation, but they are not constrained to be single clauses or sentences. While arguments of explicit connectives are allowed to be distant, those of implicit connectives are required to be adjacent. This requirement is not a theoretical one: It has been raised in order to reduce the

[1] Throughout this subsection, the term 'argument' refers only to the formal objects being related by a relation, and is not to be understood as an instance of argumentation.

annotation load, mainly, and it has been relaxed in comparable annotation efforts [Prasad et al., 2014].

Discourse connectives are often ambiguous. The connective *since* can e.g. have a temporal and a causal reading. In order to be able to distinguish between different readings, all marked relations in the corpus are furthermore annotated with a 'sense' of the discourse connective. The scheme offers a sense hierarchy with three levels of granularity. The first level distinguishes the four classes: temporal, comparison, contingency, and expansion. The second level expands these to 16 categories: E.g. contingency divides into cause, pragmatic cause, condition, and pragmatic condition; comparison divides into contrast, pragmatic contrast, concession, and pragmatic concession. The difference between semantic relations, relating two eventualities 'in the world' and pragmatic relations, relating to speech acts is made on this level.[2] The third level provides even finer granularity for some of the relations.

For analysing argumentation, two groups of coherence relations that are especially relevant: Causal relations (including both the semantic and pragmatic relation) cover argumentative support. Contrastive relations are used for attack and counter-attack configurations. Besides these, additive signals such as *in addition* or *moreover* can play an organising role when several arguments are presented sequentially. These sense groups are, however, not equally likely to be grounded in an explicit signal: Generally, the fraction of coherence relations that are explicit in text is typically reported at roughly 40% (cf. [Stede, 2011, p. 110]). While there are no specific results for argumentative text, we can expect that argumentative support ('causal') relations will often go unsignalled, whereas the attack / counter-attack configurations usually require a lexical signal to allow the reader to identify the contrastive argumentative move. This is, for instance, the case for Concession relations, which are known to require a connective such as *although* or *nonetheless*.

Another annotation level captures attributions, a relation between an attribution holder and an abstract object [Prasad et al., 2006]. The aim of this annotation level was to further minimise the span of arguments, by removing attributing clauses from them. Besides spans for the abstract object and optionally the attribution holder, attributions are classified into four aspects: the source (either the writer, an other agent specified in the text or non-specified arbitrary individuals), the attribution type (assertion propositions, belief propositions, facts, and eventualities), the scopal polarity (is negation involved), and determinancy (can beliefs be cancelled). The annotations of attributions have later been semi-automatically extended to the Penn Attribution Relation Corpus (PARC) [Pareti, 2012]. We will not go into detail here, as the theorising, annotating, and automatically recognising of

[2] A similar distinction between the informational and the intentional level of discourse relations is made by Moore and Pollack [1992]. However, they argue for a multi-level annotation of discourse, as both aspects are often present in one discourse, and annotators should not be forced to make arbitrary decisions for one or the other.

attributions in text constitutes a field on its own in computational linguistics. But we will bear in mind that the attribution level might be able to capture some of the perspective / role switches that frequently occur in argumentation.

The PDTB covers English news texts from the Wall Street Journal, and has over 18,000 relations with explicit and 16,000 relations with implicit connectives annotated. For details about the mid-term agreement studies, see [Miltsakaki et al., 2004]. Similar annotation projects have been undertaken, or are still in progress, for other languages including Arab, Czech, Chinese, French, Hindi, and Turkish, but also for other text genres as for example biomedical scientific papers; see [Prasad et al., 2014] for references. One interesting study of discourse connectives and relations in argumentative student essays has been presented by Forbes-Riley et al. [2016]. Attributions have only been annotated in the original English PDTB. The PDTB is one of the largest resource for studying discourse relations in English and has attracted a lot of attention also from the modelling perspective, aiming for automatic sense disambiguation of discourse connectives, extraction of their arguments, and identification of implicit discourse relations. Just recently, automatic discourse relation identification was chosen as one CoNLL Shared Task 2015 [Xue et al., 2015] and 2016 [Xue et al., 2016].

For German, there are similar but much smaller resources for discourse connectives and their arguments: Besides already existing annotations of syntax, coreference, and rhetorical structure the Potsdam Commentary Corpus (PCC) offers also explicit connectives and their arguments in its second release [Stede and Neumann, 2014]. Connectives are taken from a rich lexicon of 274 discourse connectives [Stede, 2002]. Around 1,000 discourse relations are marked in the 176 newspaper commentaries available. In the TüBa/DZ newspaper corpus, around 1,300 instances of seven connectives with multiple readings (mostly temporal versus comparative/contrastive readings: *nachdem, während, sobald, seitdem, als, aber, bevor*) have been annotated [Simon et al., 2011]. Another annotation layer in that corpus contains about 1,400 explicit and implicit discourse relations [Gastel et al., 2011]. About 1,100 causal connectives, their arguments, and the illocutionary type of the arguments have been annotated in a small corpus of hotel reviews [Peldszus et al., 2008, Stede and Peldszus, 2012].

The discourse connectives and their relations, as they are annotated in the PDTB and related resources, promise to be a useful representational device for describing the relations frequently found in argumentation. The scheme offers a rich and fine-grained set of relations, a subset of which will be adequate to represent the inferential relations inherent in argumentation (Cause, Result, Condition, Concession, Contrast). The first requirement is therefore met.

The dialectical dimension of argumentation cannot be captured directly. Although the scheme offers an extra annotation level for encoding attributions, the annotation of attributions is focused on agents and attributing relations found explicitly in the text, and only

1% of non-writer sources are implicit 'arbitrary individual' sources.[3] The switching of perspective in argumentation is, however, often expressed more indirectly, without reference to specific opinion holders. Nevertheless, it might still be possible to derive role switches from typical patterns of relations of adjacent segments.

With the notion of structure, we arrive at a crucial question: What kind of structures are promoted when analysing discourse connectives and their relation? The answer is: only the *local* relational structure of one connective and its arguments. The PDTB scheme is intentionally non-theoretic towards the question of larger discourse structure. Instead of enforcing a specific discourse structure, it is geared towards providing the basic elements required to investigate and evaluate the specific claims of theories of discourse structure. Not only is the scheme neglecting structure-building operations to form larger units from smaller units, the annotation procedure also assumes a certain non-dependency between the relations: All relations are marked on their own without constraints of existing or possible relations in the vicinity. For example, multiple explicit connectives and multiple implicit connectives with equal spanning arguments are allowed; only the insertion of implicit connectives concurrent to existing signals is prohibited, see [Prasad et al., 2008b, p. 19] and [Prasad et al., 2014, p. 925]. There is no common discourse segmentation or unitising governing all relations in a text. Also, there are different configurations of overlapping relations: While embedded and nested relations could be interpreted as tree-structures, other configurations such as shared arguments between two relations, properly contained arguments, or relations and crossing relations might require more complex structures [Lee et al., 2006, 2008, Demirsahin et al., 2013]. An approach that derived tree structures from PDTB-relations was reported by Zhang et al. [2016]. Note, however, that the derived structure is not intended as a discourse representation, but rather serves as a vehicle to improve downstream applications.

The requirement for long-distance dependencies is unlikely to be met. The large majority of relations are between adjacent arguments: all relations of implicit connectives, as well as 91% of the relations marked by explicit connectives. Only the remaining 9% of the external Arg1s of explicit connectives are found in a previous non-adjacent sentence, the actual distance between them not being reported in [Prasad et al., 2008a]. The requirement for non-linear representations is fulfilled, as crossing relations are possible in this approach. Note, however, that this is not due to a structuring principle less restrictive than linear trees, but due to the absence of a structuring principle.

Let us conclude this discussion of connective-centred accounts of discourse relations: We found a theory providing us with a fine-grained set of discourse relations either signalled by explicit connectives or inferred as if signalled by an implicit connective. While the dialectical dimension of argumentation could be accounted for indirectly, the main obstacle for using it

[3] See `http://compprag.christopherpotts.net/pdtb.html#attdist`

as a representation of argumentation structure is the intended lack of commitment towards a structuring principle.

### 2.2.3 Intentional and illocutionary accounts of discourse structure

When an author writes an argumentative text, she wants her text to have an effect on the reader. Her intention is to persuade her audience to accept her claim, e.g. to believe that something is the case or to accept the need for some action.

If the text as a whole has a purpose or exhibits an intention, then maybe the parts of the text have their own purposes that facilitate achieving the main purpose. To make the reader believe a proposition, the author might first need to present some evidence. Or to make her believe some action should be taken, she might first need to convince the reader this action will lead to a desirable situation. This decomposition of the overall text purpose into smaller ones might be governed by structuring principles.

The text linguistic and computational linguistic literature has brought forward several intentional approaches to discourse structure. In this subsection, we will briefly revisit these theories and evaluate their usefulness for representing the structure of argumentation.

#### Speech acts, illocutions, and intentions

Before we review the principles that allow building larger units from smaller ones, let us begin by investigating what the elementary units are that constitute such a discourse structure. What is the author's intention when producing this segment of the text, what does she want to 'do' with it?

Our starting point here is Austin's theory of speech acts [Austin, 1975]. His central claim is that there is much more we 'do with words' than just stating facts that describe the world, yielding sentences which are either true or false. Instead, whenever we say or write something, this can be understood as an action that we intend to be understood and to have an effect on our audience or the world. According to his theory, a speech act can be divided into multiple acts: the *locutionary act*, which is the act of producing a meaningful sentence; the *illocutionary act*, which captures the communicative function of the sentence intended by the producer, i.e. what she is doing *in* producing this sentence; and finally the *perlocutionary act*, which describes the effect or the psychological consequences that this sentence has on its audience.

Austin's theory may not have been as influential in linguistics, had it not been Searle's bridging speech act theory and formal semantics. Searle [1969] proposes a fine-grained analysis of the semantic aspect of the locutionary act, which is in his terminology the *propositional act*. In correspondence to the functional formalisation of reference and predication in the formal semantics tradition, he distinguishes a reference act (of referring to some

entity) from the predication act (functionally applying a predicate to an entity). The proposition of Sam smoking habitually is for example analysed as the act of referring to Sam and applying the predicate of smoking habitually to the referred entity. In a second step, he formalises illocutionary acts as a function of propositional contents [Searle, 1976]: The same propositional content, in his example Sam smoking habitually, can be used in different illocutionary acts, e.g. in *representatives* (reporting that Sam smokes habitually), in *directives* (demanding that Sam smoke habitually), *commissives* (committing to Sam smoking habitually), in *expressives* (e.g. regretting, that Sam smokes habitually), or finally in *declaratives* (although declaring that Sam smokes habitually is of course not a felicitous declaration, because the act of declaring it does not make it true).

Note that both Austin and Searle focused their work on rather 'ideal' sentences. This has two consequences: First, the notorious question of segmentation of authentic text was not tackled. The minimal units in their analysis are usually simple sentences without subordinated or coordinated clauses, parentheses, ellipses, etc. When working with authentic text, we must find criteria of deliminating minimal units that are capable to have an illocution. Furthermore, the speech act theoretical analysis is one of sentences, not of sentences of a *text*. The interactions and regularities between the illocutions of a text and the question of what makes a text coherent in terms of the illocutions expressed in it has not been studied by Austin or Searl.

The influential work of Austin and Searle has stirred up a plethora of follow-up investigations on speech acts, including different refined taxonomies of illocutions. One such refined taxonomy of illocutions has been proposed by Schmitt [2000] with the aim to be applicable to texts of various genres and domains. The taxonomy includes reportives, estimatives, evaluatives, identificatives, representatives, interrogatives, directives, commissives, declaratives, and 'relationata'. This set of illocutions is then tested for applicability in an exemplary analysis of 28 German and English texts of varying genres, including speeches, instructional text, live sport reports, letters, holiday postcards, advertisement, sermon, announcements etc. In total, about 560 German and 200 English illocutions have been identified.

Where Schmitt contrasts different text genres and domains seeking for a general, applicable scheme, Dillmann [2008] investigates the differences between languages by comparing the distribution of illocutions in German and Japanese commentaries. For this purpose, he collected a corpus of 100 German and 100 Japanese newspaper commentaries which are annotated with ten different illocution types. In total, the corpus contains about 4,900 German and 4,200 Japanese instances of illocutions. One of his findings shows that Japanese commentaries had a significantly higher rate of 'objective' factual reporting illocutions compared to 'subjective' evaluatives or estimatives than German commentaries. Furthermore, this work features a comprehensive comparison of various taxonomies of illocutions.

Corpora of authentic, non-conversational text annotated with illocutions are very rare. Besides the corpora of Schmitt [2000] and Dillmann [2008], we are only aware of the small

corpus of hotel reviews [Stede and Peldszus, 2012]. There, about 2,350 segments have been annotated for illocutions, following Schmitt's tagset. However, only segments that were arguments of causal connectives have been annotated in the corpus, so the majority of segments that are not being causally related remain unmarked and a full illocutionary classification of the text is not provided.

In computational linguistics and the study of discourse, the speech act theory was very influential in the dialogue community, where different schemes for analysing, annotating, and automatically classifying dialogue acts in conversation transcripts have been proposed. For the study of speech acts in text, however, there is much less work. Most of the approaches work with written conversations as found in emails, message boards, or forums [see Jeong et al., 2009, as one example]. Professional email exchanges could indeed be interesting from an illocutionary perspective, as they might exhibit many directives and commissives. However, as in [Cohen et al., 2004], emails are often analysed as one act without considering the individual illocutions expressed in the sentences. Often, the tagsets are focusing more on the interactional aspects of written conversation (questioning, answering, clarifying, positively or negatively acknowledging) than on illocutionary forces [as for example Arguello and Shaffer, 2015, for MOOC forum posts]. One exception is the work of Qadir and Riloff [2011], where sentences have been reliably annotated and automatically classified into four of Searle's illocutionary acts (expressives, directives, commissives and representatives).

Let us recapitulate to this point: It appears evident, that illocutions alone will not be able to represent argumentation structure. A simple tagging of text segments with illocutions, i.e. a flat, sequential structure will not be able to describe the argumentative relations between text segments. The requirements formulated above are thus not met. Also, there is of course not a simple mapping from illocutionary types to argumentative segment types such as premises, conclusions, and objections. For example, the central claim of a text should be something disputable, which rules out acts of informing or reporting non-contentious facts; but all other other illocutions are perfectly adequate for a central claim – be they directive, evaluative, estimative, commissive. Similarly, argumentative opposition can take all different illocutionary forms (evaluative, estimative, commissive, expressive, factual), excluding perhaps only directives and declaratives.

Still, illocutions might be helpful in determining the argumentation structure: The occurrence of some illocutions might serve as an indicator that there is something argumentatively relevant going on. Furthermore, illocutions might help to identify relations between these text segments. Note that these assumptions are all subject to a systematic, empirical study on a corpus that is annotated with both illocutions and argumentation structures. Unfortunately, no such corpus of monologue text is yet available.

### Constituency structures

We now move from approaches of classifying the purpose of a text segment into illocutionary types and and proposing a flat, sequential characterisation of the illocutionary structure of text to approaches employing hierarchical structures, i.e. to a description of how the author's intentions for individual text segments fuse to larger complexes, corresponding to the purpose of text parts or even of the entire text.

One of the first formulations of an intentional discourse structure is the theory of Grosz and Sidner [1986]. Although the theory was mainly applied to task-oriented dialogues, it is still interesting for us here, as the seminal paper also features an example analysis of a monologue argumentative text. Grosz and Sidner propose three different levels for describing discourse structure: In the linguistic structure, multiple subsequent elementary discourse units (EDUs, i.e. utterances, sentences, or clauses) are combined into *discourse segments*. The intention corresponding to a discourse segment – in their terminology the *discourse segment purpose* – is represented in the intentional structure. Finally, the attentional state describes the focus of attention of the reader or dialogue participant when processing the ongoing discourse.

The linguistic structure has an embedding relation. A discourse segment can consist of one or more EDUs, but also of further discourse segments. This produces a hierarchy of segments. Linguistic cues such as discourse markers or shifts in sentence mood, tense, or aspect serve as indicators for possible segmentation boundaries. Grosz and Sidner do not directly define what it means when one discourse segment embeds another, but characterise it indirectly with the strict correspondence of linguistic and intentional structure: Each discourse segment is associated with a purpose in the intentional structure, represented as e.g. the intention of the author to make the reader believe the proposition expressed in the segment. Discourse segment purposes may contribute to the satisfaction of another discourse segment purpose. In this case the first purpose is *dominated* by the latter. This domination relation imposes hierarchy on the discourse segment purposes, which is assumed to be parallel to the linguistic structure of discourse segments.

In addition to the subordinating relation of dominance, Grosz and Sidner also postulate the coordinating relation of *satisfaction-precedence*. This relation explicitly encodes the order in which the purposes have to be satisfied. While the precedence is more important in the study of task-oriented dialogues or instructional text, where there are constraints on what to do when, the example analysis of an argumentative text does not feature any instances of the satisfaction-precedence relation. However, as pointed out by Stede [2007, p. 108], more complex argumentative texts might require a certain linearisation in order to be effective.

As a result, both linguistic and intentional structures are constituency trees, which are isomorphic due to the direct correspondence between discourse segments and their pur-

poses. The parallelism is used for the recognition of discourse structure: Discourse segment boundaries signalled by linguistic cues might help inferring a relationship between the corresponding intentions, while domination relations between intentions can help to infer boundaries between discourse segments.

Note that the hierarchy is not strict in the sense that a discourse segment only consists of either other discourse segments or EDUs. For example, one discourse segment could consist of an EDU, a discourse segment, and another EDU. This allows for center-embedding structures. Grosz and Sidner discuss three different types of interruptions to the discourse flow: true interruptions of other unrelated discourse; flashbacks for introducing topics or entities that should have been introduced earlier; and digressions to related topics that do not satisfy the purpose of the current segment. While true interruptions and flashbacks are unlikely to occur in written argumentation, digressions can be found in authentic argumentation, e.g. when the author takes sideswipes at her opponent.

A similar constituency structure is proposed by Schröder [2003] in his theory of the structure of action in text: He distinguishes between sentential acts on the terminal node level of discourse trees, intermediary, and finally overall textual acts for non-terminal and the root node respectively. Schröder assumes four relations, one subordinating relation and three coordinating relations. Subordination is understood as specification of a partial aspect of an action: One action is realised *by* another, for example the author reports on an event *by* citing the official notice *by* informing about content of the official notice. Coordination is described as an *and then*-relation. Three different types of coordinating relations are distinguished: supplements, enumerations, and continuation.[4]

The sentential acts are intended to be very basic, without all the fine-grained facets brought up in the literature on types of illocutions. Since the analysis is applied to reporting text, not to commentaries or other opinion pieces, most of the sentential acts are inform acts, a sort of assertion. This directly corresponds to Grosz and Sidners typical discourse segment purpose in argumentative text: the intention of making the reader belief a proposition expressed in the segment.

More fine-grained illocution types are introduced at the constituency level, i.e. as intermediary or overall textual acts. Although Schröder's analysis is mainly focused on reporting text, featuring different types of acts of factual reporting, there is also a discussion of analysing reports [p. 233ff]. These reports are closer to commentaries in that they do not only notify about events but also interpret and evaluate them in a larger context. The textual acts in this genre might also cover claims and trigger further argumentative acts of justification. Unfortunately, due to the scope of the analysis, no detailed specification of textual acts of interpretation and commentary or even argumentation are presented.

[4]The original German terminology for subordination is 'indem'-Beziehung and for coordination 'und dann'-Beziehung.

Lets us draw an intermediary conclusion: Is a constituency structure of illocutions or intentions capable of representing all desired aspects of argumentation? The requirement of inferentiality is not met. Both theories could represent a hierarchy of supporting relations, but cannot capture the inferential regularities of objections and rejections of objections with the presented set of intentions and textual acts. For Grosz and Sidner's theory, a corresponding extension would include the specification of intentions besides the inform act, and the domination relation could not always be interpreted as a supporting relation between propositions. For Schröder's theory, the sentential and textual acts necessary for representing the commentary, interpretative, and evaluative text need to be determined. Since both theories are not geared towards representing opposition, the requirement of dialectics is currently not met. A richer intentional respectively illocutionary inventory might, however, allow a determination of the dialectical stance in retrospect.

The constituency structures proposed are compositional, but do not allow non-linear relationships between segments, even though Grosz and Sidner's structures allow center-embedding. Long-distance dependencies between segments cannot be captured directly with a constituency tree. One way to at least indirectly represent some possible long-distance relations would be to apply the nuclearity principle as in RST. Roughly the idea is as follows: For every constituent, we know which of its children is the most important one. Then, we can say that whenever a segment and a constituent are siblings, the sibling is in relation with this most important child of the constituent. We will discuss in more depth below in Section 2.2.5.

To our knowledge, no corpus of monologue text annotated with intentional or illocutionary constituency structures is available. Grosz and Sidner's theory has been applied only to task-oriented dialogue. Schröder's examples are from a source corpus of 320 news reports. He discusses 18 prototypical texts in full length, but without providing full annotations of the illocutionary structure of the text. Similarly, we did not find any agreement studies on annotating this type of structure.

#### Dependency structures

Besides accounts of illocutionary structure that represent this aspect of discourse as a constituency tree, there are approaches using a dependency structure. In such structures, the segments of the texts are the nodes of a dependency graph and each edge represents a relation holding between two segments. Instead of specifying the structural complexity by proposing intermediary node types, as e.g. the complex intentions in Grosz and Sidner's theory or the textual acts in Schröder's theory, a dependency structure specifies as set of immediate relations between segments.

A theory of illocutionary structure based on dependency structures is presented by Brandt and Rosengren [1992], who investigate complex directives in written business communi-

cation. Again, a hierarchical and a sequential structure are assumed. The hierarchical structure specifies the dominance relations between text segments; the sequential structure defines the order in which the segments occur in the text. Each text segment has an illocutionary type. The authors base their set of illocutionary type on Searle; they distinguish declarations, assertions and questions, expressives, directives, commissives, and permissions and asking for permission. In their analyses, however, they also use more specific acts that such as acts of hoping, thanking, showing interest.

The dependencies between text segments specify the function that the dominated act has to ensure the felicity of the dominating act. All functions are understood to provide some sort of support to the dominating act. Brandt and Rosengren distinguish between functions of immediate ('subsidiary') and intermediary ('complementary') support. Functions of immediate support are analysed on an axis of communicative success: first, functions that support by enabling the addressee to understand the intended illocution, e.g. when the illocution is clarified or made explicit; second, functions that support by increasing the acceptability of the dominating illocution, e.g. when providing evidence for a claim or a reason for a demanded action; third, functions that support by motivating the addressee to execute the demanded action; and finally functions that support by enabling the addressee to execute the demanded action, e.g. by providing required information. For intermediary functions, Brandt and Rosengren propose functions of clarification of the topic or the matter of business, e.g. by providing background information or referring to prior communications, and functions of ensuring cooperativity, e.g. expressions of courtesy such as greeting and thanking.

The theory does not require a single, global structure across the text. Brandt and Rosengren show example analyses where they identify multiple directives in one text, each having its own illocutionary structure, without specifying any domination relation that would make one directive the dependent of another. Multiple structures might also share dominated segments. Furthermore, multiple text segments can be be combined to a complex segment, which have an illocution when interpreted as a whole. Beyond the exemplary analysis of 14 short texts in the paper, we are not aware of an available annotated corpus.

The hierarchical approach of Brandt and Rosengren was later used by Schmitt [2000], who was already mentioned above for his refined set of illocutions. He criticises several aspects of their original approach, most importantly the lack of systematicity in segmenting the discourse into elementary units that could have an illocution, and the unconstrained set of illocution types.

In his exemplary analysis of 28 German and English texts, Schmitt not only identifies and classifies illocutions, but also analyses the illocutionary structure. Approximately 100 relations were marked and classified into the illocutionary functions of Brandt and Rosengren. About 10% of these functions involve complex segments, e.g. when multiple segments support another or multiple segments are supported.

Another example of dependency-based illocutionary structures is the analysis of Motsch [1987]. This work builds on a theory of illocutions presented in Motsch and Pasch [1984], where they distinguish between eleven different types of illocutions, and already sketch the idea of a global structure of illocutions with coordinating relations and three types of subordinating support-relations (ensuring understanding of an intention, increasing the acceptance of an intention, or enabling the execution of an intended action). Focusing on texts which contain mainly declarative sentences, Motsch [1987] then presents a refined taxonomy of declarative illocution types and exemplifies this in a detailed analysis for a longer text, including a specification of its illocutionary structure. The text segments are classified into illocutionary types: Due to the informing character of the text these are mostly acts of informing, as well as a few claims and constatives (i.e. redundant statements of uncontroversial propositions).[5] The illocutionary functions used in this analysis are not explicitly defined, however. The illocutionary structure itself is quite rich and includes several arguments, yet does not cover argumentative objections or counter-objections. Especially interesting with regards to argumentation is the analysis of the sequential structure of illocutions in Motsch [1996]. Motsch presents a detailed analysis of the support structure for directives in business communication and evaluates the felicity of different linearisations of the same illocutions, discussing the indicators needed to mark entangled structures. This analysis is more oriented towards argumentation, as it features reconstruction of enthymemes, i.e. of presupposed premises of the arguments, and includes adversative relations into the set of illocutionary functions. Still, the argumentative analysis is only in an early stage: The segmentation is not consistently fine-grained (clauses or phrases representing a reason are split off sometimes, but not always) and the adversative relation is only used for a semantic contrast between two propositions, not for an argumentative opposition. Furthermore, it is only exemplified in two short texts.

Finally, a similar structure is exemplified in the analysis of a newspaper commentary in Lenk [2011]. Again, text segments are the nodes in a dependency graph; the linear order of the segments determines the sequential structure. Each node has one or more illocutionary types. Dominance relations are interpreted as support. Instead of defining different subtypes of supporting functions, Lenk distinguishes between dominance and vague dominance. Additionally, there is an even weaker dependency of being 'related' with regards to content, and the symmetric relation of being in contrast, again only semantically. The scheme also allows conjunctions of two segments with individual illocutions to a complex, that is then addressed in the structure. Also, Lenk identifies one segment as the central

[5] Note, however, that the analysed text's more general intention is to provide the reader with evidence and arguments, enabling her to agitate against some issue. Motsch himself is aware of the problem that segments may only be presented by the author as factual information, although they might be highly controversial [Motsch, 1987, p.65].

illocution of the text and thus assigns one global structure, in contrast to multiple parallel structures in the analysis of Brandt and Rosengren.

Let us conclude with an appraisal as to whether the presented approaches of dependency structure of illocutions conform with our requirements and desiderata to represent argumentation structure. The inferentiality requirement is not met, for the same reason as in the case of constituency structures: The dependency relation types used in these theories cover different kinds of supporting relations, but are not capable of representing the opposition when presenting and arguing against possible objections. Lenk introduces symmetric contrast relations, which is used in his example, however, only for representing semantic contrast. Dialectical roles are not made explicit in any theory. Extending the set of relations correspondingly could of course address this shortcoming and at least allow an indirect reconstruction of dialectical roles based on the sequence of attacking relations. The structural requirements, however, are all met: Dependency structures are compositional. They allow non-linearity, since edges can cross. Finally, long-distance dependencies are easy to handle, as nothing forbids a dependency spanning from one end of a text to the other.

Considering our desiderata of text genre and domain independence, only the account of Schmitt has really been applied to different texts, while the others are focused on their text genre (business communication for Brandt and Rosengren, and news commentaries for Lenk). Unfortunately, no scheme has been tested for reliability in annotation experiments. Annotated corpora are not available – perhaps with the exception of Schmitt, whose data are available only in the printed text of his dissertation, not in a machine-readable distribution.

Concluding this subsection, we could say we were *nearly* there. An analysis of the illocutions in an argumentative text provides us with a general description of the communicative purposes of the different segments of the text. But, as we argued, this would not be enough to determine the structure of argumentation. Hierarchical structures of illocutions have been proposed – constituency and dependency structures – and we argued that dependency structures fit out needs better, as they allow for long-distance-dependencies and non-linearities between the segments. The relation sets provide fine-grained distinctions of support, but do not cover argumentative attacks and are thus not able to (directly or indirectly) represent the dialectical roles involved in argumentation. Here, the theories would require refinements. Finally, there is rarely any annotated corpus available with illocutionary structures, nor is there experimental evidence for the reliability of the corresponding schemes.

### 2.2.4 Syntactic accounts of discourse structure

Some accounts of discourse structure have been influenced by syntactic theories. We will consider two examples. The first is LDM, which largely focuses on the structure building of

discourse trees, without assuming an intentional or illocutional analysis of the discourse. The other is D-LTAG, which extends a model of syntax to represent relations across sentences.

Polanyi and Scha [1984], Polanyi [1988] proposed the 'linguistic discourse model' (LDM), aiming for a formal model of the structure of both monological and dialogical discourse. They envisage this general model to be the basis on which different phenomena involved in understanding discourse (such as anaphora, temporal, and deictic expressions) could be described.

According to their theory, the structure of discourse is represented by a constituency tree. Clauses representing propositional meaning are the elementary units, as well as discourse operators. As in previous approaches using tree-structures, the two types of formation principles are conjunctive structures and subordinate structures. Sequential relations, such as in lists, topical, or temporal chains are represented by coordination of discourse constituents, while dominance relations, such as for causality or elaborations, are constructed as subordinate constituents. Finally, so-called 'n-ary' structures are used to represent fixed-sized, conventionalised constructs, such as if-then conditions.

Most importantly, their model focuses on the incremental process of constructing these discourse trees, and sketches a paradigmatic discourse parser. This parser consumes the clauses of the discourse in a left to right fashion and integrates the current clause into the growing discourse tree. Not all existing nodes are accessible to attach a new clause to, though: The *right frontier constraint* regulates that only the last node and all nodes dominating it are available for attachment.

Whether and how a clause is attached to one of the available nodes is determined by semantic rules, which compare the semantic representation of the clause and the possible attachment site. Semantics are specified as higher order logic representations. If for example the logic predicate in the current clause can be seen as a subtype of the logic predicate in the preceding clause, this would be a valid condition of an elaboration, which in turn would trigger a subordinate structure. Attachment sites on the right frontier are tested from the leaf nodes (the last clause) up to the root, so local attachment is preferred. If no semantic rule applies, the current clause is subordinated locally by default.

A more recent presentation of the theory [Polanyi et al., 2004b] elaborates on segmentation rules, separates intra-sentential and inter-sentential discourse parsing processes, and presents a more detailed description of the attachment rules. Based upon this, a discourse parser was implemented [Polanyi et al., 2004a] and used for the downstream tasks of text summarization of technical reports. In addition, the use of LDM-theoretic discourse structures for modelling sentiment in film reviews was discussed [Polanyi and van den Berg, 2011], which may involve a certain degree of argumentation.

Could a LDM structure serve as a representation of argument structure? First of all, for our requirement of inferentiality, we would need a set of discourse relations that could

be mapped to argumentative relations. These may be defined in a LDM grammar and its semantic rules, but additional work would be required, as we have not found a grammar detailed enough to clearly separate supporting, attacking, and non-argumentative relations types. Given that these relations types are well separated, the dialectics could again be derived.

In terms of structure, the tree building principles are clearly compositional. They do not, however, allow for non-linearity, as this would violate the right frontier constraint. As in previous constituency tree based approaches, long-distance dependencies could only be indirectly represented if the siblings receive different importance, as we will see in the RSTs nuclearity principle below.

From our desiderata, we find genre and domain independence provided; however no corpora or reliability studies have been published to the best of our knowledge.

We now move on to the discourse structure theory of D-LTAG, a lexicalised tree-adjoining grammar for discourse that was initially presented in [Webber and Joshi, 1998, Webber et al., 1999]. The theory aims to link the sentential grammar with its syntax and semantics to the grammar of discourse with its own syntax and semantics, and transfers concepts established in sentential grammar to the discourse level.

D-LTAG identifies three mechanism which drive the syntax and semantics of discourse. The first is composition: The grammar features a set of initial trees, which define basic structural complexes for coordinate and subordinate structures. They have two open slots for the segments to be related and one for a discourse connective. Once a connective is found that is considered to trigger this elementary structure (such as e.g. the subordinating connective 'because'), the initial tree is instantiated and integrated into the overall structure. These trees are interpreted to have a compositional semantic. The second mechanism is that of inference. When two adjacent segments are not related by an explicit discourse connective, an implicit discourse connective can be inferred. The corresponding operation to integrate the next segment into the discourse structure is the adjunction of an auxiliary tree.

Until now, by composition and by adjunction, a constituent tree has been built to represent the discourse structure. The authors, however, stress that a third mechanism of building discourse structure exists, which introduces additional relations that can violate the tree principles: They identify certain connectives, especially adverbial such as 'for example', which should rather be interpreted anaphorically as bearing a presuppositional semantics. For these relations, the antecedent has first to be identified by semantic means.

Could a D-LTAG analysis be useful in representing argumentation structure? A first obstacle would be that the theory does not specify an inventory of discourse relations, but that every relation is connective-based. We already discussed above that discourse connectives often need to be disambiguated. Given that we had disambiguated connectives, it would be necessary to group them into signals of argumentative relations such as support or attack

in order to fulfil the requirement of inferentiality. The dialectics would then be required to be derived from the sequence of support and attack relations.

Structurally, D-LTAG is a rich approach. The full representation of the discourse with both structural and anaphorically triggered relations is a potent graph structure, which is compositional but also allows for non-linear and long-distance relations by anaphoric interpretation. However, in order to derive this full structure an extensive semantic reasoning is required. We argue, that for the purpose of representing 'just' the argumentation structure, the necessary reasoning is too fine-grained. This becomes evident for instance when considering the different interactions between discourse adverbials and explicit or implicit structural connectives identified by Webber et al. [2003]. Without the anaphorical relations, however, the resulting tree structure would be quite uninformative as a representation of argumentation, as it would be missing many argumentative relations triggered by discourse adverbials.

The desiderata of genre and domain independence are well met. In terms of corpora, however, we are not aware of any available resource focusing on D-LTAG structures. Note that when the work on the PDTB started, D-LTAG was considered the underlying theory, an assumption which was later changed. As described above in Chapter 2.2.2, PDTB was then built up as a theory neutral resource that focuses on local relations rather than on larger discourse structures.

### 2.2.5 Rhetorical Structure Theory

One approach to discourse structure that obviously is a candidate for representing argumentation is Rhetorical Structure Theory (RST) [Mann and Thompson, 1988]. According to RST, the structure of the text is described by coherence relations connecting text segments in a global tree structure. RST has been conceived as an empirical tool for practical text analysis, and the developers originally justified their design decisions with the claim that a fairly large number of texts from different genres had been successfully analysed. Moreover, RST is geared towards pragmatic description, since the definitions of coherence relations make reference to the underlying intentions of the speaker or writer. For these reasons, one could argue, RST might constitute an adequate framework for the representation of argumentation structure. In the following, we first provide a brief outline of the main ideas of RST and critically discuss work by researchers who proposed extensions of RST for representing argumentation structure. Then we will evaluate the appropriateness of the original and the extended proposals, given our requirements and desiderata.

As for most other theories of discourse structure, the central notion for RST is that of a *coherence relation* [Hobbs, 1979], i.e., the idea that adjacent spans of text stand in a semantic or pragmatic relationship to one another, such as causality or contrast (see also Section 2.2.2 on local structures of coherence relations). This plausible intuition then needs

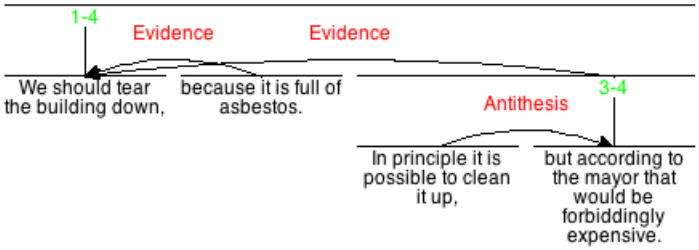



Figure 2.1: RST analysis for a short text.

to be operationalised, and this is one point where theories differ: Is there a finite, and reasonably small, set of such relations so that 'any' text can be analysed in this way? How are coherence relations to be defined? Mann and Thompson [1988] proposed a specific set of 25 relations, yet they pointed out that the set should be regarded as in principle open to extension. However, the manifold practical uses of RST over the past three decades (see [Taboada and Mann, 2006]) have shown that the relation set can in fact be regarded as relatively stable. It contains, for example, several adversative, causal, temporal, and additive relations.

The main criterion for judging whether a relation holds between two segments is the reconstruction of the writer's intention: Why did he state the two segments, why did he place them in adjacency, what did he want to achieve with this combination? In this regard, a major claim of RST is that for the vast majority of relations, the two segments do not have equal status for realising the underlying intention. Instead, one segment (called 'nucleus') is the central one, whereas the other ('satellite') plays only a supportive role. Mann and Thompson point out that when all satellites are removed from a text, the main message can still be reconstructed (notwithstanding certain problems in cohesion, i.e. information flow). When nuclei are removed, however, the text will be simply incoherent. As an exception to the general rule, there are a few multi-nuclear relations that do not make this distinction – an example is the temporal relation Sequence, which applies to segments where one event is described to take place after another event.

For illustration, Figure 2.1 shows an RST analysis of a short text. It follows the notation suggested by Mann and Thompson and was produced with the RSTTool software.[6] Curved lines connect a satellite to a nucleus, with the arrowhead pointing to the nucleus, which is also indicated by a vertical line. Horizontal lines demarcate a larger segment arising from the fusion of smaller segments. Despite the unusual notation, the structure is in effect a tree. In RST, the main structural claims are that only adjacent segments can be connected by a relation, that the complete text needs to be covered by the analysis (there are no gaps), and that no crossing edges arise. (For a more thorough formal analysis of RST's predictions,

[6] `http://wagsoft.com/RSTTool`

**Relation:** EVIDENCE
**Constraints on nucleus:** reader might not believe nucleus to a degree satisfactory to writer
**Constraints on satellite:** reader believes satellite or will find it credible
**Constraints on the N+S combination:** reader's comprehending satellite increases reader's belief on nucleus
**Effect:** reader's belief on nucleus is increased
**Locus of the effect:** nucleus

Figure 2.2: RST-Definition of the 'Evidence' relation, following Mann and Thompson [1988].

see Marcu [2000]). The same set of coherence relations, therefore, is used to describe the relationships between the small segments (minimal units of the analysis) and recursively of the larger text segments.

A relation is defined by constraints on the nucleus, satellite, and on the combination of both, as well as by the intentions of the writer. As an example, Figure 2.2 shows the definition of the Evidence relation as provided by Mann and Thompson. In general, the descriptions of constraints and intentions can refer to a variety of semantic and pragmatic aspects and in this sense do not constitute a very systematic framework. Clearly, they appeal to the intuitions of the analyst, and in the end the RST tree for a text will be the representation of one possible interpretation of a text. Nonetheless, when sufficiently strict annotation guidelines are provided, it is possible to achieve acceptable agreement among human analysts (e.g., Carlson et al. [2003]).

Another important characteristic of RST to be mentioned here is the distinction between 'subject-matter' and 'presentational' relations. The former refer to relationships that hold in the world and are merely being reported in the text; in these cases the intention of the writer is of the form 'Reader recognises that X'. As an example consider this causal relation: *Tom's train was delayed, and therefore he didn't make it to the meeting.* It describes the a relation between two events, which is why this is also referred to as a 'semantic' relation. Presentational relations, on the other hand, are employed by the writer to actually change the beliefs or attitude of the reader. The Evidence relation, shown above, is an example. The causal relation here is a 'pragmatic' one, since it holds between speech acts. Another one is Motivation, where the intention is 'reader's desire to perform the action described in the nucleus is increased'. A very similar distinction between the informational and the intentional level of discourse relations is made by Moore and Pollack [1992], see the discussion above. Obviously, it is the latter family of relations – the presentational, pragmatic, or intentional – that proves particularly relevant to representing argumentation structure.

Finally, we emphasise an observation made by Marcu [2000], which is important for interpreting RST trees, and which has gained wide acceptance in the community. The 'compositionality criterion' states that when a relation holds between large segments, it also holds

between its 'most important units', where importance is defined by a maximum degree of nuclearity. When starting at the root node of the tree representing the segment and following only nucleus lines downward, one ends up at the most important elementary units of that segment. As a consequence, when applying this to the whole text, one is supposed to find the central statement(s) of the text.[7]

Several corpora exist that are annotated with rhetorical structures: The most famous one is probably the RST Treebank [Carlson et al., 2003] consisting of articles from the English Wall Street Journal. For German, there is the aforementioned Potsdam commentary corpus (PCC) [Stede, 2004, Stede and Neumann, 2014], in which newspaper commentaries are annotated on multiple levels, including RST structures and local relations based on discourse connectives. Apart from these, a handful of smaller corpora exist for a variety of languages, including Chinese, Spanish, Portuguese, Basque etc. On the application side, the automatic recognition of rhetoric structures, RST parsing, is an established task an has evolved over the last 20 years. A more detailed overview of the approaches will be given in Chapter 6.1.8.

**RST for argument representation?**

With its focus on speaker intentions and changes in reader attitudes, RST is by design well-suited for studying argumentative text. While purely descriptive texts (e.g., encyclopedia entries) or narrations, including news reports, often make for relatively 'boring' RST analyses, the description of argumentative text can reflect the way this text works in an interesting way. Thus, even though the tasks of explaining the coherence of a text (the goal of RST) and capturing the argumentation found in a text are not identical, it is tempting to employ RST for representing the argumentation structure of texts and thereby to eliminate the need for a distinct theory specific to argumentation.

In this vein, Azar [1999] argued that the nucleus-satellite distinction is crucial to distinguish between premises and conclusions in an argumentative relationship, and that, in particular, five RST relations should be regarded as 'argumentative' in the sense that one segment is a conclusion (or an opinion), and the other segment is an premise brought forward to, for instance, increase the reader's belief in the conclusion: Motivation for calls for action; Antithesis and Concession for increasing positive regard toward a stance; Evidence for forming a belief; Justify for readiness to accept a statement. Azar illustrates his idea with a few short sample texts, which he analyses in terms of RST trees using these relations, and he claims that the argumentation found in those texts is represented adequately. However, in one of the examples, Azar uses an interesting twist in the tree representation: The 14 minimal units are labelled '1-6' (indicating that the substructure among these units is not relevant for the argumentation), '7+12', '8', '9', '10', '11', '13', and '14'. The inter-

[7] This will be a single elementary unit when no multinuclear relation is involved on the path from root to leaf.

esting one is '7+12', which represents the central claim of the text, supported by Evidence relations from '8' to '14'. The claim is thus split between two non-adjacent text units, and Azar simply fuses them into a single node. This is clearly in conflict with a central principle of RST, whose object of study is the actual linear sequence of textual units and their coherence. Azar therefore seems to regard RST more as a notation (which can be adapted to one's purposes) rather than as a theoretical framework.

In a somewhat similar fashion, Green [2010] borrows certain aspects from RST (several relation definitions and the nucleus / satellite distinction) for her 'hybrid' representation that is supposed to capture both the argumentation structure and the rhetorical structure. Green studies medical patient letters that explain a diagnosis and provide reasons for recommendations on the patient's behaviour. The ultimate goal of the project is the automatic generation of these letters, but text representation is a significant part of it. The tree structure suggested by Green consists of text segments participating in the argumentation, which are linked by RST relations, but this information is supplemented by decidedly argumentative annotations: First, links from RST relation to the segments are labelled not only as nucleus and satellite but also with labels from other argumentation theories (the role the segment is playing in the scheme of Toulmin [1958], or with names of argumentation schemes [Walton et al., 2008], both of which we will learn more about in Section 2.3). When this argumentative analysis is done, the annotator reconstructs the enthymemes, i.e. the *implicit* propositions, and adds them to the tree as leaf nodes on a par with the minimal text segments. Moreover, a leaf may contain a text segment copied from elsewhere, to cover cases where old information from the text is needed to complete an argument. The tree structure in the end consists of potential redundant text segments and formulations of implicit statements. On the whole, the representation thus takes some inspiration from RST but serves a different purpose than that intended by Mann and Thompson (i.e., to reflect the coherence of a text, with segments taken in the order of appearance).

More recently, Green [2015] argued that the hybrid representation does not readily translate to a different text genre (biomedical research articles), and she concluded that RST and argumentation structure operate on two levels that are subject to different motivations and constraints, and thus should be kept distinct.

We now return to the question whether the original, *bona fide* Rhetorical Structure Theory tree can be an appropriate device for representing argument structure. On the one hand, the interesting parallels between RST's presentational relations and argumentative moves, which had already been noted by Azar, make RST a promising candidate for a text-oriented argumentation representation. There are different relations to express argumentative support and attacks which could be mapped correspondingly. We therefore consider the inferentiality requirement satisfied. The dialectical roles of proponent and opponent are not directly represented in RST. Nonetheless we would propose as before to derive the

switches of argumentative role from the series of attacking relations. Compositionality is also provided by the tree building principles of RST.

Having said this, several observations from our own text analysis work with the Potsdam Commentary Corpus [Stede, 2004, Stede and Neumann, 2014] indicate that there are (at least) two principal limitations. The first concerns long-distance dependencies and non-linearities of various kinds. This is the problem Azar circumvented by creating an artificial node covering two non-adjacent text segments. We found that quite often arguments in text are not linearised in a straightforward way: Pro- and contra arguments may be dispersed across the text and need to be linked to their common conclusions, which can violate RST's ruling out of crossing edges. Similarly, the end of a text often repeats, or slightly extends, the main thesis that has been stated earlier, necessitating the two segments to be brought together.[8] Consequently, we posit that the underlying phenomenon of non-adjacency creates a problem for RST-argumentation mapping in general, i.e., it is not limited to discontinuous claims: Both Support and Attack moves can be directed to material that occurs in non-adjacent segments.

The other observation concerns the structural configuration of attack and counter-attack, which is found frequently in the PCC. Consider the text given in Figure 2.1. Segment 1 and 2 present a simple argument: The building should be torn down, because it is full of asbestos. Segment 3 then attacks this argument in a way which we will later identify as an 'undercutter'. It presents an exception to the rule used in the argument from 2 to 1: If the contamination can be cleaned up, then a demolition does not necessarily follow. This objection is then countered in segment 4, again by an undercutting relation. The author grants that a cleanup might be possible, but that it will not happen because it is too costly.

Our RST analysis in Figure 2.1 does not reflect the attack- and counter-attack configuration at all. Firstly, the fact that 3 attacks 1-2 is not represented. Instead, 3 will always be understood as the satellite of 4, be it as an Antithesis or as a Concession. The 3-4 complex thus somewhat represents the counter-attack of 4 on 3, but it does so from the perspective of 4. As a result, the complex of 3-4 can only be linked to 1 with a supporting relation, and we lose the information that there is an attack and a counter-attack in the text. In this case, the intrinsic dialectics of argumentation could not be recovered from the RST relations.

An RST tree that better reflects the attack- and counter-attack configuration would need to link segment 2 first to 3-4, for example via Antithesis with 2 as nucleus, which would then make 2-4 the Evidence for 1. From the text-descriptive viewpoint, this would be a bit unfortunate (the main 'break' in the representation is in the middle of the first sentence rather than at a sentence boundary), but would still yield a plausible analysis. If the text were somewhat more complex, the overall configuration still makes sense if Marcu's com-

[8] We do not imply here that those texts are 'bad' – they are perfectly easy to understand and have straightforward RST analyses. But from that analysis, the argumentation structure cannot be read off without adding 'deeper understanding' and rewriting the representation.

positionality criterion is applied: The most-nuclear segment in 2-4 is 2, which is the major Evidence for 1. Segment 3, however, is only a satellite to 4 (and rightly so), and therefore the compositionality principle cannot yield the information that 3 is closely linked to 1-2 (as a direct attack) – this information is simply lost in the RST tree.

Examples of this kind are not uncommon, and unfortunately the problem can get even worse. Imagine a linearisation variant of the text where the premise precedes the conclusion: *[The building is full of asbestos,* $]_1$ *[so we should tear it down.* $]_2$ *[In principle it is possible to clean it up,* $]_3$ *[but according to the mayor that would be forbiddingly expensive.* $]_4$ Segment 1 can only be analysed as Evidence for 2. The Antithesis between 3 and 4 stays the same, and it is then another Evidence for 2. But we have no way of capturing the attack relation between 1 and 3-4, as long as we adhere to RST's principles. Again, this is not a criticism of RST – the analysis for *its* purposes is perfectly plausible – but an observation on the limitations of its accounting for 'deeper' structural configurations in argumentative text.

Concerning our desiderata, we consider text genre and domain independence granted, given the generality of the relation set and the successful application of the theory to different types of texts. Furthermore, the reliability of annotation has been shown and several corpora in different languages exist and have been used to automate the recognition of rhetorical structure.

The main obstacle for using RST to represent the structure of argumentation, to summarise, is the conflict between segment *adjacency* (a central feature of RST's account of coherence) and *non-adjacency* (a pervasive phenomenon in argumentative function of portions of text). We thus also subscribe to the view that (at least for many text genres) distinguishing rhetorical structure and argumentation structure is important for capturing the different aspects of a text's coherence on the one hand, and its pragmatic function on the other.

### 2.2.6 Semantic accounts of discourse structure

The Discourse Representation Theory (DRT) [Kamp and Reyle, 1993] proposes a syntax-semantic interface for deriving a model-theoretic semantic representation not only on the sentential level, but also for a discourse of multiple sentences. A discourse representation structure is defined as a set of discourse referents and a list of logical predicates. For adding the representation of a new sentence to the already existing discourse context, either the referents and predicates associated with the new sentence are appended to the overall discourse structure context, or a new subordinated structure is integrated into it. A dynamic logic [see e.g. Kamp, 1981, Staudacher, 1987] is used to interpret the structure dynamically through subsequent updates of the context. The main aim of the theory is to model the accessibility of discourse referents for anaphora, but similar dynamic semantic theories have also been used to model tense and presuppositions.

What this discourse structure lacks is a clear separation of the semantic contributions of each segment of the discourse, as well as a notion of discourse coherence, which makes explicit the relations between the segments. Both has been provided in the Segmented Discourse Representation Theory (SDRT) [Asher and Lascarides, 2003]: Every segment of the discourse has its own structure which is labelled with a logical identifier, enabling reference to these structures. Furthermore, coherence relations between these discourse segments may be expressed as logical predicates, in the very same way as DRT uses logical predicates expressing relations between discourse referents. The discourse structure derived by SDRT can be formally represented as a directed acyclic graph: It consists of the labelled coherence relations on the one hand, and of subordinating relations between the different SDRSs on the other.

There exists several corpora with discourse structures annotated according to SDRT, most of which also report inter-annotator agreement scores: Baldridge et al. [2007] annotated parts of the MUC-6 and ACE-2 corpora for their experiments. Afantenos et al. [2012] presented a French collection of news and Wikipedia articles annotated with SDRT structures. English multi-party dialogues have been collected and marked up from chats in online sessions of the game 'The Settlers of Catan' by Afantenos et al. [2015]. Finally, Benamara et al. [2016] presented a corpus of French film / product reviews and letters to the editor, as well as English film reviews for a study of discourse and sentiment.

SDRT has been used to model a variety of discourse phenomena, although most of them are situated in the context of dialogue rather than monologue (with a few notable exception). The derivation of SDRT structures, however, from the syntax-semantic to the semantic-pragmatic interface is a quite complex process: Multiple different logics and reasoning processes are involved and lexical knowledge is required for determining a maximally coherent discourse representation, with the corresponding structure and relation types.

For automatic recognition of these structures, most researcher therefore rather relied on machine learning approaches than on implementing the elaborate logics. A more detailed review of the corresponding approaches will be given in Chapter 6.1.8.

Let us now revisit out initial question as to whether SDRT meets our requirements for representing argumentation structure: The coherence relations proposed for monologue text do not explicitly represent the argumentative moves of support and attack. Nonetheless, there are various relations which we expect to find frequently in argumentative texts, such as causal relations of result, explanation, and goal, as well as the contrast relation. As in previously reviewed approaches of coherence relations, a corresponding mapping to argumentative relations would be required to fulfil the inferentiality requirement. The dialectical level could be derived as well from the contrast relations, yet with less precision, as not every contrast is used argumentatively. The structural requirements we consider all to be fulfilled: The representations are compositional and allow for non-linearities and

long-distance dependencies. Furthermore, the theory is independent of genre and domain. Several annotated corpora exist including reliability reports, and at least one of them, the corpus of film reviews, features a text genre that is very likely to contain monologue argumentation.

In comparison to the approaches reviewed above, SDRT has a lot to offer to argumentative analysis. There are no structural restrictions, and the theory is established in the field both concerning its theoretical modelling, as well as for its empirical studies and computational modelling using available corpora. Nonetheless, it is not the perfect candidate: An adequate representation of argumentation structure requires more work on the set of coherence relations and an appropriate mapping. Furthermore, the graph structure might be too powerful: It is not clear whether the nesting of substructures is required for representing argumentation. While it is necessary for the theoretical modelling of the accessibility of discourse referents and the availability of substructures for attachment, practical approaches to SDRT-based discourse parsing typically factor the nesting out for efficiency.

### 2.2.7 Conclusions

It is now time to cast a look back at all the theories of discourse structure we have reviewed in search for a suitable candidate for representing argumentation structure. First, we observe that no theory turned out to completely fulfil all our requirements and desiderata. This was expected, in that none of the theories were explicitly developed to model argumentation structure.

The zoning approaches, as well as non-hierarchical illocutionary labellings were clearly not structurally potent enough to represent the rich relational structure of argumentation. In contrast, a PDTB-style analysis of the coherence relations could model these relations, but the theory is intentionally agnostic of any form of global discourse structure. From the different approaches that propose tree structures, the majority employed constituency trees (such as several illocutionary accounts, LDM, and RST). These typically have the problem that non-linearities and long-distance dependencies can only be represented using extra constructs, such as special nodes for parallel constructions or extra information like nuclearity required for the Nuclearity Principle. Dependency structure has proven to be more flexible here. Finally, D-LTAG and SDRT proposed graph structures that are richer than trees. While this gives additional flexibility, it remains to be investigated in which cases this extra representational power is really needed, as it might complicate the discourse parsing later on.

In terms of inferentiality and dialectics, we have found many approaches that provide us with a useful set of discourse relations for various forms of support. In contrast, attacking relations are typically not covered well: Sometimes they are not considered at all, often they are only potentially signalled by a contrast relation which could, though, just as well be a

non-argumentative, i.e. purely semantic contrast relation. In all cases, the dialectics would have to be inferred, since no theory allows for a dialectical distinction between discourse segments.

Before we draw a final conclusion, we first want to review the approaches to argumentation structure that the argumentation community has brought forward.

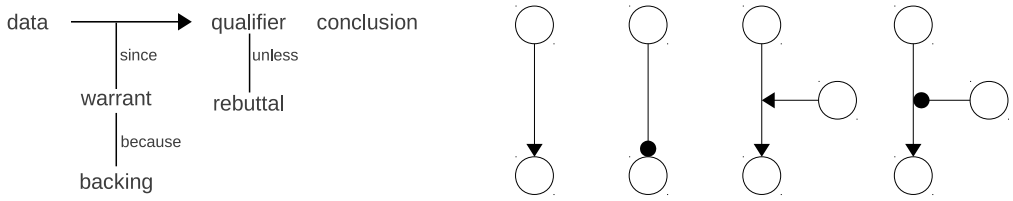
## 2.3 Structure of argumentation

After having reviewed the literature on discourse structures and investigated whether these could be used to represent argumentation structures, we now turn to theories decidedly developed to represent argumentation. The literature on argumentation is vast, and we consider here only work that has a clear focus on proposals for formal graphical *notations*. A very useful earlier overview of the use of argument diagramming techniques to represent the structure of arguments has been given by Reed et al. [2007]. The authors review the theories and diagramming schemes in logic, law, and artificial intelligence and cover many important aspects relevant to (automatic) evaluation of arguments. In contrast, our aim is of a rather descriptive nature and our secondary focus is on the linguistic realisation of argumentation, especially on the representation of argumentative opposition.

Due to our restriction of scope, some important research will not be included here. A very influential approach, which has inspired the whole field of computational argument in artificial intelligence, is the framework for argumentation graphs proposed by Dung [1995]. While superficially similar to the kinds of graphs we are interested in here, this framework applies to a different level: Dung is interested in a representation of arguments that allows for formally modelling reasoning processes. Most importantly, arguments are the *elementary* units in his attack graphs. Hence the relations represented in such graphs are not argumentative relations between propositions, but between instances of argumentation, i.e. between inferences. Our interest, however, is in representation geared toward modelling the *textual presentation* of arguments and thus the relation between the expressed propositions.

### 2.3.1 From Toulmin's roles to dialectical, compositional structures

An important step in the younger history of the development of a theory of argumentation is Stephen E. Toulmin's influential analysis of argument [Toulmin, 1958]. Dissatisfied with the simple analysis of an argument into premises and conclusion, he investigated the actual use of arguments with the aim to identify different roles that utterances can play in arguments, i.e. the way they contribute to its persuasive force. Toulmin proposed a scheme with six functional roles (see Figure 2.3a): On the grounds of some evidence ('data') and a possibly implicit but defeasible generalisation ('warrant') a conclusion is derived. The




(a) Toulmin (b) Grewendorf

Figure 2.3: Diagramming techniques in theories of argumentation of Toulmin and Grewendorf.

conclusion can be 'qualified' by a modal operator, indicating the strength of the inferential link. Furthermore, a 'rebuttal' can specify an exceptional condition that would undermine the inference if it holds. Finally, the warrant can be further supported ('backing').

Toulmin's scheme certainly represents the inferential and dialectical relations we are interested in. It does so, however, in a flat and non-compositional way. The relations are implicitly assumed between the different roles, such as for example the inference from data and warrant to the conclusion. The overall structure is regarded a single unit of analysis, and not considered to be recursively embedded. Since Toulmin did not aim to present a model of the structure of *text*, he makes no commitment as to how the propositions in his examples should be linearised in a text. We thus conceive the requirements of non-linearity and long-distance dependencies to be met per default.

There exist corpora annotated according to Toulmin's scheme, which are also in view of the reliability of annotation, see [Habernal and Gurevych, 2017] as one example (of a similar scheme). The inferences captured by Toulmin's scheme are quite general, so we consider it to be genre and domain independent, something that was also empirically studied by the authors of the latter paper.

Of the immense amount of literature on Toulmin's theory in different disciplines (ranging from philosophy and pedagogy to legal studies, linguistics, artificial intelligence, and others) we want to focus on a number of critiques that have been articulated, addressing problems of the application of the theory on complex, authentic argumentations, of the distinction of the functional roles, and of the representation of the opponent.

Both Öhlschläger [1979] and Kienpointner [1983] address the lack of compositionality, and suggest that the backing of the warrant should better be represented as a new, connected argument. In this new argument, the backing serves as the data and the warrant as the conclusion. Since a combination of several arguments is necessary for the representation of complex argumentation anyway, Öhlschläger presents a scheme that allows to recursively chain arguments together, thus building a serial structure, or 'multi-level argumentation' in the terms of Kopperschmidt [1989]. Also, multiple arguments can be presented in favour of the same conclusion, which Kopperschmidt calls 'multi-threaded' argumentation. However,

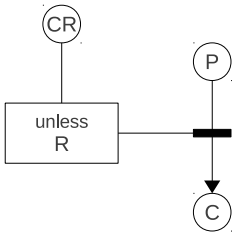
Öhlschläger's scheme does neither integrate Toulmin's rebuttal, nor the qualifier. Similarly, Klein [1980] argued for a recursively applicable argumentation scheme. Furthermore, he claimed that the distinction between Toulmin's data and warrant cannot always be drawn precisely. He proposed a representation of argument that can be conceived basically as a support tree, with the root node as the main claim and supporting arguments in the unfolding tree structure.

However, all of the schemes discussed so far lack a proper representation of the opponent. Due to its dialectical nature, argumentations often refer to an explicitly mentioned or at least supposed opponent, as for instance in the rebutting of possible objections. Wunderlich [1980] thus interpreted Klein's support-tree as a 'decision'-tree, where the root node is the 'quaestio', i.e. the question to be decided on. From there, not only arguments *for* and but also *against* the decision unfold recursively. Since there can be pro and contra for every node in the tree, the opponent's role is integral to this representation.

Grewendorf [1980] then offered a dialogue-oriented diagram method that also demonstrates the origin of arguments: It is possible to distinguish between counterarguments that are brought up by the opponent as explicit attacks, from those that the proponent himself anticipated in order to refute them. In addition, Grewendorf replaces the tree structure with a graph, so that nodes can participate in multiple support or attack relations. In the diagram, the polarity is no longer an attribute of the node but an attribute of the link depicted by different types of arrows (see Figure 2.3b). Those with an arrowhead denote support, those with a circle an attack. Finally, Grewendorf makes the interesting move to allow support and attack not only for statements (nodes) but also (recursively) for support and attack relations. Hence it is in principle possible to also represent meta-communicative disputes. Note that Grewendorf's account is the first to fulfil all our requirements. However, Grewendorf provides only a rough outline of his diagram method and no formal specification. One of the aspects missing is amongst others a specification for conditions of a node having multiple support by a series of nodes. As a consequence, authors who took up his proposal sometimes produced ambiguous graphs that are difficult to interpret, as for instance in [Adachi-Bähr, 2006].

### 2.3.2 Freeman's macrostructure of argumentation

A detailed examination of Toulmin's theory has been presented by Freeman [1991], whose goal was to integrate Toulmin's ideas into the argument diagramming techniques of the informal logic tradition (see Beardsley [1950] and its refinement by Thomas [1974]). Recently, an updated but compatible version of the theory has been presented in [Freeman, 2011]. If necessary, we will distinguish between both versions in the following discussion, but otherwise simply speak of Freeman's theory.




(a)

*Proponent:* *We should tear the building down.*
*Challenger:* *Why?*
*Proponent:* *It is full of asbestos.*
*Challenger:* *What makes you sure about your claim, given that it is in principle possible to clean it up?*
*Proponent:* *According to the mayor, this would be forbiddingly expensive.*

(b)

Figure 2.4: Freeman's representation of a rebuttal and counter-rebuttal and the corresponding dialectal exchange.

The central claim of Freeman's theory is that the so called macrostructure of argumentation, i.e. the different ways in which premises and conclusions combine to form larger complexes, can be modelled as a hypothetical dialectical exchange between a proponent, who presents and defends claims, and a challenger, who critically questions them in a regimented fashion. Every move in such a *basic dialectical situation* corresponds to a structural element in the argument diagram. The analysis of an argumentative text is thus conceived as finding the corresponding critical question of the challenger that is answered by a particular segment of the text. Freeman's theory thus makes an explanatory valuable connection between the focus on arguments *as process*, such as found in the study of dialectical dialogues in philosophy or in rhetorics, or even as a special form in judicial proceedings, and arguments *as product*, such as found in the study of persuasive text in newspaper commentaries, in scientific writing, or even advertising. The dialectical process serves as a model for the studied product. This combination of process-based dialectics and product-oriented structure representation makes the theory accurately fitting our goal of argument structure recognition for authentic text.

Freeman presents critical questions a challenger would ask in a basic dialectical situation for the structural complexes typically assumed in the informal logic tradition: for serial, linked, convergent, and divergent structures.[9] In his reception of Toulmin's theory, Freeman rejects the distinction between data and warrant: Although they are obviously discernible in arguments as process, he argues in an extensive discussion that the distinction is not applicable for arguments as product (see Ch. 3 in both Freeman [1991, 2011]). Freeman thus assumes one category of premises subsuming both data and warrant. Also, the category of backing is dropped in favour of a compositional analysis of serially connected arguments. The qualifier is integrated into Freeman's theory as a property of the inferential links between premises and conclusion.

[9] A more detailed description of those complexes follows in Chapter 3.

Especially interesting for us is Freeman's integration of Toulmin's rebuttal. As described above, the *rebuttal* is an exception of the generalisation presented as the warrant. It specifies a condition under that the inference from premise to conclusion would not hold. Typically, an author mentions a possible exception to preempt his critics and then in turn rebuts that anticipated objection, in Freeman's terms a *counter-rebuttal*. In the basic dialectical situation, the challenger asks a critical question with a possible objection, and thereby forces the proponent to defend her argument accordingly. A diagram featuring a simple sequence of claim, premise, rebuttal and counter-rebuttal and the corresponding hypothetical dialectical exchange is shown in Figure 2.4a. Additionally, Freeman [1991] allows the challenger to make her possible objection stronger by supporting it with further premises, which Freeman terms 'defended rebuttals'. Although Freeman identifies different ways to attack and defend an argument, his use of the terms 'rebuttal' and 'counter-rebuttal' is rather general and corresponds to argumentative attack and counter-attack.

Freeman's diagramming technique is not perfect, however, especially for certain complex combinations of features: Freeman lists all rebuttals of one argument in a single rebuttal-box. In order to relate the counter-rebuttals to their target, co-indexation is introduced instead of representing the relation by links in the diagram. Also, to represent a defended rebuttal that is countered by attacking its support, Freeman [1991, p. 199f] introduces further diagram features such as crossed-out boxes additionally containing the defeated rebuttal. These artificially increase the structural complexity, even though a more simple representation with relations, analogous to the rebuttal, would be possible.

Another issue concerns the representation of the distinction between what Pollock [1995] calls *rebutting* and *undercutting defeaters* in defeasible reasoning. Freeman indeed makes equivalent distinctions: For attacks by the opponent, he acknowledges both the denial of the conclusion and the undercutting of its support (i.e. exceptional 'rebutting' in Toulmin's sense). For counter-attacks by the proponent, he considers both denying the exception's holding and denying the exception's undercutting force. But these differences are not reflected in the argument diagram. Rebutting and undercutting attacks of the challenger are represented uniformly in Freeman's 'rebuttal'-box (although Freeman [2011, p. 23] lists them with different prefixes). Rebutting and undercutting counter-attacks of the proponent are represented uniformly as nodes attached to the 'rebuttal'-box. Only [Freeman, 2011, p. 57] chose to represent undercutting counter-attacks in a new fashion in order to visualise the difference.

Finally, it is worth noting that Freeman now integrates uncountered attacks, or *counter-considerations* in the terminology of Govier [1985] into his theory. Here, instead of rebutting or undercutting a possible objection, it is left uncommented because it is understood as being outbalanced by more weighty reasons in favour of the claim. While Freeman [1991, p. 173] argued that such counterconsiderations need not be represented in argument structure, because they could be seen as rhetorical accessory logically not effecting the case for

the claim, they are now represented as a special 'even though' rebuttal in [Freeman, 2011, p. 29]. This extension of the theory was in our view an advantageous move, as this argumentative strategy appears frequently in argumentative text and could not be adequately represented before. There are many more noteworthy features of Freeman's approach that are beyond the scope of this discussion, as for instance the elaborate and ongoing discussion of the linked-convergent distinction or the representation of suppositions.

Freeman's theory fully represents the aspects of inferentiality and dialectics we are interested in, and the formation of complexes is compositional. In terms of linearisation and long-distance dependencies, again, no commitment is made, as the theory did not aim to be a text structural model. We thus find all requirements to be fulfilled. The desiderata of genre and domain independence are also met. What is yet missing is firstly the proof that argumentation structures can be reliably annotated according to this theory, and secondly a corresponding resource that can be used to study and model the argumentation in text.[10]

### 2.3.3 Pragma-Dialectics

The pragma-dialectical theory of argumentation [van Eemeren and Grootendorst, 1984, 1992, 2004] aims to combine the study of argumentation *as a product* (the perspective typically taken in logical analysis and also in descriptive linguistics), and the study of argumentation *as a process* (the perspective on the communicative and interactional aspects) into a holistic investigation of the discourse activity of argumentation. Argumentation is understood as a complex speech act performed in order to resolve a difference in opinion. This complex activity is realised by more elementary speech acts or illocutions, such as those we already studied in Chapter 2.2.3.

The theory has a strong focus on a normative characterisation of rational argumentation: An ideal model of a critical discussion is presented, where based on rationality and reasonableness the two roles of the protagonist and the antagonist resolve their conflicting positions in a regimented fashion. First of all, the argumentative discussion is divided into four stages: In a confrontation phase participants establish that they have conflicting positions. The opening phase serves the purpose of determining a common ground. In the argumentation phase then the protagonist presents and defends her claims, while the antagonist critically questions them. Whether and how the difference of opinion is being resolved is finally settled in the concluding stage. Most importantly, the authors posit ten rules all rational participants should ideally follow. Violations of these rules are considered as *fallacies*. Three different argumentative schemes are allowed based on causality, comparability, and symptomaticity. The theory is used normatively in relying on its general code of

[10] The only work we are aware of that applied Freeman's theory to text in the field of linguistics is [Stede and Sauermann, 2008], where ten commentary texts were analysed.

conduct of critical discussions, but it is also applied in more a descriptive way for describing failed discussions, or instances of fallacious argumentation.

The examples used by the authors are often dialogue excerpts or transcripts, and it is indeed the *dialogical* interaction that is predominantly investigated. However, the theory is open to monological argumentation as well. Here the distinction between 'explicit' and 'implicit' discussion is important: In monological discussion, the role of the antagonist is not impersonated. The criticism usually brought forward by the antagonist is anticipated by the author and communicated implicitly. This view is similar to Freeman's hypothetical dialectical exchange.

In terms of argumentation structure, the pragma-dialectical theory distinguishes between the following types of complexes which are used to generate large argumentation diagrams [see van Eemeren and Grootendorst, 1992, p.73-89]: 'single' argumentation, with one premise and one conclusion; 'multiple' argumentation, with several independent premises for one conclusion; 'coordinatively compound' argumentation, where several premises are required to be accepted in order to support the conclusion; and 'subordinatively compound' argumentation, providing a serial chain of subordinated arguments. Finally, the authors also include unexpressed premises in their diagrams. These complexes are comparable to the simple, convergent, linked, and serial structures of Freeman. A discussion of the subtle differences is provided by Snoeck Henkemans [2000].

Interestingly, the antagonist and the corresponding critical questions or objections are not directly represented in the argument diagrams. Only supporting relations are shown in the structure, and claims associated with an attack occur there only indirectly. For instance, consider the examples discussed by Snoeck Henkemans [2003]: An attack of the antagonist that was successfully rebutted by the protagonist is simply dropped and not represented in the structure. In another example an undercutter of the form 'But couldn't we do X?' was (since it was successfully rebutted) transformed into the negated assertion 'We can't do X', or as in a different example turned into a meta-claim 'The objection that we can do X is not a sound argument'. The fact that the antagonist and his attacks are not represented in the structure may be related to the 'Closure Rule', one of the rules of the code of conduct. It states that a successful defence obligates the antagonist to retract his doubts (and likewise that a failed defence obligates the protagonist to retract his standpoint). This would mean, though, that the structure only reflects the argumentation as the *resolved product* and does not reflect the dialectics of its production.

Although the pragma-dialectic theory is akin to descriptive inquiry and provides a lot of examples from authentic sources, neither an annotated corpus, nor a proof of reliability of annotation are available yet.

### 2.3.4 Argumentation schemes

A theory that has been used extensively for the analysis of argumentative text is that of argumentation schemes [Walton, 1996, Walton et al., 2008]. Here, arguments are understood as instances of abstract argumentation schemes, a large and detailed catalogue of which is provided. Each scheme specifies the required premises, implicit assumptions, and possible exceptions. The premises are all required to be accepted in order to lend support to the conclusion. Assumptions are considered to be accepted implicitly, unless they are questioned explicitly. Exceptions, finally, may undercut the argument: If one of the exceptions holds, the argument is rendered ineffective, even though all other premises might be accepted. For every premise, assumption, or exception there is a corresponding critical question specific to the argumentation scheme. The theory is integrated as a diagramming technique in argument visualisation tools such as Carneades [Gordon, 2010].

The main focus of the theory of argumentation schemes is on defining the premises, assumptions, and exceptions of the various schemes. Yet, the formation of structural complexes is acknowledged, in the way that each of the components of an applied scheme may itself be the target of an argumentative relation. Consequently, serial and multiple argument structures are possible. Linked structures are represented scheme-internally, as all premises in one argument are understood as being linked. Since the theory does not assume a sequence of text segments as elementary units, no commitment is made towards linearisation constraints. We thus conceive the requirements of inferentiality and compositionality met, and find no restrictions towards non-linearities or long-distance dependencies specified.

In terms of dialectics, it is worth considering how argumentative attacks and counter-attacks are represented [see Walton, 2011, 2012]. Rebutting attacks are expressed as arguments of opposite polarity against a claim. Undercutting attacks are expressed by claiming that one of the defined exceptions holds. These attacks can itself be supported or rebutted again. Since the theory distinguishes between support and attack, the dialectical role of each unit can be derived. It should, however, be noted that the polarity of the relation targeting the exception depends on the way the exception is expressed: In the context of arguments from expert opinion, we found examples where the exceptions were framed as stating the exception ('expert is not trustworthy'), so the argumentative attack was actually a *support* for the proposition that the exception *holds*. Conversely, we also found examples where the exceptions were framed by stating the norm ('expert is trustworthy'), so the argumentative attack was indeed an *attack* against the proposition that the exception does *not hold*. Note that this remark is not intended as a criticism, but rather as a reminder that the framing of exceptions needs to be consistent and explicit if one is to derive the dialectical role from the polarity of argumentative relations.

Argumentation schemes have been annotated in the AraucariaDB, which is distributed in [AIFdb, 2016], following different scheme-sets, e.g. that of Walton et al. [2008]. Schemes

have also been the subject of automatic recognition, which we will report on later in Chapter 6.1. Whether analysts were able to annotate argumentation scheme (structures) reliably has, however, not yet been shown to the best of our knowledge. We consider the inventory of schemes as general enough to be applied to texts of different genre and in different domains.

The analyses following the argumentation schemes are clearly much more fine-grained than those of the approaches reviewed above, which often only distinguish between support and attack. We thus conceive an analysis at the level of argumentation schemes as a subsequent step after a coarse-grained analysis of argumentation structure. In general, however, it should be noted that in practice argumentation schemes often serve a purpose that is different from ours. Research around the theory of argumentation schemes has shown increased interest in building tools to support argumentation as a process for instructional, legal, or decision-making ends and in advancing techniques of argument evaluation. Having said that, our search for a representation of argumentation structure is focused on more descriptive aspects, suitable for corpus-linguistic studies and ultimately for computational applications such as argumentation mining.

### 2.3.5 Inference Anchoring Theory

A link between the speech acts in argumentative dialogue and the structure of argumentation and its underlying propositions is made in the Inference Anchoring Theory (IAT) [Reed and Budzynska, 2010]. Although the focus of the theory emphasises dialogical argumentation even more strongly than the Pragma-Dialectics, we consider it here, because it combines logical and illocutionary characterisations of argumentation in an exemplary way.

The theory assumes three levels of analysis: On the *locutionary* level acts of utterance are represented such as *Bob says 'We should do X.'*, *Wilma says 'Why?'.*, or *Bob says 'Because of Y.'*. Locutions are connected by transitions, which correspond to rules of a dialogue protocol. On the *inferential* level the propositions expressed in the locutions are represented, such as *We should do X* or *Y* and they are connected by applications of inference rules, such as those of an argumentation scheme for instance. We thus have the bare locutionary acts licensed by a dialogue protocol on the one side, and the logical representation of an argument with propositions and inferences on the other. The key idea of IAT is to ground (or 'anchor') the inferential structure on the *illocutionary* level, which serves as an interface between locutions and propositions. Two types of illocutions are considered: Illocutionary acts of e.g. asserting, challenging, or questioning connect locutions with propositions (see also Chapter 2.2.3). So called indexical illocutionary forces then connect transitions with inferences. The act of arguing is considered as such: It connects the transition (from the challenging

why-question to the substantiating assertion) with the inference (from the proposition Y to X).

All three levels with their units and relations are represented in one graph with differently typed nodes. The representation is compositional and allows for all structural complexes we have seen so far: serial, multiple, and linked relations. As expected for a theory aiming to represent authentic multi-party debates, IAT does not impose any linearisation-specific constraints on the structure, and is thus prepared to represent non-linearities and long-distance dependencies. In contrast to Freeman and Pragma-Dialectics, IAT does not assume idealised dialectical roles of proponent and opponent. Instead, the actual participants are represented in the locutions, and – as so often in argumentative dialogues – each of them will have his own (sometimes even shifting) perspective on the argumentation. Dialectical roles could, however, as previously be derived from the series of supporting and attacking relations, depending on the participant's initial standpoint.

Overall IAT representations can be quite complex. For someone only interested in the structure of argumentation, the additional but theoretically necessary and informative levels of locutions and illocutionary force might be overwhelming. Furthermore, since the level hierarchy usually consumes the horizontal axis of the graph, the often non-linear argumentation structures are forced to unfold rather linearly on the vertical axis.

IAT has been applied to dialogue transcripts, such as radio debates in the 'Moral Maze' programme, annotations of which are available [Budzynska et al., 2014]. The authors also report on inter-annotator agreement on the different levels and constructs. We thus consider all desiderata met, though with the important caveat that all this only applies to dialogue and not to monologue text.

This brings us to our final remark. IAT has been developed and applied for dialogue, where it was used to successfully model complex argumentative interactions (including e.g. the use of assertive questions and rhetorical questions in debates). It would be highly desirable, we think, to apply IAT to monologue argumentative text, but we are not aware of attempts approaching this issue. Similar to Pragma-Dialectics, it might involve some notion of an *implicit* dialogue, and a thorough investigation would be required to determine which transitions of locutions are possible according to a decidedly monologue argumentation protocol, and which illocutionary forces are correspondingly expressed in the text.

## 2.4 Conclusion

Let us conclude this survey of theories of the structure of discourse and of argumentation. We first defined the criteria we applied in our review as five requirements (inferentiality, dialectics, compositionality, and the abilities to represent non-linearity, and long-distance dependencies), and four desiderata (text genre, and domain independence, reliability of

annotation, and availability of annotated corpora). We found that none of the reviewed theories fully fulfilled all requirements and desiderata.

We will first recapitulate our review of *discourse* structures. It started with theories proposing a flat partitioning of the text into argumentative or topic zones. These might carry information useful also to argumentation, but generally failed to represent argumentative relations. The analysis of local coherence relations in the manner of the PDTB was able to represent those, but the theory intentionally makes no assumptions towards a global discourse structure. Structures based on constituency trees (used for example in illocutionary accounts, in LDM, and RST) could adequately describe the composition into larger complexes, but could not handle non-linearities or long-distance dependencies without additional annotations such as nuclearity. These structures involving non-adjacency could be better handled by dependency trees (used also in illocutionary accounts). Even richer graph structures had been proposed by D-LTAG (given that one includes the important anaphoric relations) and by SDRT, but these come at the cost of more complicated discourse parsing.

While most of the approaches provided us with a set of relations that could be partially mapped to the relation of argumentative support, many theories did not cover well attacking relations adequately – and even if there were relations that could correspond to attacks they often could not be mapped without ambiguity. The distinction between semantic and pragmatic relations is an important issue here, as well as the disambiguation of discourse connectives. For all discourse structures, the dialectical role of a segment could only be derived from the sequence of supporting and attacking relations. Bear in mind, though, that the representation of argumentation structure was not a central objective for most of the theories of discourse structure.

Reviewing the theories of *argumentation* structure, we started with Toulmin's influential model, a labelling of the different roles a segments can have in an argument. It has been mainly criticised for its lack of compositionality, but also for the often impractical distinction between data and warrant. Several authors proposed compositional tree structures in reaction to these issues, but disregarded the rebutter and hence the distinction of dialectical roles. Grewendorf's graphs were the first structures that fulfilled all our requirements. However, the scheme was only rudimentarily sketched, and when applied to authentic texts it led to ambiguous representations. The theory presented by Freeman likewise fulfilled all our requirements, but in comparison to Grewendorf it is more extensively and more elaborately defined. Our only criticism concerns the elegance of representation of certain complicated structural configurations. In terms of our desiderata, a proof of reliability and a larger annotated resource have not yet been provided, though. We also considered the structures described by the normatively oriented pragma-dialectical theory. Here, our main concern was that the opponent was not explicitly represented in the argumentation structures, although this role is an integral part of the situation of a critical discussion, and that text segments clearly referencing this role were either excluded from the structure or

transformed. The theory of argumentation schemes was also a promising candidate that fulfilled our requirements, but one that aims for representations that are more detailed than we require in the first place. Nevertheless, we consider this fine-grained analysis into the different argumentation schemes and their corresponding critical questions as a subsequent step that could follow after a coarse-grained analysis of argumentation structure has been derived. Finally, rather as an outlook, we considered IAT as a means to represent the argumentation structure in monologue text. It has been developed for and successfully applied to argumentative dialogue and offers a powerful analysis on the locutionary, illocutionary and inferential level. It is, however, not clear yet how this analysis could be translated for application to monological text.

Before we choose a candidate theory and move on, it should be highlighted that there has been considerable effort – approaching the different ways of representing argumentation from a practical, computational perspective – to represent argumentation in a more general and theory-agnostic way, an endeavour which lead to the specification of the Argument Interchange Format (AIF) [Chesñevar et al., 2006]. Among other things, AIF offers a more abstract representation in the sense that no strong structural commitments (e.g. by restricting to trees, or acyclic graphs) or detailed conceptual assumptions (e.g. by restricting to specific relation, scheme, or inference types) are made. This way, it is able to subsume representations of different theories that we reviewed above, such as for example instances of the Toulmin scheme, or Freeman-like structures. While this is a clear demonstration of the representational power of AIF, its main focus is to serve as a common data interface for multi-agent argumentative systems and to allow easier exchange of data between argumentation tools. We thus conceive it more as a valuable representation format, rather than as a theory that lends itself to annotation.

To conclude, we consider Freeman's theory as the most prominent and directly suitable candidate to represent the structure of argumentation for the aims of this work. Our next goals are therefore, first, to derive an annotation scheme for this theory. This will be the topic of the next chapter. We will then demonstrate that this scheme can yield reliable annotations of argumentation structure in Chapter 4, and finally present a text corpus annotated with these structures in Chapter 5.

# 3 A synthesised scheme of the structure of argumentation

In the previous chapter we have reviewed different theories of the structure of discourse and of argumentation. As a promising candidate we considered the theory of Freeman [1991, 2011], which takes the moves of the proponent and opponent role in a basic dialectical situation as a model of the structure of argumentation in texts. We also identified smaller issues with the structural representation of the rebutting and undercutting distinction and of complex attack- and counter-attack constellations. In this chapter, we present a synthesised scheme, which largely builds upon Freeman's work but aims to mitigate the aforementioned issues through a more elegant representation. Note that this chapter focuses on the scheme itself and its features. Its implementation and application in annotation guidelines will be addressed in the following chapter.

**Previously published material**

The scheme has been initially proposed in [Peldszus and Stede, 2013b]. For annotation guidelines based on this schemes see Appendix A or the extended version of the guidelines published as [Peldszus et al., 2016].

## 3.1 Basics

We define an **argument** to consist of a non-empty set of premises supporting a conclusion. We thus use the term 'argument' not for premises, but for the complex of one or more premises put forward in favour of a claim. Premises and conclusions are propositions expressed in the text segments. We can graphically present an argument in an argument diagram, with propositions as nodes and the relation as an arrow linking the premise nodes to the conclusion node.

When we speak of **argumentation**, we mean the structure that emerges when multiple arguments are related to each other and form larger complexes. The manner in which arguments combine into larger complexes can be generally described as either supporting, attacking, or counter-attacking. In the following sections we will describe each of them.

## 3.2 Support

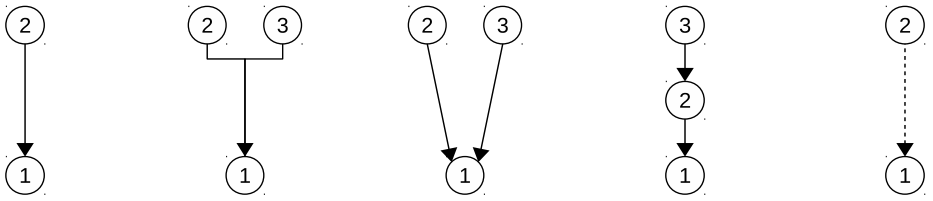
**Simple support:** The most simple configuration of an argument would consist of two propositions, one conclusion that is supported by exactly one premise, as in example (1). The corresponding structure is shown in Figure 3.1a. In the basic dialectical situation, the support relation is triggered by the opponent challenging the presented claim by asking for a reason. In our example, the argument is linearised by first presenting the conclusion, then the premise. Of course the argument could also be presented with the premise preceding the conclusion, as in example (2).

(1) [We should tear the building down.]$_1$ [It is full of asbestos.]$_2$

(2) [The building is full of asbestos.]$_1$ [We should tear it down.]$_2$

**Linked support:** If an argument involves multiple premises that support the conclusion only if they are taken together, we have a *linked* structure in Freeman's terminology. Only if both premises are accepted, are they able to support the conclusion. In the basic dialectical situation, a linked structure is induced by the opponent's question as to why a premise is relevant to the claim. The proponent then answers by presenting another premise explicating the connection. Building linked structure is thus to be conceived as completing an argument. A typical example for linked premises are inference rules, such as in (3).[1] Linked support is shown in the diagram by connecting the premises before they link to the conclusion (see Figure 3.1b).

[1] As discussed in Section 2.3, both Toulmin's data and warrants are represented as linked premises by Freeman. One might argue that data and warrant should not be linked according to this definition, for an argument might be fully functional without a premise corresponding to warrant. However, in this case the warrant would simply be implicitly assumed by the author. Since we aim to describe the author-relative argument as product, we postpone the issue of representing implicit premises for now.




(a) simple sup. (b) linked sup. (c) multiple sup. (d) serial sup. (e) example sup.

Figure 3.1: Basic support relations and complex formation.

(3) [We should tear the building down.]$_1$ [It is full of asbestos,]$_2$ [and all buildings with hazardous materials should be demolished.]$_3$

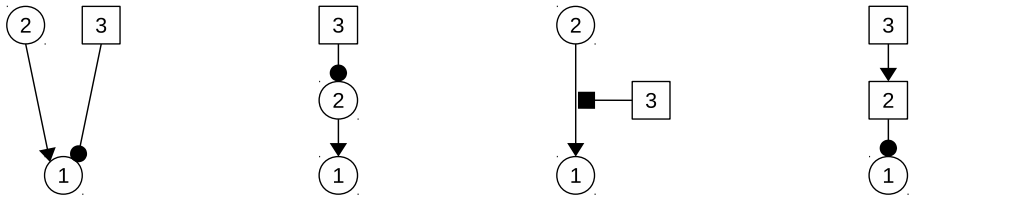
There are different ways to provide further support to the conclusion. One is to bring up a separate argument for the same conclusion, the other is to further develop the argument already given.

**Multiple support:** Let us start with the strategy where the author puts forward a separate, new argument for the same conclusion. For instance, consider example (4). Both arguments stand for themselves and each of them could be put forward by the author without the other. Both arguments are independent from another in the sense that the supporting force of one argument would not be impaired if the supporting force of the other is undercut.[2] On the dialectical level, the opponent asks: 'Can you give me an additional argument for that conclusion?', and the proponent answers by offering a new argument accordingly. We call this structure *multiple* support, in order to prevent confusion with Freeman's convergent structure. In our case, we could say there are two *arguments* converging on the same conclusion. In contrast, what Freeman identifies as convergent structures are two *reasons* converging in one and the same argument. A discussion of that difference can be found in [Freeman, 2011, Ch. 5]. Bringing forward a new argument for the same conclusion is graphically represented as a separate arrow linking the premises of the new argument to the common conclusion, as in Figure 3.1c.

(4) [We should tear the building down.]$_1$ [It is full of asbestos.]$_2$ [Also, people in the neighbourhood have always hated it.]$_3$

**Serial support:** Another way to provide further support to the conclusion is to further develop the argument already given by supporting one of the argument's premises. This is the case in example (5). The author presents a new argument to convince the reader of the acceptability of a premise. By directly supporting the premise, she is indirectly giving support to the conclusion. The role of the supported text segment is then twofold, on the one

[2]However, the arguments are not required to be independent in the sense of premise acceptability: If both arguments share a premise or have semantically interconnected premises, it may turn out that evaluating a premise in one argument as unacceptable also renders one in the other unacceptable.




(a) rebut a conclusion (b) rebut a premise (c) undercut an arg. (d) a defended attack

Figure 3.2: Opponent's attacks of the proponent's argument.

hand serving as a premise in the original argument, on the other serving as the conclusion of the following argument. On the dialectical level the opponent asks: 'Why should I accept that premise?', and the proponent answers by offering a new argument accordingly. Following the terminology of Freeman and others we call the resulting structure *serial*. Such serial structure is presented in the argument diagram by a new arrow linking the premises of the new argument to their conclusion, which is one of the premises of the original argument (see Figure 3.1d).

(5) [We should tear the building down.]$_1$ [It is full of asbestos.]$_2$ [The commission reported a significant contamination.]$_3$

**Example support:** A special form of lending support to a claim is that of giving *examples*. If the author claimed a generalisation, she can provide evidence that it proved to apply correctly at least in the given example, as it is the case in (6). Those arguments are based on inductive reasoning. Since Freeman represents inductive reasoning as convergent premises, there is no special type of question for the opponent in her conception. To make this more fine-grained distinction, it is reasonable to assume the opponent would simply be asking: 'Do you have an(other) example?' As with all supporting arguments, we represent the example arguments with an acute arrowhead, though with a dashed instead of a solid line; see Figure 3.1e.

(6) [A citizens' initiative can force the mayor to tear the building down.]$_1$ [In Munich such a group forced the local authorities to tear down an old office building!]$_2$

These supporting structures are not only available to the proponent to support his own claims, but likewise to the opponent.

## 3.3 Attacks

Now that we have presented the different ways to support an argument, we focus on ways to attack it. One is to present an argument against the conclusion irrespective of the support

for it; the other is to attack the cogency of the given argument by attacking its premises or by diminishing its supporting force. Both of these strategies can be used by the opponent to attack the proponent's arguments and by the proponent to counter the opponent's attacks. In the argument diagram, attacks will be indicated by solid arrows with a round or square arrowhead. Furthermore, we think it is useful to be able to clearly distinguish between the opponent's and the proponent's attacks. Thus, in allusion to Freeman's rebuttal box, segments corresponding to attacks of the opponent will be represented as box nodes, while those corresponding to counter-attacks of the proponent as circle nodes. We now describe attacks of the opponent, and consider the proponent's counter-attacks in the section 3.4.

**Rebutting:** Let us start with the strategy to provide a new argument against the conclusion, an example of which is given in (7). The author anticipates that there are premises supporting the negation of the conclusion. In accordance with Pollock and Freeman, we call this type of attacking argument *rebutters* directed against the conclusion. As mentioned in Section 2.3, Freeman does distinguish between rebutting and undercutting attacks by the opponent. However, he still represents both by the same structure in the argument diagram and does not provide us with an opponent's question specific to rebutting attacks. Since we want to represent this distinction structurally, the corresponding opponent's question would be: 'What makes you sure about your claim in the light of the following counter-evidence?' Such rebutters are depicted in the argument diagram as arrows with round arrowhead from the opponent's premise to the proponent's conclusion. An example is shown in Figure 3.2a.

(7) [We should tear the building down.]$_1$ [It is full of asbestos.]$_2$ [On the other hand, many people liked the view from the roof.]$_3$

Instead of rebutting the conclusion, the opponent could also attack the given argument by rebutting one of its premises, as in (8). Technically, this is not a new structure: Whether the attacked claim serves as a premise or as a conclusion in some argument is irrelevant for it being rebutted. However, this can be regarded as a different strategy. By rebutting the argument's premises, the opponent argues against the argument's cogency. For a corresponding argument diagram see Figure 3.2b.

(8) [We should tear the building down.]$_1$ [It is supposed to be full of asbestos.]$_2$ [Yet, nobody really made a precise assessment of the degree of contamination.]$_3$

**Undercutting:** Another way to attack the argument's cogency is in questioning the supporting force of the premises for the conclusion. On the text level, the author anticipates a possible exception (to an implicit or explicitly stated rule) that could defeat her argument if it would hold. For instance, see the example (9). On the dialectical level, the opponent argues for the invalidity of the inferential step from premises to conclusion by pointing to a possible exception. In doing so, he is neither rebutting the premise nor the conclusion,

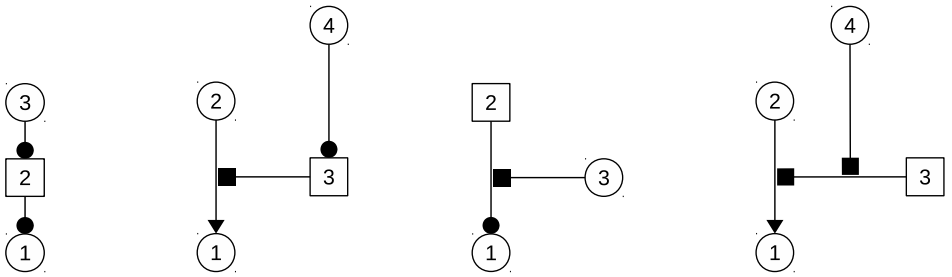
but restricting the applicability of the argument. The opponent's question presented by Freeman is: 'Why do your premises make you so sure in light of the following condition?' With Pollock and Freeman, we call this type of attacking arguments *undercutters*. They are represented diagrammatically as arrows with a square arrowhead directed to the body of the arrow representing the attacked relation. Figure 3.2c shows an example.

(9) [We should tear the building down.]$_1$ [It is full of asbestos.]$_2$ [The building might be cleaned up, though.]$_3$

Rebutting and undercutting attacks can sometimes be hard to distinguish on the opponent side: Is the given segment to be understood as an exception of the inferential move from premises to conclusion, or as an argument in favour of the conclusion's negation? A convenient way to tell them apart is to focus on the attacker's commitment to the conclusion. If the attacker presents a possible argument for the negation of the conclusion, this is a clear indicator for a rebutting attack.[3] Furthermore, contrary to rebutters of the conclusion, undercutters must be semantically related to the premise in some way. A possible test would therefore be to see how felicitous the attack is if the premise turns out to be false, is suspended, or is omitted. A rebutter of the conclusion will presumably be unaffected, while an exception without inference seems questionable. As an example, consider (7): When we omit the premise (segment 2), the attack is still a valid move. In (9) on the other hand, omitting the premise leads to an infelicitous attack.

**Defended attacks:** Freeman [1991] permits the opponent to provide support to his attacks and so do we. As an example, consider (10). On the text level this means that the author not only has the chance to present an anticipated argument against her conclusion or an anticipated exception to her argument, but also to strengthen it by explaining why it is worth taking this objection into account. All sorts of supporting relations described in the previous subsection are available for that purpose. Dialectically, this support of an attack is modelled by a temporal role switch between opponent and proponent. In our argument diagram these temporal role switches are already resolved, in that all supporting and attacking arguments are related to proponent and opponent according to the main claim. An example is shown in Figure 3.2d, where a rebutting argument is supported by an additional premise.

[3]Note that the opponent can only propose *possible* arguments conflicting the proponent. He is not allowed to assert a proposition in the basic dialectical situation, as his role is defined very restrictively to that of a constructive partner testing the proponent's argumentation by asking critical questions. His goal is to wrench the best possible argument for the main claim from the proponent. He will thus never argue out of his own interest to convince the proponent of some claim. Consequently, he can neither claim that the negation of the conclusion holds, nor that some exception holds. He can only present *possible* arguments in favour of the conclusion's negation or *possible* exceptions to some inference from premises to conclusion, in order to provoke a corresponding reaction of the proponent.




(a) rebut a rebutter (b) rebut an undercutter (c) undercut a rebutter (d) undercut an undercutter

Figure 3.3: Proponent's counter-attacks of the opponent's attack.

(10) [We should tear the building down.]$_1$ [On the other hand, many people liked the view from the roof.]$_2$ [On weekends in summer, the roof is usually crowded with sunset partygoers.]$_3$

## 3.4 Counter-Attacks

How can the proponent respond to these challenges? Which possibilities are available to the author to counter the anticipated attacks? Freeman identified several ways to defend an argument. We will present what we regard as the most important ones.

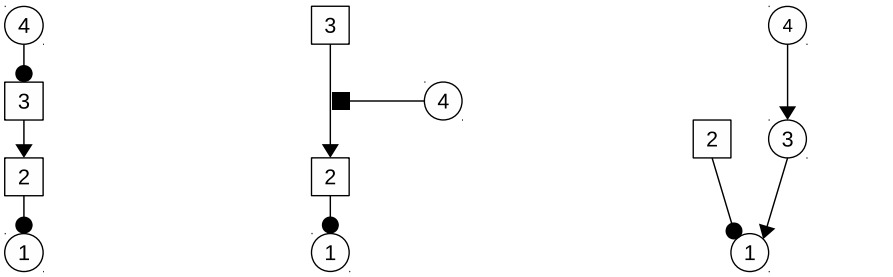
**Rebut a rebutter:** If the attack itself was a rebutter, then the counter-rebutter is an argument for the negation of the rebutter, i.e. the author is for some reason denying the anticipated argument against her original claim, as in example (11). For the corresponding structure see Figure 3.3a.

(11) [We should tear the building down,]$_1$ [even though it's supposed to be some touristic attraction.]$_2$ [But, I've never seen any visitor groups there!]$_3$

**Rebut an undercutter:** If the attack itself was an undercutter, then the counter-rebutter is an argument for the negation of the undercutter, i.e. the author is denying that the exception holds. This is the case in (12). It may be that the exception would undercut her argument if it were true, but it is not. An example diagram is shown in Figure 3.3b.

(12) [We should tear the building down.]$_1$ [It is full of asbestos.]$_2$ [Some new scientific study reportedly considers asbestos harmless,]$_3$ [but that is probably only a hoax.]$_4$

**Undercut a rebutter:** Undercutting a rebutter means to present an exception to the argument for the negated conclusion. The author not only shows that the anticipated argument against her claim needs to be restricted, but also that the argument is irrelevant for her claim, because the exception holds. An example would be (13). Figure 3.3c illustrates this structure.




(a) rebut an attack's defence (b) undercut an attack's defence (c) counterconsideration

Figure 3.4: Further strategies of counter-attacks.

(13) [We should tear the building down,]$_1$ [even though it's supposed to be some touristic attraction.]$_2$ [They'll surely build something more attractive on the site.]$_3$

**Undercut an undercutter:** Undercutting an undercutter correspondingly means to present an exception to an exception. The author does not even need to address whether the anticipated exception to her argument holds or not, because she can show that the anticipated exception itself is rendered irrelevant due to an exception. This constellation is shown in Figure 3.3d. For instance, consider the following argumentation in (14):

(14) [We should tear the building down.]$_1$ [It is full of asbestos.]$_2$ [In principle it is possible to clean it up,]$_3$ [but that would be forbiddingly expensive.]$_4$

While distinguishing rebutters from undercutters seemed possible though not trivial for the opponent's attacks, we expect it to be an easier task for the proponent's counter-attacks. Since the basic dialectical situation only forbids the opponent to assert but not the proponent, it is likely that strong linguistic signals are found when (in rebutting) the negation of the target is actually claimed, or when (in undercutting) the exception is actually claimed to be holding.

**Undermine an attack's defence:** Given that the opponent provided support for his objection by additional arguments, another strategy to counter his objection is in attacking those supporting arguments. In this case, the proponent is arguing against the cogency of the argument in favour of the objection and thus diminishing its strength. The argument can be attacked either by rebutting the premise in favour of the objection (Figure 3.4a), or by undercutting the support of the premise for the objection (as shown if Figure 3.4b). For the sake of brevity we will not present further full examples. The interested reader is invited to extend the given examples accordingly.

**Counterconsiderations:** The last possibility to react to an attack is to leave it uncountered. At first glance this seems counterproductive to the author's goal to convince the

reader of her main claim. However, this appears frequently in commentary text. By leaving a rebutter uncountered, the author assumes that the arguments presented in favour of the claim will outbalance the arguments against the claim, either because the rebutting attack is conceived to be of only minor strength, or because the pro arguments are seen as especially important. This had been discussed under the term *counterconsiderations* in Section 2.3. Since the observation that a rebutter is for some reason not countered can only be made retrospectively, no additional structure is required to represent counterconsiderations in our scheme. This is trivially shown in Figure 3.4c, where a rebutting attack is simply followed by premises directly and indirectly supporting the main claim, leaving the rebutter uncountered.

## 3.5 Extensions: Applying the scheme to authentic text

So far, we have presented the 'pure' scheme, arguing from the need to represent abstract configurations of argument. When it comes to annotating authentic text, a few extensions are in order. These largely concern the role of segmentation. Argumentation theories often assume a clean list of the claims found in a real text, i.e. they do not work with the propositions expressed in the segments of the original text, but with a set of summarised extracts opportune for argumentative analysis. In contrast, our scheme presented here applies to real text segments. However, it requires these segments to be argumentative discourse units (ADUs), i.e. units corresponding to propositions that are argumentatively relevant and have their own argumentative function. Yet, not all elementary discourse units (EDUs) directly correspond to ADUs. Very often a practical, EDU-based annotation has to cope with the linguistic style of the author and the peculiarities of the segmentation process. We therefore propose three tools that enable the annotator to handle these typical problems and form ADUs from EDUs.

**Non-argumentative segments:** Not all segments have an argumentative function. Often there are parts of the text only serving the purpose to set the scene, to introduce the topic to be discussed, or to state factual background information that is not used in any argument. Sometimes, the author is simply digressing from the topic, or dealing a side-blow. All these segments are considered non-argumentative (relative to the main claim). Since these segments will not constitute an ADU, no node with this segment number will appear in the argumentation graph. As an example consider (15), where the first segment is setting the scene, rather than presenting an argumentatively relevant claim:

(15) [Take out your pencils and write that down:]$_1$ [Dictation tests are barely improving pupil's spelling skills.]$_2$ [The obligation for dictation exams at least once a year is antiquated.]$_3$

**Joining adjacent EDUs:** From time to time, texts contain segments that only present a partial proposition. There are examples where these have to be mentally completed by the annotator, who then has to decide whether the completed proposition is argumentatively relevant or not. We also find instances where the partial proposition expressed in a segment can be completed by joining it with a segment next to it. We thus allow the annotators to combines multiple adjacent EDUs into one ADU, if the resulting unit can be understood as one proposition. Consider the three segments in example (16). All of them can be read as a single proposition. In an argument diagram, these three segments would thus be represented by a single ADU node, which is named after the involved segments, separated by commas.

(16) [The building is contaminated with asbestos.]$_1$ [In every single corner.]$_2$ [From the first to the last floor!]$_3$

**Restatements:** Finally, we sometimes find *restatements*, usually of the main claim of the argumentation. Therefore, if two (non-adjacent) segments appear to express the same proposition, both segments can be represented as one node, with an equation of the involved segments as its label. By doing this we ensure that there are no duplicate ADUs in our argumentation structure. Restatements typically occur in longer texts that start with the main claim, and after having developed the argument come back to the final conclusion in the end. An example is given in (17):

(17) [We should tear the building down.]$_1$ [Not only is it ...]$_2$ [Also, ...]$_3$ [Finally, ...]$_4$ [It's time for a demolition.]$_5$

As an illustration of those extensions to handle text segmented into EDUs, see Figure 3.5. It shows the argumentation structure of an imaginary text consisting of six ADUs which are based on a total of nine EDUs. The first segment was not argumentatively relevant and hence does not occur in the graph. The second segment represents the main claim, and a simple argument follows. Another argument follows and is undercut by the fifth segment. This is countered by a rebuttal through the segments six and seven, which have been joined to one ADU. The eighth segment is a last supporting argument, followed by a restatement of the main claim.

## 3.6 Conclusion

In this chapter, we have presented a synthesised scheme for representing the structure of argumentation in authentic text. It is based to a large part on the theory of Freeman [1991,

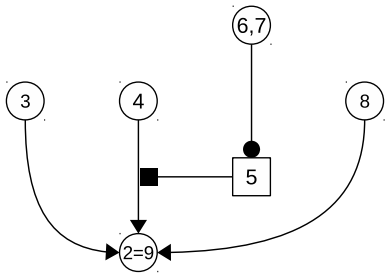



Figure 3.5: Supplemental features

2011], but we revised it to represent complex structures, as we think, more elegantly.[4] The main difference to Freeman's original formulation are:

1. We represent each argumentative attack or counter-attack by an individual relation. Co-indexation of arguments is not required any longer.

2. We make an explicit distinction between rebutting and undercutting attacks, in the argument diagram and in the formulation of dialectical challenges of the proponent.

3. We represent defended attacks using the very same structural principles of support that we have on the proponent side, i.e. we prefer a relational representation over the introduction of complex nested node types.

4. We specified extended features to cope with segmentation issues typically occurring in authentic text: The process allows the annotators to build a segmentation into ADUs from the text's EDUs, including cases of non-argumentative segments, ADUs spanning over multiple adjacent EDUs, and restatements.

The scheme fulfils our requirements defined in Chapter 2: We have explicit means for representing the inferential relations of support, attack, and counter-attacks. Dialectical roles are directly represented in the proponent and opponent node types. The complex formation is compositional and not restricted in a way that prohibits non-linear or long-distance dependencies. From our desiderata, we furthermore consider text genre and domain independence fulfilled, given the generality of Freeman's theory. Before we can start to create an annotated corpus of argumentation structures using this scheme, however, we first have to prove that this scheme yields reliable annotations. This question will be investigated in the next chapter.

[4] It can be considered as a subset of what can be represented in the Argument Interchange Format [Chesñevar et al., 2006].

# 4 Agreeing on the structure of argumentation

In the previous chapters, we first reviewed theories of the structure of argumentation and then devised an annotation scheme, based on the theoretic work of Freeman [1991, 2011]. This scheme is formulated and has been published as a scientific contribution, but is not meant as an instructive guide to be used for the actual annotation. For this purpose, we formulated annotation guidelines (or a coding manual).

The purpose of this chapter is to report on our annotation experiments and ultimately to show that the scheme presented in the previous chapter can be used reliably for the creation of a corpus of argumentation structures. The required methodology will be introduced in the next section. We will then present the results of three annotation experiments, one on annotating full argumentation structures in short texts (Section 4.2), one on distinguishing certain types of argumentative attacks in short texts and in pro and contra commentaries (Section 4.3), and finally one experiment on argumentative zones for pro and contra commentaries (Section 4.4).

**Previously published material**

Section 4.2 contains results that have been previously published in different workshop papers: The results for naive annotators had been presented in [Peldszus and Stede, 2013a], those for the expert annotators in a less elaborate way in [Peldszus, 2014]. The intermediary group of annotators was not publicly reported on yet. Our approach to cluster annotators in order to study structure of agreement (Section 4.1.3) was initially presented in [Peldszus and Stede, 2013a]. Sections 4.3 and 4.4 have not been published yet. Annotation guidelines can be found in [Stede, 2016a].

## 4.1 Methodology

### 4.1.1 Measuring reliability

Various coefficients have been defined for assessing the inter-annotator agreement. A good overview of the coefficients, their properties, use cases, and sometimes misleading terminological history is given by Artstein and Poesio [2008] and we will follow their formalisation. One important property of those coefficients is that they are chance corrected, i.e. they estimate how much agreement is to be expected by chance due to the frequency of the categories, and then calculate the agreement above chance. In our annotation (and also classification) experiments, we will use the following coefficients:

- **Cohen's kappa $\kappa$:** A coefficient for two coders that estimates the expected agreement based on the individual annotator's category distribution has been presented by Cohen [1960]. The general form of this coefficient is $\kappa = \frac{\text{AO}-\text{AE}}{1-\text{AE}}$, where AO stands for the observed agreement and AE for the expected agreement.[1]

- **Fleiss' kappa $\kappa$:** Prior to Cohen's $\kappa$, Scott [1955] presented the $\pi$ coefficient, which estimates the expected agreement differently: Instead of using the category distributions of the individual annotator, it assumes a single category distribution over all annotators. Fleiss [1971] later generalised this metric for multiple annotators, likewise giving it the name $\kappa$. The general form of the coefficient is equal to Cohen's $\kappa$.

- **Krippendorff's alpha $\alpha$:** A complimentary formalisation of an agreement coefficient, which is based on measuring *disagreement*, has been proposed by Krippendorff [1980]. The estimation of disagreement relies on a single category distribution over all annotators, similar to Fleiss' $\kappa$. Krippendorff's $\alpha$ allows multiple annotators. Furthermore, the disagreement can be weighted. In principle every distance function can serve to define the weight of disagreement, which makes $\alpha$ a very versatile coefficient. The general form is $\alpha = 1 - \frac{\text{DO}}{\text{DE}}$, where DO stands for observed disagreement and DE for expected disagreement.

The agreement measured with these coefficients ranges between -1 and 1, where 1 is perfect agreement, 0 is chance agreement and values below 0 signal agreement below chance.

The interpretation of this scale is not trivial. Krippendorff [1980] interprets the strength of agreement values quite conservatively and considers only values above 0.8 as good reliability and values above 0.67 to only allow "highly tentative and cautious conclusions". Landis and Koch [1977] on the other hand propose a more permissive interpretation, where values above 0.4 are considered as moderate agreement, above 0.6 as substantial, and above

[1] We will use this metric only for reporting agreement in automatic classification, where there are only two coders – the gold standard and the system's prediction.

0.8 as perfect agreement. The achievable agreement most certainly also depends on the complexity and difficulty of the task: Agreement in discourse annotation typically does not reach the level achieved in shallower annotation task, as e.g. in part of speech tagging. A POS-tagging study that only reaches an agreement of 0.7 above chance is not acceptable, while this value might be already satisfactory for highly interpretatory tasks of discourse annotation such as subjectivity, coherence relation, rhetorical, or argumentative structure. Often, this level can only be surpassed by rather extensive training of the annotators.[2]

Besides *inter*-annotator agreement, these coefficients are also used to measure the stability of a scheme over time: *intra*-annotator agreement. To this end, the same annotator codes the same items at two different points in time, with a reasonable pause in between. This measure is not reported frequently in the CL literature, with some notable exceptions [Teufel, 2010]. Usually, before and after agreement scores are reported only when the guidelines have been improved or the annotators have been trained more extensively.

### 4.1.2 Investigating confusions

When it comes to investigation of the reasons of disagreement, the informativeness of a single inter-annotator agreement value is limited. We want to identify sources of disagreement in both the set of annotators as well as the categories. To this end, contingency tables (confusion matrices) are studied, which show the number of category agreements and confusions for a pair of annotators.

For larger numbers of annotators, studying the confusion matrices for each pair of annotator is infeasible. One solution to nevertheless get an overview of typical category confusions, is to build an **aggregated confusion matrix**, which sums up the values of category pairs across the normal confusion matrices of all possible pairs. This aggregated confusion matrix can be used to derive a **confusion probability matrix**, as proposed in Cinková et al. [2012]. It specifies the conditional probability that one annotator will annotate an item with (column) category, given that another has chosen a (row) category, so the rows sum up to 1. The diagonal cells then display the probability of agreement for each category.

Krippendorff [1980] proposed another way to investigate category confusions, sometimes referred to Krippendorff diagnostics. Two different tests are proposed. Both systematically compare the agreement on the original category set with the agreement on a reduced category set. They differ in how they collapse categories:

The first is the **category definition test**, where all but the one category of interest are collapsed together, yielding a binary category distinction. When measuring the agreement with this binary distinction only confusions between the category of interest and the rest

[2] Agreement of professional annotators on 16 rhetorical relations was $\kappa$=0.64 in the beginning and 0.82 after extensive training [Carlson et al., 2003]. Agreement on 'argumentative zones' is reported $\kappa$=0.71 for trained annotators with detailed guidelines, another study for untrained annotators with only minimalist guidelines reported values varying between 0.35 and 0.72 (depending on the text), see Teufel [2010].

count, but no confusions between the collapsed categories. If agreement increases for the reduced set compared to the original set, that category of interest is better distinguished than the rest of the categories.

The other of Krippendorff's diagnostics is the **category distinction test**, where two categories are collapsed in order to measure the impact of confusions between them on the overall agreement value. The higher the difference, the greater the confusion between the two collapsed categories.

### 4.1.3 Ranking and clustering annotators

In experiments with many annotators, it is of interest to get a better understanding of the characteristics of the annotators. Are there especially good or bad annotators, are there groups of annotators each following a common pattern? Are there ambiguities in the guidelines which cause a systematically different annotation result? All this cannot be read from a single inter-annotator agreement score, nor from confusion matrices or Krippendorff's diagnostics accumulated over all annotators.

A first step towards this is to analyse and compare the individual annotator's category distributions. They can be easily understood and visualised without the need for a combinatorial exploding pair-wise comparison. From this one can learn about the preferences or biases of all annotators.

We propose another way to investigate the similarities and differences between the annotators and thus identify the structure of agreement in the group of annotators: Our idea is to use **agglomerative hierarchical clustering** to group annotators. The clusters are initialised as singletons for each annotator. Then agreement is calculated for all possible pairs of those clusters. The pair of clusters with the highest agreement is merged. This procedure is iterated until there is only one cluster left. In contrast to normal clustering, the linkage criterion does not determine the distance between complex clusters indirectly as a function of the distance between singleton clusters, but directly measures agreement for the unified set of annotators of both clusters.

An example clustering is shown in the dendrogram in Figure 4.1a. The x-axis marks the different annotators, the y-axis specifies the agreement for the merged clusters of annotators. The dendrogram gives an impression of the possible range of agreement: The best pair of annotators being N06 and N08 with a $\kappa \approx 0.85$, the whole group of annotator achieving a $\kappa \approx 0.81$. Furthermore, it allows us to check for ambiguities in the guidelines: If there were stable alternative readings in the guidelines, we would expect multiple larger clusters that can only be merged at a lower level of $\kappa$. This is simulated in Figure 4.1b, where half of the annotators exhibit a category confusion: We observe two larger clusters, both at an agreement level of around $\kappa \approx 0.79$. Both clusters can be merged only at a significantly lower level around $\kappa \approx 0.67$. The existence of such cluster configuration can thus serve

as an indicator for systematic differences across annotators which in turn might be due to ambiguous annotation guidelines.

## 4.2 Experiment 1: Argumentative structure of microtexts

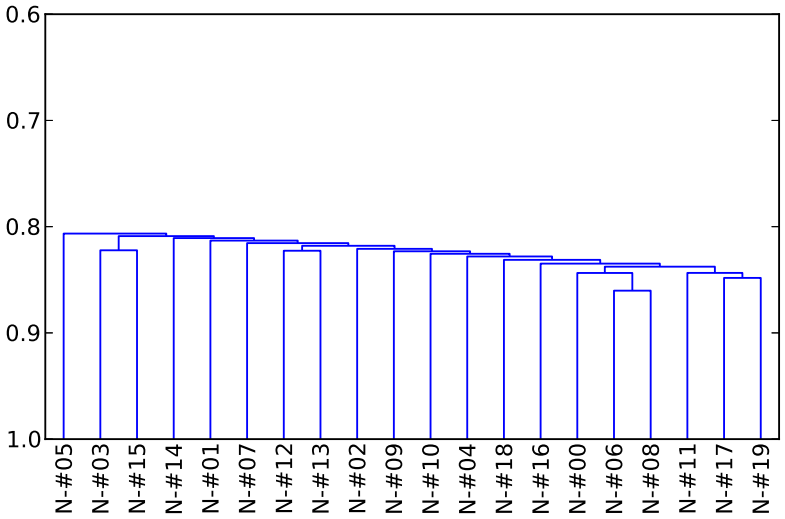
Our first annotation experiment will focus on evaluating the scheme for the structure of argumentation that we presented in Chapter 3. Annotators will have to come up with a full argumentation structure, which amounts to identifying the central claim of the text, determining the argumentative role (proponent versus opponent) for each text segment and then deciding how these segments are related to each other: Which segments are supporting the main claim or one of its premises? Where are possible objections and how are they countered?

### 4.2.1 Experimental setup

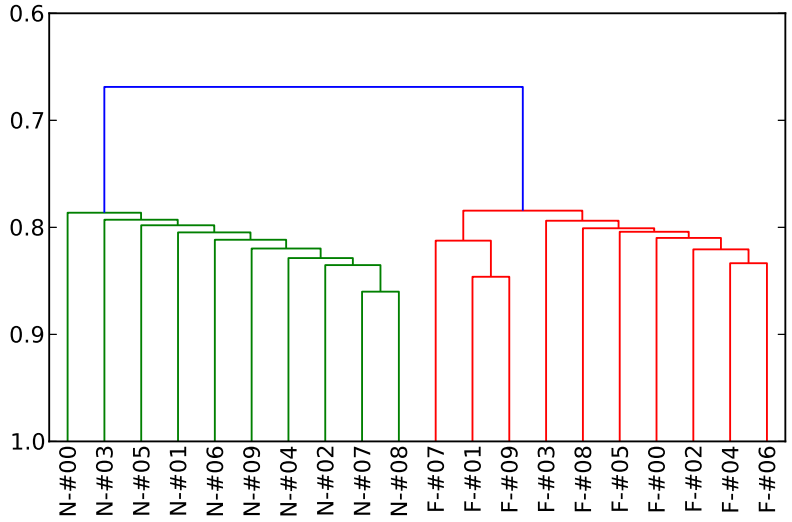
We developed annotation **guidelines** based on the theory presented in Chapter 3, see Appendix A. They are about six pages long. An extended version of them has been published as [Peldszus et al., 2016]. The guidelines contain text examples and the corresponding graphs for all basic structures, and they present different combinations of attack and counter-attack. The annotation process is divided into three steps: First, one segment is identified as the central claim of the text. The annotator then chooses the dialectical role (proponent or opponent) for all remaining segments. Finally, the argumentative function of each segment (is it supporting or attacking) and the corresponding subtypes have to be determined, as well as the targeted segment.

Three groups of **subjects** participated in our experiment: First, a group of 26 *students* (A01 - A26) participated in the experiment in the context of an undergraduate university course. Completion of the experiment was obligatory. All of them were native speakers of German. Another group of six students of a higher semester (T01 - T06), which we refer to as *experienced* annotators, participate in the study in the context of a course on rhetorical and argumentation structure of text. They had more experience with discourse analysis, and the annotation scheme had been covered in the course. Finally, three *expert* annotators (E01 - E03) completed the experiment, all of which were familiar with various discourse annotation tasks in general and with argumentative analysis. Two of them were the authors of the guidelines, the third a post-doc in computational linguistics.

For the **source material**, we considered newspaper arguments found 'in the wild' to be too challenging for a start, for applying the scheme demands a detailed, deep understanding of the text. We therefore chose to first evaluate this task on short and controlled instances of argumentation. For this purpose we built a set of 23 constructed German texts, where each text consists of only five discourse segments. While argumentative moves in authentic texts




(a) noise only




(b) noise and systematic disagreement

Figure 4.1: Clusterings of simulated annotators: Figure (a) shows a simulation of annotators each with a noisy version of a given labelling. In Figure (b), half of the simulated annotators exhibit systematic disagreement with the given labelling by introducing a category confusion.

are often surrounded by material that is not directly relevant to the argumentation, such as factual background information, elaborations, or rhetorical decoration, in the constructed texts all segments were clearly argumentative, i.e. they either present the central claim, a reason, an objection, or a counter-attack. Merging segments and identifying restatements was thus not necessary. The texts covered several combinations of the basic constructs in different linearisations, typically one central claim, two (simple, combined or exemplifying) premises, one objection (rebutting a premise, rebutting the conclusion, or undercutting the link between them), and a possible reaction (rebutting or undercutting counter-attacks, or a new reason that renders the objection uncountered). A (translated) example of a micro text is given in Figure 4.2.

The **procedure** of the annotation experiment was as follows: All annotators received only minimal training in the experiment: A short introduction (5 min.) was given to set the topic. After studying the guidelines (~30 min.) and a very brief opportunity to address questions, the subjects annotated the 23 texts (~45 min.), writing their analysis as an argumentative graph in designated areas of the questionnaire.

#### Interpreting argumentation graphs as segment labels

Since the annotators were asked to assign one and only one argumentative function to each segment, every node in the argumentative graph has exactly one out-going arc. The graph can thus be reinterpreted as a list of segment labels.

Every segment is labelled on different levels: The 'role'-level specifies the dialectical role (proponent or opponent). The 'typegen'-level specifies the general type, i.e. whether the segment presents the central claim (thesis) of the text, supports or attacks another segment. The 'type'-level additionally specifies the kind of support (normal or example) and the kind of attack (rebutter or undercutter). Whether a segment's function holds only in combination with that of another segment (combined) or not (simple) is represented on the 'combined'-level, which is roughly equivalent to Freeman's 'linked premises'. The target is finally specified by the segment identifier ($1 \ldots 5$) or relation identifier ($a \ldots d$) on the 'target'-level.

The labels of each separate level can be merged to form a complex tagset. We interpret the result as a hierarchical tagset as it is presented in Figure 4.3. The label 'PSNC(3)' for ex-

> [*Energy saving light bulbs contain a significant amount of toxins.*]$_1$ [*A commercially available bulb may contain for example up to five milligrams of mercury.*]$_2$ [*That's why they should be taken off the market,*]$_3$ [*unless they're unbreakable.*]$_4$ [*But precisely this is unfortunately not the case.*]$_5$

Figure 4.2: A translated example micro text (micro_d21)

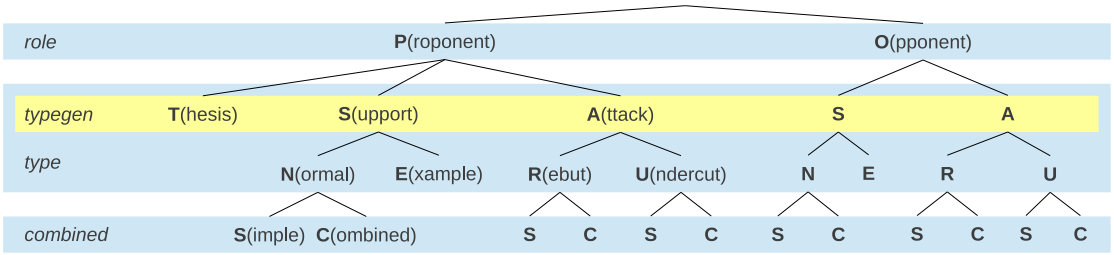



Figure 4.3: The hierarchy of segment labels.

ample stands for a proponent's segment, giving normal support to segment 3 in combination with another segment, while 'OAUS($b$)' represents an opponent's segment, undercutting a relation $b$, not combined. Note that this hierarchy is implicit in the annotation process. The annotators were neither confronted with a decision-tree version nor the labels of this tag hierarchy.

To sum up, the annotated graphs of the 23 texts can be transformed to nine different labellings (five for the basic levels and additional four combinations of them) over the in total 115 segments (see also Table 4.1).

### 4.2.2 Results

One question that arises before evaluation, especially in our setting, is how to deal with missing annotations, since measuring inter-annotator agreement with a $\kappa$-like coefficient requires a decision of every annotator (or at least the same number of annotators) on each item. One way to address this is to exclude annotators with missing annotations, another to exclude items that have not been annotated by every subject. In our experiment only 11 of the 26 subjects of the *student* group annotated every segment. Another ten annotated at least 90% of the segments, five annotated less. In the *experienced* group, there are likewise some, but fewer annotations missing. Excluding some annotators would be possible in our setting, but e.g. keeping only 11 of 26 annotators from the student group is unacceptable. Excluding items is also inconvenient given the small dataset. We thus chose to mark segments with missing annotations as such in the data, augmenting the tagset with the label '?' for missing annotations. We are aware of the undesired possibility that two annotators 'agree' on not assigning a category to a segment. However this false agreement occurred only very infrequently in our evaluations and we consider its impact on the overall result to be negligible.

| | | students | | | experienced | | | experts | | |
|---|---|---|---|---|---|---|---|---|---|---|
| level | # | $\kappa$ | AO | AE | $\kappa$ | AO | AE | $\kappa$ | AO | AE |
| role | 2 | .521 | .783 | .546 | .604 | .820 | .546 | 1.00 | 1.00 | .614 |
| typegen | 3 | .579 | .719 | .334 | .658 | .766 | .318 | .945 | .965 | .363 |
| type | 5 | .469 | .608 | .261 | .518 | .634 | .240 | .866 | .901 | .267 |
| comb | 2 | .458 | .727 | .497 | .449 | .697 | .449 | .895 | .948 | .503 |
| target | (9) | .490 | .576 | .169 | .568 | .638 | .162 | .850 | .878 | .185 |
| role+typegen | 5 | .541 | .656 | .251 | .641 | .729 | .245 | .952 | .965 | .269 |
| role+type | 9 | .450 | .561 | .202 | .509 | .607 | .199 | .875 | .901 | .209 |
| role+type+comb | 15 | .392 | .491 | .162 | .413 | .503 | .153 | .842 | .867 | .156 |
| role+type+comb+target | (71) | .384 | .436 | .084 | .424 | .471 | .083 | .831 | .846 | .088 |

Table 4.1: Agreement for all annotator groups for the different levels. The number of categories on each level (without '?') is shown in the second column (possible target categories depend on text length).

**Overall agreement**

The agreement in terms of Fleiss's $\kappa$ of all annotators on the different levels is shown in Table 4.1. The group of 26 student annotators reached an agreement $\kappa$>0.45 for all basic levels. Combining the levels to a complex tagset reduces the agreement. On the full task, which covers all aspects of the argumentation graph in one tagset, the students only agreed with $\kappa$=0.384. According to the scale of Krippendorff [1980], the annotators of the student group did neither achieve reliable ($\kappa \geq 0.8$) nor marginally reliable ($0.67 \leq \kappa < 0.8$) agreement in our experiment. On the scale of Landis and Koch [1977], most results can be interpreted to show moderate correlation ($0.4 < \kappa \leq 0.6$), only the two most complex levels fail. Although typical results in discourse structure tagging usually reach or exceed the 0.7 threshold, we expected lower results for three reasons: First, the minimal training of the naive annotators only based on the guidelines; second, the varying commitment to the task of the annotators in the obligatory setting; and finally the difficulty of the task, which requires a precise specification of the annotator's interpretation of the texts.

The more annotators of the *experienced* group achieve 5 to 7 points better results in general, except for the 'comb' level. In contrast, the three expert annotators achieve a very good agreement. On the basic levels the agreement is substantial with values around $\kappa\approx$0.9 and yet perfect agreement for the proponent opponent distinction. Even for the full task, the expert annotators' agreement is substantial, with $\kappa$=0.831.

For the complex levels we additionally report Krippendorff's $\alpha$ [Krippendorff, 1980] as a weighted measure of agreement. We use the distance between two tags in the tag hierarchy to weigh the confusion (similar to Geertzen and Bunt [2006]), in order to capture the intuition that confusing for instance PSNC with PSNS is less severe than confusing it with OAUS. The results are shown in Table 4.2. As expected, the agreement figures improve. We

| | | students | | | experienced | | | experts | | |
|---|---|---|---|---|---|---|---|---|---|---|
| level | # | $\alpha$ | DO | DE | $\alpha$ | DO | DE | $\alpha$ | DO | DE |
| role+typegen | 5 | .534 | .280 | .601 | .628 | .225 | .605 | .969 | .017 | .560 |
| role+type | 9 | .500 | .333 | .667 | .581 | .281 | .671 | .930 | .044 | .638 |
| role+type+comb | 15 | .469 | .378 | .710 | .531 | .335 | .715 | .903 | .067 | .690 |
| role+type+comb+target | (71) | .425 | .454 | .789 | .473 | .419 | .795 | .865 | .105 | .779 |

Table 4.2: Weighted agreement for all annotator groups for the combined levels.

observe an increase of 3 to 6 points on all complex levels, except 'role+type', for all groups of annotators. The increase is of course highly depending on the definition of the distance function, which is why Artstein and Poesio [2008] point out that the resulting values can neither be properly interpreted on the strength scales, nor should they be compared directly with unweighted scores.

**Category confusions**

The 'role' and 'comb' levels are binary decisions, for which an analysis of category confusions is of limited informativeness. For the other levels, we focus our discussion on confusions on the on the 'role+type' category level, since this is the most informative level in terms of granularity.

The high number of annotators in our study makes it infeasible to study the individual confusion matrices of all different pairs of annotators; it would be 325 pairs alone for the *student* group of annotators. We thus built an aggregated confusion matrix, which sums up the values of category pairs across all normal confusion matrices, and derived from it a **confusion probability matrix** [Cinková et al., 2012]. Table 4.3 shows the matrix for all three groups of annotators. Comparing these matrices reveals that the higher agreement of the experienced annotators and then the experts correlates with less confusion and much more probability mass on the diagonal cells. As an example, the probability that another annotator agrees when one annotator chooses the OAR label for an opponent's rebuttal is 0.339 for the student group, 0.478 for the experienced, and 0.794 for the expert group. Some less frequent labels, such as proponent's example support (PSE) or the opponent strengthening his own argument (OSN), caused confusions in the student and the experienced group, but were very reliably annotated by the experts. Note that the opponent supporting his own argument by examples (OSE) is a possible category, but not supposed to be found in the texts.

Two confusions are especially important: All annotator groups confused (to varying degrees) the attack subtypes rebutter and undercutter. Distinguishing them is very hard for both student and experienced annotators and still challenging for the experts. We will study this more deeply in a dedicated follow-up annotation experiment (see Section 4.3).

| | PT | PSN | PSE | PAR | PAU | OSN | OSE | OAR | OAU | ? |
|---|---|---|---|---|---|---|---|---|---|---|
| PT | **.625** | .243 | .005 | .003 | .002 | .006 | | .030 | .007 | .078 |
| PSN | .123 | **.539** | .052 | .034 | .046 | .055 | .001 | .052 | .021 | .078 |
| PSE | .024 | .462 | **.422** | .007 | .008 | | | .015 | .001 | .061 |
| PAR | .007 | .164 | .004 | **.207** | .245 | .074 | | .156 | .072 | .071 |
| PAU | .007 | .264 | .005 | .290 | **.141** | .049 | | .117 | .075 | .052 |
| OSN | .016 | .292 | | .081 | .046 | **.170** | .004 | .251 | .075 | .065 |
| OSE | | .260 | | | | .260 | | .240 | .140 | .100 |
| OAR | .033 | .114 | .004 | .070 | .044 | .102 | .001 | **.339** | .218 | .076 |
| OAU | .017 | .101 | | .069 | .061 | .066 | .002 | .469 | **.153** | .063 |
| ? | .179 | .351 | .031 | .066 | .041 | .055 | .001 | .157 | .061 | **.057** |

(a) student

| | PT | PSN | PSE | PAR | PAU | OSN | OSE | OAR | OAU | ? |
|---|---|---|---|---|---|---|---|---|---|---|
| PT | **.654** | .234 | .007 | | | | | .005 | | .100 |
| PSN | .120 | **.549** | .068 | .032 | .036 | .033 | .008 | .060 | .005 | .088 |
| PSE | .033 | .582 | **.264** | .011 | | | .033 | | | .077 |
| PAR | | .160 | .006 | **.218** | .410 | | | .032 | .045 | .128 |
| PAU | | .160 | | .366 | **.314** | | | .029 | .069 | .063 |
| OSN | | .406 | | | | **.328** | .172 | .047 | .047 | |
| OSE | | .300 | .150 | | | .550 | | | | |
| OAR | .006 | .136 | | .014 | .014 | .009 | | **.478** | .241 | .101 |
| OAU | | .027 | | .047 | .080 | .020 | | .553 | **.200** | .073 |
| ? | .183 | .315 | .032 | .091 | .050 | | | .160 | .050 | **.119** |

(b) experienced

| | PT | PSN | PSE | PAR | PAU | OSN | OSE | OAR | OAU | ? |
|---|---|---|---|---|---|---|---|---|---|---|
| PT | **.918** | .082 | | | | | | | | |
| PSN | .048 | **.905** | | | .048 | | | | | |
| PSE | | | **1.00** | | | | | | | |
| PAR | | | | **.680** | .320 | | | | | |
| PAU | | .182 | | .242 | **.576** | | | | | |
| OSN | | | | | | **1.00** | | | | |
| OSE | | | | | | | | | | |
| OAR | | | | | | | | **.794** | .206 | |
| OAU | | | | | | | | .667 | **.333** | |
| ? | | | | | | | | | | |

(c) experts

Table 4.3: Confusion probability matrix on the 'role+type' level for (a) 26 student annotators, (b) six experienced annotators and (c) three expert annotators. Empty cells correspond to zero probability. The '?' category stands for unlabelled annotation items.

| | students | | | experienced | | | experts | | |
|---|---|---|---|---|---|---|---|---|---|
| category | $(\Delta)\kappa$ | AO | AE | $(\Delta)\kappa$ | AO | AE | $(\Delta)\kappa$ | AO | AE |
| *base* | .450 | .561 | .202 | .509 | .607 | .199 | .875 | .901 | .209 |
| PT | +.265 | .912 | .691 | +.232 | .919 | .689 | +.071 | .983 | .677 |
| PSN | +.082 | .785 | .539 | +.042 | .796 | .545 | +.048 | .965 | .546 |
| PSE | +.128 | .970 | .928 | -.112 | .961 | .936 | +.125 | 1.00 | .901 |
| PAR | -.148 | .924 | .891 | -.189 | .929 | .896 | -.078 | .977 | .886 |
| PAU | -.240 | .930 | .912 | -.068 | .930 | .876 | -.166 | .959 | .861 |
| OSN | -.198 | .927 | .903 | -.028 | .975 | .952 | +.125 | 1.00 | .917 |
| OSE | -.451 | .999 | .999 | -.515 | .988 | .988 | -.875 | 1.00 | 1.00 |
| OAR | -.027 | .858 | .754 | +.077 | .896 | .748 | -.015 | .959 | .709 |
| OAU | -.229 | .916 | .892 | -.213 | .930 | .901 | -.400 | .959 | .922 |

Table 4.4: Krippendorff's category definition diagnostic for the level 'role+type'.

Furthermore, there are confusions between undercutting counter-attacks of the proponent (PAU) and normal support (PSN). This is a typical annotation conflict that is not always easy to resolve. The annotators have to decide whether a proponent's segment following a possible objection is intended to undercut the attack inference of the objection, rendering it irrelevant for a reason, or whether it is opening up a new supporting argument for the attacked claim, this merely indirectly out-weighing the aforementioned possible objection.

To assess the quality of each category's definition, we apply Krippendorff's **category definition test** (see Table 4.4). It shows the highest distinguishability is found for PT, PSN, and PSE for both students and experts; the latter also exhibit a high value for OSN. Rebutters are better distinguished for the opponent role than for the proponent role. Undercutters seem equally problematic for both roles. The extreme value for OSE is not surprising, given that this category was not supposed to be found in the dataset and was only used twice by the students and never by the experts. It shows, though, that the results of this test have to be interpreted with caution for rare categories, since in these cases the collapsed rest always leads to a very high chance agreement.

The other of Krippendorff's diagnostics is the **category distinction test**, where two categories are collapsed in order to see how much agreement is lost due to confusions between them. We report the results for this test in Table 4.5, showing only the problematic pairs with a positive $\Delta\kappa \geq 0.005$. The highest gain of nearly 5 points $\kappa$ is observed when collapsing rebutting and undercutting attacks on the opponents side. On the proponent side this confusion is exists but less pressing. This holds for all annotator groups. The distinction between direct counter-attack and new (potentially outweighing) support (PAU versus PSN) has an impact for the expert annotators, but is not affecting the agreement of the student

| category pair | | $(\Delta)\kappa$ | AO | AE |
|---|---|---|---|---|
| *base* | | .450 | .561 | .202 |
| OAR | OAU | +.048 | .608 | .219 |
| PAR | PAU | +.026 | .585 | .208 |
| OAR | OSN | +.018 | .583 | .217 |
| PSN | PSE | +.012 | .586 | .229 |
| OAR | PAR | +.007 | .576 | .219 |
| PSN | OSN | +.007 | .587 | .239 |
| PAR | OSN | +.005 | .568 | .208 |

(a) students

| category pair | | $(\Delta)\kappa$ | AO | AE |
|---|---|---|---|---|
| *base* | | .509 | .607 | .199 |
| OAR | OAU | +.052 | .655 | .214 |
| PAR | PAU | +.042 | .644 | .206 |
| PSE | PSN | +.025 | .638 | .222 |
| OSN | PSN | +.008 | .622 | .216 |
| OSE | OSN | +.008 | .613 | .199 |

(b) experienced

| category pair | | $(\Delta)\kappa$ | AO | AE |
|---|---|---|---|---|
| *base* | | .875 | .901 | .209 |
| OAR | OAU | +.050 | .942 | .223 |
| PAR | PAU | +.028 | .925 | .218 |
| PAU | PSN | +.015 | .919 | .261 |

(c) experts

Table 4.5: Krippendorff's category distinction diagnostic for the level 'role+type'.

and the experienced group, which is probably because they confuse PAU and PSN with various other labels (see their confusion matrix in Figure 4.3). A problem that only exists for the student group is the confusion between opponents rebuts and supports (OAR versus OSN): This occurs because some annotators mixed up the distinction between argumentative role and argumentative function type.

### Comparison with gold data

After the experiment, the expert annotators compared their annotations and agreed on a gold standard. In the following, we will compare the results of the first group of student annotators with this gold standard. For each annotator and for each level of annotation, we calculated the F1 score, macro-averaged over the categories of that level. Figure 4.4 shows the distribution of those values as boxplots. We observe varying degrees of difficulty on the basic levels: While the scores on the 'role' and 'typegen' are relatively dense between 0.8 and 0.9, the distribution is much wider and also generally lower for 'type', 'comb' and 'target'. Especially remarkable is the drop of the median when comparing 'typegen' with 'type': For the simpler level, all values of the better half of annotators lie above 0.85, but for the more complex level, which also requires the distinction between rebutters and undercutters, the median drops to 0.67. The figure also shows the pure F1 score for identifying the central claim (PT). While the larger part of the annotators performs well in this task, some are still below 0.7. This is remarkable, since identifying one segment as the central claim of a five-segment text does not appear to be a challenging task.

### Ranking and clustering the annotators

Until now we have mainly investigated the tagset as a factor in measuring agreement. The widespread distribution of annotator scores in the comparison with gold standard, however, showed that their performance differs greatly. As described in the section on the experimental setup, participation in the study was obligatory for our student subjects. We thus want to make sure that the differences in performance are a result of the annotators' varying commitment to the task, rather than a result of possible ambiguities or flaws of the guidelines. The inter-annotator agreement values presented in Table 4.1 are not so helpful in answering this question, as they only provide us with an average measure, and not with an upper and lower bound of what is achievable with our annotators. Consequently, the goal of this section is to give structure to the set of annotators by imposing a (partial) order on it, or even by dividing it into different groups and investigate their characteristic confusions.

**Central claim:** During the conversion of the written graphs into segment label sequences, it became obvious that certain annotators nearly always chose the first segment of the text as the central claim, even in cases where it was followed by a consecutive clause with a

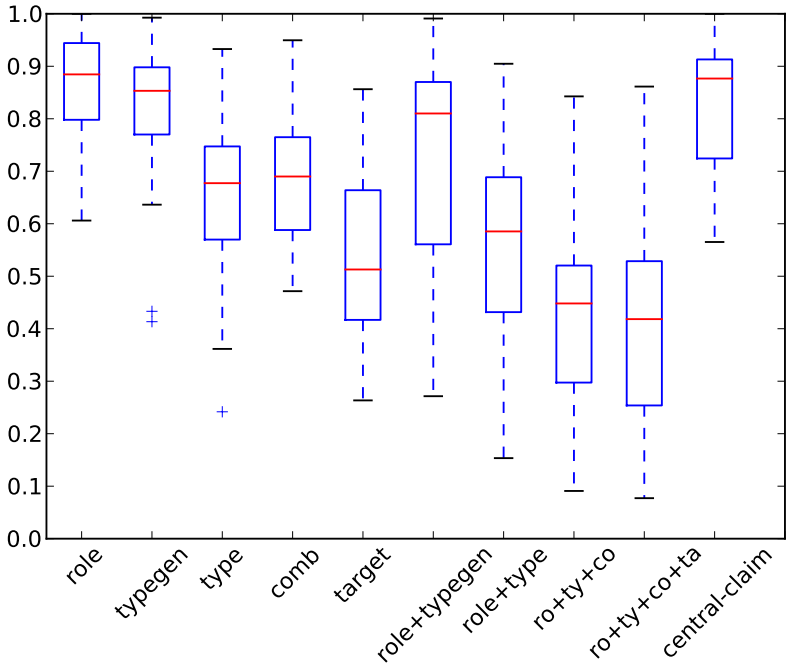



Figure 4.4: Comparison of student annotations with gold standard: For each level we show a boxplot of the F1 scores of all annotators (each score macro-averaged over categories of that level). Also, we present the F1 score for the recognition of the central claim.

discourse marker. Therefore, our first heuristic was to impose an order on the set of annotators according to their F1 score in identifying the central claim. This not only identifies outliers but can additionally serves as a rough indicator of text understanding. Although this ordering requires gold data, producing gold data for the central claim of these texts is relatively simple and using them only gives minimal bias in the evaluation (in contrast to e.g. 'role+type' F1 score as a sorting criterion). In our case, gold data are available, as described above. However, consider a scenario where outlier-annotators were to be sorted out without having a full gold annotation at hand. In this scenario central-claim annotations might serve as a convenient and simple vehicle for filtering. With this ordering we can then calculate agreement on different subsets of the annotators, e.g. only for the two best annotators, for the ten best, or for all. Figure 4.5 shows $\kappa$ on the different levels for all $n$-best groups of annotators: From the two best to the six best annotators the results are quite stable. The six best annotators achieve an encouraging $\kappa$=0.74 on the 'role+type' level and likewise satisfactory $\kappa$=0.69 for the full task, i.e. on the maximally complex 'role+type+comb+target' level. For increasingly larger $n$-best groups, the agreement decreases steadily with only minor fluctuations. Although the central claim F1 score

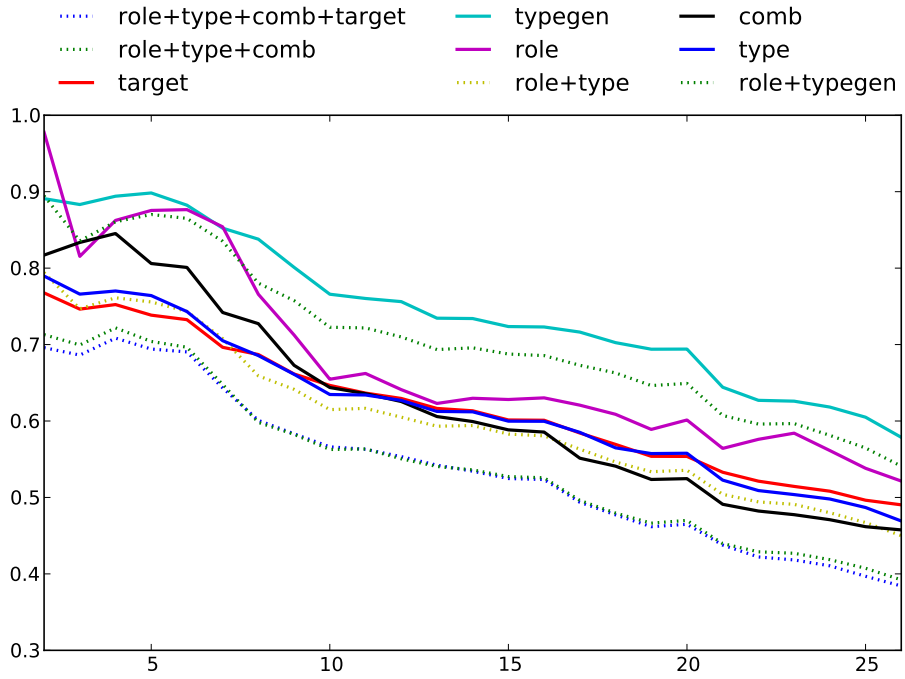



Figure 4.5: Agreement in $\kappa$ on the different levels for the $n$-best annotators ordered by their F1 score in identifying the central claim.

proves to be a useful sorting criterion here, it might not work as well for authentic texts due to the possibility of restated, or even implicit central claims.

**Category distributions:** Investigating the annotator bias is also a promising way to impose structure onto the group of annotators. A look on the individual distribution of categories per annotator quickly reveals that there are some deviations. Table 4.6 shows the individual distributions for the 'role+type'-level, as well as the average annotator distribution, and that found in the gold data. We focus on three peculiarities here. First, both annotators A18 and A21 refrain from classifying segments as attacking. Although they make the distinction between the roles, they give only supporting segments. Checking the annotations shows that they must have mixed the concepts of dialectical role and argumentative function. Another example is the group of A04, A20, and A23, who refrain from using proponent attacks. Although they make the distinction between the argumentative functions of supporting and attacking, they do not systematically attribute counter-attacks to the proponent. Finally, as pointed out before, there are several annotators with a different amount of missing annotations. Note that missing annotations must not necessarily signal an unmotivated annotator (who skips an item if deciding on it is too tedious). It could very well also be a diligent but slow annotator. Still, missing annotations lead to lower agreement

| | categories | | | | | | | | | | deviation | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| annotator | PT | PSN | PSE | PAR | PAU | OSN | OSE | OAR | OAU | ? | $\Delta^{gold}$ | $\Delta^{\varnothing}$ |
| A01 | 23 | 40 | 5 | 13 | 0 | 6 | 0 | 24 | 0 | 4 | 17 | 15.6 |
| A02 | 22 | 33 | 7 | 8 | 11 | 3 | 0 | 23 | 1 | 7 | 17 | 16.9 |
| A03 | 23 | 40 | 6 | 4 | 12 | 5 | 0 | 16 | 9 | 0 | 7 | 11.8 |
| A04 | 21 | 52 | 6 | 1 | 0 | 0 | 0 | 14 | 11 | 10 | 25 | 20.5 |
| A05 | 23 | 42 | 5 | 15 | 2 | 5 | 0 | 20 | 3 | 0 | 10 | 14.2 |
| A06 | 24 | 39 | 6 | 6 | 9 | 7 | 0 | 15 | 9 | 0 | 7 | 10.9 |
| A07 | 22 | 41 | 1 | 12 | 8 | 5 | 0 | 13 | 8 | 5 | 13 | 9.4 |
| A08 | 23 | 35 | 6 | 6 | 14 | 6 | 1 | 17 | 7 | 0 | 9 | 13.3 |
| A09 | 23 | 43 | 2 | 6 | 7 | 7 | 0 | 15 | 12 | 0 | 9 | 10.8 |
| A10 | 23 | 51 | 3 | 3 | 4 | 8 | 0 | 8 | 15 | 0 | 21 | 21.2 |
| A11 | 21 | 41 | 3 | 2 | 1 | 1 | 0 | 22 | 9 | 15 | 21 | 16.6 |
| A12 | 23 | 42 | 6 | 15 | 5 | 3 | 0 | 13 | 4 | 4 | 13 | 11.7 |
| A13 | 23 | 40 | 4 | 16 | 0 | 7 | 0 | 17 | 8 | 0 | 14 | 13.3 |
| A14 | 19 | 33 | 6 | 10 | 4 | 4 | 0 | 11 | 8 | 20 | 26 | 20.2 |
| A15 | 19 | 37 | 2 | 6 | 7 | 3 | 0 | 18 | 3 | 20 | 20 | 16.9 |
| A16 | 20 | 31 | 4 | 7 | 10 | 7 | 0 | 14 | 5 | 17 | 22 | 16.9 |
| A17 | 22 | 53 | 2 | 4 | 3 | 0 | 0 | 20 | 6 | 5 | 17 | 15.1 |
| A18 | 23 | 51 | 5 | 0 | 0 | 34 | 1 | 0 | 1 | 0 | 39 | 40.4 |
| A19 | 24 | 41 | 7 | 13 | 2 | 5 | 0 | 20 | 3 | 0 | 10 | 14.5 |
| A20 | 21 | 41 | 4 | 0 | 1 | 2 | 0 | 31 | 5 | 10 | 22 | 18.2 |
| A21 | 16 | 40 | 0 | 1 | 0 | 20 | 0 | 0 | 1 | 37 | 52 | 44.8 |
| A22 | 22 | 34 | 7 | 5 | 10 | 6 | 0 | 17 | 9 | 5 | 12 | 10.3 |
| A23 | 23 | 52 | 0 | 1 | 0 | 0 | 0 | 32 | 6 | 1 | 24 | 27.1 |
| A24 | 23 | 41 | 6 | 6 | 9 | 5 | 0 | 22 | 3 | 0 | 4 | 11.8 |
| A25 | 23 | 38 | 4 | 5 | 15 | 0 | 0 | 7 | 23 | 0 | 24 | 27.1 |
| A26 | 23 | 44 | 5 | 8 | 4 | 4 | 0 | 21 | 3 | 3 | 9 | 10.2 |
| ∅ | 22.0 | 41.3 | 4.3 | 6.7 | 5.3 | 5.9 | 0.1 | 16.5 | 6.6 | 6.3 | | |
| gold | 23 | 42 | 6 | 6 | 8 | 5 | 0 | 19 | 6 | 0 | | |

Table 4.6: Distribution of categories for each annotator in absolute numbers for the 'role+type' level. The last two rows display gold and average annotator distribution for comparison. The two right-most columns specify for each annotator the total difference to gold or average distribution $\Delta^{gold/\varnothing} = \frac{1}{2}\sum_c \Delta_c^{gold/\varnothing}$.

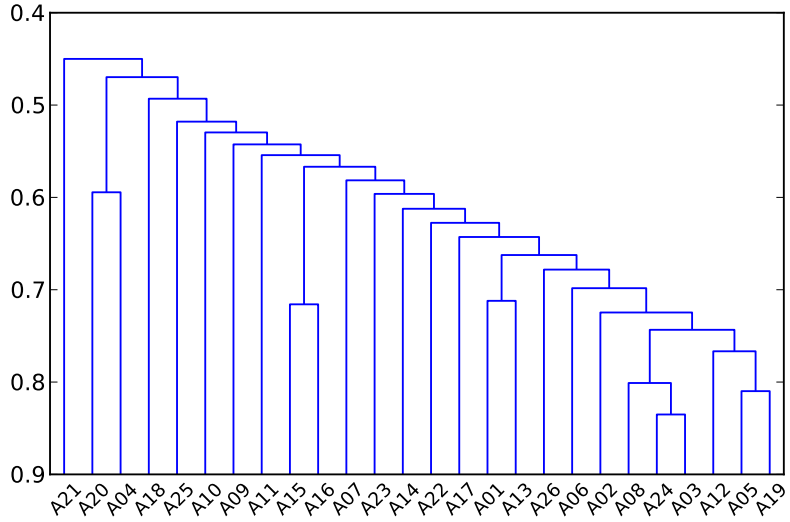
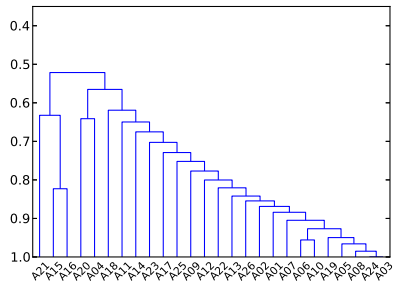



Figure 4.6: Clustering of the student annotators for the 'role+type' level.

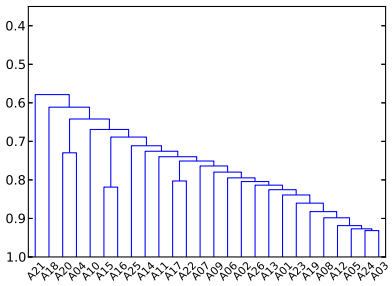
in most cases, so filtering out the severe cases might be a good idea. Most of the annotators showing deviations in category distribution could be identified, when annotators were sorted by deviation from average distribution $\Delta^{\varnothing}$, which is shown in the last column of Table 4.6. Filtering out the seven worst annotators in terms of $\Delta^{\varnothing}$, the resulting $\kappa$ increases from 0.45 to 0.54 on the 'role+type'-level, which is nearly equal to the 0.53 achieved when using the same size of annotator set in the central claim ordering. Although this ordering suffices to detect outliers in the set of annotators without relying on gold data, it still has two drawbacks: It only maximises to the average and will thus not guarantee best agreement scores for the smaller $n$-best sets. Furthermore, a more general critique on total orders of annotators: There are various ways in which a group agrees or disagrees simultaneously that might not be linearised this way.

In order to investigate the structure of agreement, we use **agglomerative hierarchical clustering** over the annotators. An overview of the clusterings for all possible levels of annotation is presented in Figure 4.7. Figure 4.6 depicts the clustering on the 'role+type'-level in more detail: The clustering grows steadily, maximally incorporating clusters of two annotators, so we do not see the threat of ambiguity in the guidelines. Furthermore, the clustering conforms with central claim ordering in picking out the same set of six reliable and good annotators (with an average F1 of 0.76 for 'role+type' and of 0.67 for the full task compared to gold), and it conforms with both orderings in picking out similar sets of worst annotators.

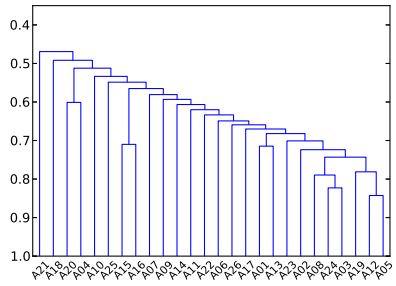
With this clustering we now have the possibility to investigate the agreement for subgroups of annotators. Since the growth of the clusters is rather linear, we choose to track



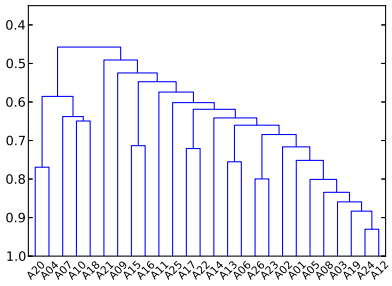

(a) role



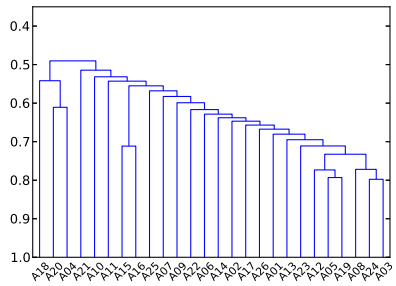

(b) typegen



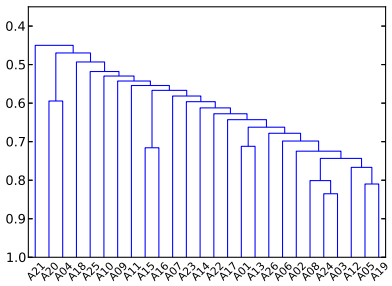

(c) type



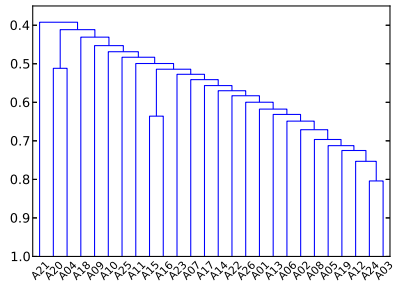

(d) comb



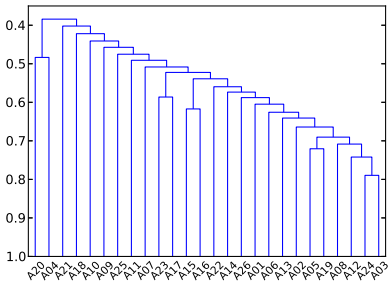

(e) target




(f) role + type




(g) role + type + comb




(h) role + type + comb + target

Figure 4.7: Clustering of the student annotators for all possible levels.

the confusion over the best path of growing clusters, i.e. starting from the best scoring {A24,A03} cluster to the maximal cluster. It would be interesting to see the change in Krippendorff's category distinction diagnostic for selected confusion pairs. However, this value not only depends on the amount of confusion but also on the frequency of these categories[3], which cannot be assumed to be identical for different sets of annotators. We thus investigate the confusion rate $\text{conf}_{c_1,c_2}$, i.e. the ratio of confusing assignments pairs $|c_1 \circ c_2|$ in the total set of agreeing and confusing assignments pairs for these two categories:

$$\text{conf}_{c_1,c_2} = \frac{|c_1 \circ c_2|}{|c_1 \circ c_1| + |c_1 \circ c_2| + |c_2 \circ c_2|}$$

Figure 4.8 shows the confusion rate for selected category pairs over the path from the best scoring to the maximal cluster. The confusion between rebutters and undercutters is already at a high level for the best six best annotators, but increases when worse annotators enter the cluster. A constant and relatively low confusion rate is observed for PSN+PAU, which means that distinguishing counter-attacks from new premises is equally 'hard' for all annotators. Distinguishing normal and example support as well as central claims and supporting segments is not a problem for the six best annotators. It becomes slightly more confusing for more annotators, yet ends at a relatively low level around 0.08 and 0.13 respectively. Confusing undercutters and support on the opponent's side is only a problem of the low-agreeing annotators, with a confusion rate at nearly 0 for the first 21 annotators on the cluster path. Finally, note that there is no confusion typical for the high-agreeing annotators only.

### 4.2.3 Conclusions

In our first annotation experiment, we asked minimally trained students, more experienced annotators, and very experienced expert annotators to annotate the structure of short argumentative texts according to the scheme presented in Chapter 3. The annotations of the three experts were very reliable, with a $\kappa$=0.83 for the full task, covering all aspects of the argumentation structure. This leads us to conclude, that this scheme is stable enough to be used to create a resource with similar short texts annotated with argumentation structure.

The training that annotators receive, however, has a strong impact on the reliability of annotation. The 26 annotators of the student group, only by reading the guidelines and without any other prior experience, reach only a $\kappa$=0.38 for the full task. Yet, we could identify a subgroup of annotators reaching a reliable level of agreement and good F1 scores in comparison with gold data by different ranking and clustering approaches and investigated which category confusions were characteristic for the different subgroups. The ex-

[3] 20% confusion of frequent categories have a larger impact on agreement than that of less frequent categories.

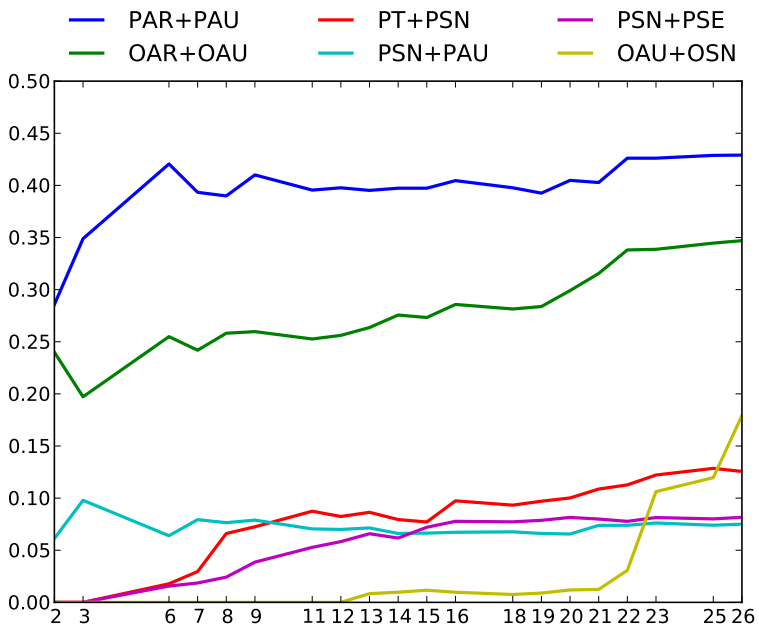



Figure 4.8: Confusion rate for selected category pairs in the growing clusters, with the numbers of annotators in the cluster on the x axis.

perienced group reach a better $\kappa$=0.42 for the full task, but are still quite far from expert performance.

The experiment presented here has several limitations: Although there are many annotators, the number of annotated items could be higher. 115 decisions are technically enough to come up with conclusive agreement scores in general, but given the skewedness of the labellings on certain levels, we should be careful in making conclusions about the reliability of annotation for less frequent categories. The second limitation of this experiment is the text genre. We chose those short texts advisedly as a starting point for an investigation on how we agree on argumentation structure, yet we are aware that these results cannot be readily generalised to other text genres such as e.g. newspaper pro and contra commentaries. We will address this question throughout this chapter to some extent, but have to leave a final answer to future work.

Nevertheless, we have learned a lot about the challenges of annotating argumentation structure even in these short texts. One of the main confusions in this experiment – the distinction between rebutting and undercutting attacks – will be the subject of our next annotation experiment.

## 4.3 Experiment 2: Classifying the type of argumentative attacks

One of the more difficult decisions in annotating argumentation structure is the distinction between rebutting and undercutting attacks. In our annotation studies we found a considerable confusion of these function types for student annotators, as well as for expert annotators.

In order to investigate this issue more deeply, we devised a follow-up experiment, where the annotators only classify argumentative attacks into either rebuttals or undercutters, a simple binary choice. For each decision, we presented to the annotators the whole text as the context and highlighted one segment as the attacker and another as being attacked. The annotators then had to decide whether the attacked segment was directly rebutted, or whether its argumentative function was undercut. To make this decision, the annotator had to understand the (supporting, attacking or thesis-stating) function of the target segment.

This task has both chances and challenges: On the one hand, annotators can concentrate on the desired task and are not distracted by other annotation decisions that would be required when annotating the full structure (choosing the central claim, distinguishing proponent and opponent, as well as support and attack). Also, to reduce the cognitive load of the annotators, we decided not to confront them with graphical representations of the remaining argumentation structure and instead simply show the text. On the other hand, we had to ensure a considerable level of text understanding, i.e. we had to provide enough information to convince our annotators that this item is indeed an instance of an attack and that the targeted claim is either rebutted or its argumentative function is undercut. We thus coloured segments as being the central claim, proponent, opponent or non-argumentative background segments, in order to provide at least a coarse-grained overview of the structure.

### 4.3.1 Experimental setup

In total, 12 **subjects** participated in the annotation experiment: One student annotator served in the pilot experiment (P). Nine student annotators (A1 - A9) of varying fields of study could be recruited over public announcements to participate in the experiments. They received experiment credit points obligatory in their studies or were remunerated for their effort. None of them participated in earlier experiments of this work. Finally, two expert annotators (E1, E2) completed the experiment, both experienced with discourse and argumentation annotation, one being the author of the guidelines. All subjects are native German speakers.

The annotation was done in a custom **web interface**, see Figure 4.9. Each text is presented in three views: The first gives the title of the text, respectively the trigger questions; and for the procon commentaries a (shortened) version of the background text is shown

to introduce the subjects to the discussed issue. The second view presents the segmented text without segments highlighted. Subjects are encouraged by the guidelines to fully read the text before advancing to the next view. The third view is the annotation view, which shows the segmented text. Each text segment is coloured (blue represents the central claim or restatements of it, green proponent, red opponent, and grey non-argumentative background segments). For each of the items to be decided, the attacking segment is marked by a fist symbol, the attacked segment by a cross-hair symbol. This is the pair for which the annotator has to decide whether it is a rebutting or an undercutting attack. If two segments form an ADU or if two ADUs are linked, more than one attacker or attackee symbols mark the corresponding segments. The annotator chooses the type of attack and continues to the next item of the text. Annotators can navigate through the items and revise earlier decisions. All actions of annotators in the annotation interface are recorded and time-stamped. Note that every annotator has a unique session ID, which technically allows her to log-out and continue her annotation after a later re-log-in. The text order was not randomised across annotators.

The **annotation guidelines** were four pages long. They first introduced the concepts of argumentative role (proponent vs opponent) and of argumentative function (support vs attack) and then elaborated on the distinction between rebutting and undercutting attacks. They provided examples for rebutting and undercutting in the opponent's voice, as well as for the proponent countering these attacks by rebutting or undercutting. The colouring of the segments was explained, as well as joint and linked segments. Finally, annotators were made aware of the need to resolve the proposition from rhetorical questions.

As **source material**, 23 German texts were selected, 17 short microtexts as well as six longer pro and contra commentaries from the ProCon section of the Potsdam Commenary Corpus [Stede, 2004, Stede and Neumann, 2014]. Additionally two microtexts and one ProCon text were selected as training material. From existing annotations of the argumentation structure of these texts, we extracted the segmentation, the segment types (central claim and its restatements, proponent and opponent and non-argumentative segments), and also the attacks, i.e. the items to be annotated. In total, the study comprises 59 attacks to be annotated. A detailed summary is given in Table 4.7.

The **procedure** of the experiment is as follows: First, annotators read the guidelines. Then, in a very short training phase they annotate two microtexts and one ProCon text with a total of eigth annotation items. They receive feedback from the experiment supervisor. Expert annotators skip the training phase. After that, individual participants annotate on their own.


Figure 4.9: Annotation environment in the rebut vs. undercut experiment.

### 4.3.2 Results

One aim of the pilot study was to measure the time required to finish the experiment. For the pilot study subject, it took 76min (14min for studying the guidelines, 13min for the training, and 49min for annotating the texts).

The minimally trained student annotators completed the experiment in two groups (A1, A4, A5, A8, and A2, A3, A6, A7, A9). Due to a network problem in the computer lab, the second group of annotators was interrupted after annotating for 20-30min and could not complete the experiment in one run. Thanks to the feature of the web-frontend to continue an annotation sessions, all participants finished the annotations remotely in the following days, but unfortunately not under controlled conditions.

| | attackee type | | | |
|---|---|---|---|---|
| | central claim | proponent | opponent | $\sum$ |
| microtext | 9 | 10 | 14 | 33 |
| ProCon | 8 | 4 | 14 | 26 |
| $\sum$ | 17 | 14 | 28 | 59 |

Table 4.7: Number of annotation items per text genre and type of the attacked segment.

| | students | | | experts | | |
|---|---|---|---|---|---|---|
| | $\kappa$ | AO | AE | $\kappa$ | AO | AE |
| microtext | .194 | .590 | .491 | .818 | .909 | .502 |
| ProCon | .171 | .594 | .511 | .842 | .923 | .512 |

Table 4.8: Agreement results per text genre in terms of Fleiss' $\kappa$.

The **agreement** of the students in terms of Fleiss' $\kappa$ is only marginal with $\kappa$=0.186 (AO=0.592; AE=0.498). Although the task was relatively hard and although the students received only minimal training, this result is below our expectations. The result for the first group of students is $\kappa$=0.194; the second group that experienced the interruption produced a result of $\kappa$=0.145. Whether this difference is to be attributed to the interruption remains unclear, as the small number of items and annotators does not allow testing with reasonable significance levels. Contrary to the students results, the agreement of the expert annotators is reliable at $\kappa$=0.830 (AO=0.915; AE=0.501).

A **clustering** of the student annotators is shown in Figure 4.10. There is a divide into two clusters of annotators, which does not coincide with the two groups of annotators. Such a divide into clusters might signal a systematic difference in the annotators' interpretation of the guidelines. However, the agreement inside of each cluster is still very low (with $\kappa$=0.224 in the left and $\kappa$=0.267 in the right cluster). We thus conclude that the low agreement of the student annotators is not due to a systematic ambiguity in the guidelines, but rather caused by other factors.

In order to investigate whether there are **genre specific** impacts on the annotation result, we calculated the agreement only on annotation items of the microtexts or of the ProCon texts (see Table 4.8 for the results). We observe no considerable difference in the result between genres for both student and expert annotators. The latter have a minimally higher agreement for ProCon texts as for microtexts, while the students' agreement is lower for ProCons.

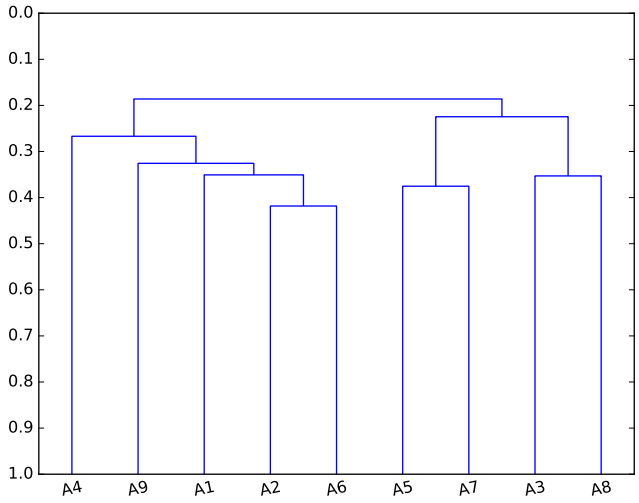



Figure 4.10: Clustering of the 9 student annotators IAA results in terms of Fleiss' $\kappa$.

In a similar vein, we can investigate the impact of the **attacked segment type**, i.e. we present separate results for annotation items where the attacked segments is the central claim, another proponent's segment, or an opponent's segment (this is only the case for counter-attacks). It is to be expected that the agreement for attacks on the central claim is very high: It can only be a rebutting attack, since the central claim has no argumentative function that targets any other claim, i.e. it does not constitute a relation that could be undercut. Note that this was not stated explicitly in the guidelines. The results are shown in Table 4.9.

Interestingly, there is only a small difference between the segment types for the student annotators, with proponent's segments being attacked showing the smallest agreement. The expected raise of agreement for central claims was not found. Looking at the decisions of the annotators it becomes evident that the rebutter and undercutter assignments are quite equally distributed over those annotation items. This allows the conclusion that the guidelines could not instruct the student subjects to apply the concept of undercutting relations appropriately for central claims, i.e. that central claims do not introduce a relation that could be undercut. One hypothesis is that annotators choosing an undercutter in this case consider it attacking a relation *from another* (non-highlighted) segment *to* the central claim, which would be in conflict with the guideline's instructions.

For the expert annotators, we can observe a more substantial difference across the segment types. The results for central claims being attacked is very good, although the $\kappa$ value is negative: In effect, the expert annotators disagreed only for one of the 17 central claim

| | students | | | experts | | |
|---|---|---|---|---|---|---|
| | $\kappa$ | AO | AE | $\kappa$ | AO | AE |
| central claim | .174 | .582 | .494 | -.030 | .941 | .943 |
| proponent | .144 | .591 | .522 | .836 | .929 | .564 |
| opponent | .201 | .598 | .497 | .727 | .893 | .608 |

Table 4.9: Agreement results per type of the attacked segment in terms of Fleiss' $\kappa$.

items, corresponding to an observed agreement of AO=0.941. Due to the nearly identical label distribution, the agreement expected by chance is likewise very high (AE=0.943), which in turn results in a negative $\kappa$ value representing an agreement below chance. When estimating the expected agreement not only on the central claim items but on all annotation items (AE=0.501), the $\kappa$ values are $\kappa$=0.882 for central claim, $\kappa$=0.858 for proponent and $\kappa$=0.786 for opponent. From this we conclude that expert annotators could better agree on potential objections presented by the author (where central claim or proponent segments are attacked) than on the author's counters of these attacks (where opponent segments are attacked). This is an interesting finding, since the Krippendorff diagnostics in the experiment on full structure annotation (see Tables 4.4 and 4.5) indicated that the confusions between opponent attacks on the proponent (OAR and OAU) have a stronger impact on the overall agreement than proponent counter-attacks on the opponent (PAR and PAU).

Finally, the annotators' results can be **compared to a gold standard**. After the annotation experiments, the experts compared their decisions and agreed on a gold standard. They disagreed on six items, most of which were edge cases. A comparison of all annotators with the gold standard, reporting both F1-scores as well as Fleiss' $\kappa$, is given in Table 4.10. It is not astonishing that the expert annotators achieve the highest agreement with the gold standard. From all other annotators, the student participating in the pilot study scores best with $\kappa$=0.649, much better than the best annotator from the regular student groups with $\kappa$=0.489. Interestingly, there is quite clear divide between five annotators with macro avg. F1-scores around 0.72, whose agreement with the gold standard can be considered moderate, and four annotators with scores around 0.50, the agreement of which is close to or below chance agreement. These two groups of better and worse performing annotators do not coincide with the annotator groups. Also, they do not align with the two identified clusters, which indicates that they are in general nearer to the gold standard, but not in a systematic way.

| | rebut F1 | undercut F1 | macro avg. F1 | $\kappa$ |
|---|---|---|---|---|
| E1 | .982 | .984 | .983 | .966 |
| E2 | .931 | .933 | .932 | .864 |
| P | .792 | .857 | .825 | .649 |
| A6 | .727 | .762 | .745 | .489 |
| A2 | .746 | .714 | .730 | .450 |
| A1 | .680 | .765 | .723 | .445 |
| A9 | .667 | .771 | .719 | .438 |
| A5 | .667 | .719 | .693 | .385 |
| A3 | .525 | .491 | .508 | .016 |
| A4 | .473 | .540 | .507 | .012 |
| A8 | .483 | .500 | .492 | -.017 |
| A7 | .385 | .515 | .449 | -.100 |

Table 4.10: Comparison of all annotators to gold standard.

### 4.3.3 Conclusions

Let us summarise the results of this annotation study: We asked annotators to classify argumentative attacks as either rebutters or undercutters, only given the text with central claims and argumentative roles highlighted, without a detailed diagram of the argumentation structure. The goal was to shed more light on the most frequent confusion in prior experiments. We found that expert annotators, who are familiar with the task and with correlated tasks of annotating the structure of argumentation could annotate this reliably. Student annotators with minimal training did not overall achieve reliable results, although the quality of their annotations in comparison to a gold standard is diverse, ranging from agreement below chance to a respectable agreement above chance in the mid-sixties.

This study can only be considered as a tentative first investigation of the complexity of argumentative attack type annotation. One obvious weakness of the experiment is its low power. The results might have been more conclusive with more items and annotators. Also, the control over the experiment was not optimal due to external technical problems.

A continuation of this research would certainly include a revision of the guidelines: Most students either did not understand that central claims do not on their own introduce a relation that could be undercut or they had an inverse relation in mind. Both should be clarified in the guidelines. In addition, a step by step decision protocol would help to give the annotator more confidence during the annotation process. Regarding the experimental setup, it would be wise to let annotators mark items for which they cannot understand the attack relation to be annotated. This would allow to single out misunderstandings of the argumentation of the text and focus on the evaluation on clear examples. Finally, in order

to investigate the questions whether a structural representation of the argumentation is required for making this distinction reliably or whether it is achievable from text without diagrams, a second condition is necessary, where annotators start with both the text and a (partial or full) argumentation graph to decide between a rebutting or undercutting attack.

## 4.4 Experiment 3: Argumentative zones in pro and contra commentaries

The argumentative nature of newspaper pro and contra commentaries, such as those in the ProCon section of the Potsdam Commentaries Corpus [Stede and Neumann, 2014], is to some extent different from what we have analysed so far in the argumentative microtexts (as we have already shown some in the experiment of the previous section). The inferences made by the author are not marked as explicitly, e.g. through discourse connectives. The persuasiveness is often carried by subtle signals, metaphors, and associations that the professional author advisedly weaves into the text, rather than direct and transparent reasoning. Even the most important information, the central claim, is not necessarily formulated explicitly. Since the critical issue is already provided in the headline, posed as a question, the stance towards it may be evident for the reader only by vocabulary choice, tone, or the reasons brought forward. Furthermore, these commentaries do not argue in every single sentence, they regularly have parts that are argumentatively not relevant, but serve other purposes such as introducing the issue, providing background information, setting the mood, or simply mentioning something as a side note. Finally, some of the arguments, even though explicitly signalled, may not be easy for the reader to integrate into the larger picture, because understanding them requires specific domain or background knowledge.

When we try to agree on the argumentative gist of these texts, it is useful to start with a shallower representation. In Section 2.2.1, we discussed the usefulness of argumentative zoning approaches, where the text is partitioned into functional zones. There have been elaborate proposals for an inventory of functional zones for scientific articles [see for instance Teufel, 1999], and also film reviews have been studied [Bieler et al., 2007], but we are not aware of any published proposal for an inventory of functional zones for argumentative commentaries. One first step towards this, especially for pro and contra commentaries, had been worked out in a student project [Bachmann and Brandt, 2005], the scheme of which we will build on in this section. What was missing in these experiments is a proof of the reliability of the scheme.

### 4.4.1 Scheme

Building on this preliminary study, we devised a scheme with the following zone categories (the corresponding segment colouring in the annotation interface is given in parentheses):

- **central claim**: The segment which best describes the author's position to the critical issue, without containing an other argument. Sometimes, the author does not formulate her claim explicitly in a separate clause, but rather relies on the reader to understand it from the trigger questions expressed in the headline and tenor of the text. In this case, the headline can be marked as the central claim, which otherwise remains unlabelled. (blue)
- **proponent**: All segments in the proponent's voice, directly or indirectly in favour of the central claim. This includes segments supporting the central claim or other proponent claims, refutations of possible objections, and implicit objections that are directly refuted in the very same segment (typically nominalisations, as in 'The idea of doing X is not helpful here, ...'). (green)
- **opponent**: All segments in the opponent's voice, directly or indirectly against the central claim. This includes possible or cited objections to the central claim or its premises, typically brought forward by the author to be refuted. (red)
- **background**: Some segments do not have an argumentative function but introduce the topic and the critical issue to the reader, state factual background information not pertaining to an argument, or simply represent digression from the topic. (grey)
- **upshot**: Often, the author ends her argument with a short summarising statement that is in some sense restating the central claim in a 'crisp' or metaphorical favourable way. If a central claim has already been marked in the text, this final restatement is marked as an upshot. (purple)
- **unlabelled**: Only the headline segment should remain unlabelled in texts with an explicit main claim, see above. (white)

Although this zoning scheme only produces a flat labelling of the EDUs, this representation already covers essential steps of the argumentative analysis: The central claim and restatements of it have been identified; arguments in favour of and against the central claim are marked with the corresponding argumentative role; and all segments that do not contribute to the whole argumentation because they are not relevant have been excluded. What is missing is the relational linking between the segments, which is left as a future step of analysis; but the building blocks for this more fine-grained analysis are set. In the following experiment, we will investigate how reliably annotators can agree on these basic labels given pro and contra commentaries.

### 4.4.2 Experimental setup

In the experiment there are two groups of **subjects**: One group of annotators are 50 *students*, who are obliged to participate in the experiment as an exercise in a (computational) linguistics course on text structure. In order to restrict the expenditure of time for each student annotator to a reasonable level for an exercise of about one hour of work, the students

are randomly assigned to ten groups of five annotators each, where each group marks a different set of texts containing two texts only marked by this group and one common text that is marked by all groups. The majority of the students, but not all of them, are native speakers of German. Since the course is taught in German, we can assume appropriate proficiency in German among the participants in the experiment. Concerning their expertise in the task, the students have learned about discourse structure in general in the course and especially devoted one session to text zoning approaches, which have been discussed in the context of scientific articles and film reviews. Finally, two *expert* annotators participated in the study, both experienced in discourse and argumentation annotation and being the authors of the guidelines. They annotated all texts of the experiment.

For the annotation, we use the **web interface** that has already been used in the previous study, see Section 4.3.1. Again, the text is shown in three views: one with the title and a short background text, the second with the segmented text to be read once before annotating, and then the annotation view. In contrast to the previous study, the annotation view presents the segmented text without any colouring or highlighting. When the annotator hovers over on of the text's segments, a classification menu with the possible categories appears right beside it. If one category is chosen, the text segment is coloured correspondingly. The annotator can freely choose which segments to annotate, in any order even across texts. Segments without a decision are later interpreted as bearing the category 'unlabelled'. An example annotation view with some segments already classified is shown in Figure 4.11.

The **annotation guidelines** used in the experiment are seven pages long; a slightly extended version is published in [Peldszus and Stede, 2016c]. Besides introducing and exemplifying the categories of the scheme, they discuss how to arrive at a propositional reading a segment in the case of fragmentary segments, rhetorical questions, and discourse anaphora. The annotators are encouraged to follow a step by step procedure, where they first identify the central claim and the upshot, then consider proponent and opponent arguments, and mark all remaining segments as background. Finally, an example ProCon text is analysed.

As **source material**, we chose 21 ProCon texts in this experiment. They were selected from the ProCon corpus according to two heuristics: 14 texts had already been used in a students' experiment on zoning in ProCon commentaries [Bachmann and Brandt, 2005]. The remaining seven texts were selected because they feature a potential high number of counter-arguments signalled by contrastive discourse markers which had been identified and disambiguated. The texts have been manually segmented into EDUs.

The **procedure** of the experiment is as follows: After receiving an introduction on argumentative zoning in scientific articles and film reviews in the course, a 15min presentation introduced the students to the idea of a similar zoning scheme for pro and contra commentaries, the scheme described above. The presentation explained all categories with an example text and made the students familiar with the web interface. The annotation itself was done as a homework, i.e. the students took the guidelines home and completed the ex-

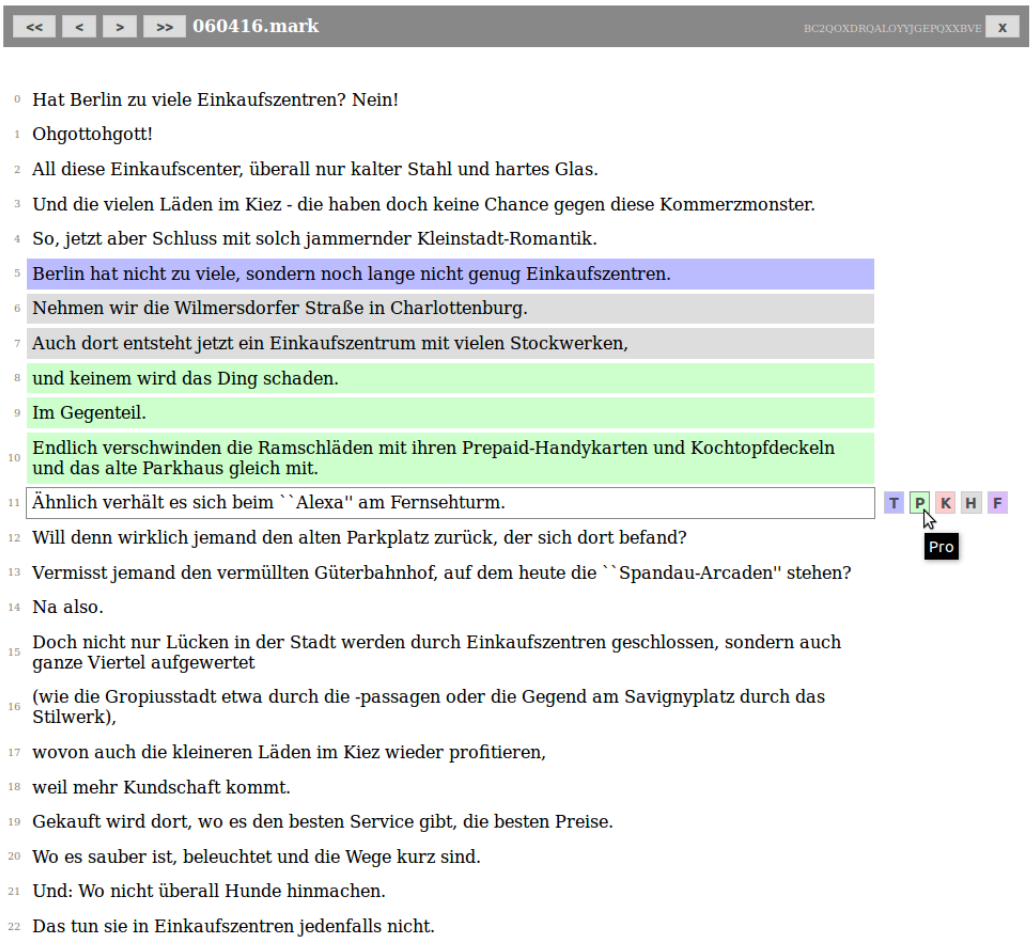



Figure 4.11: Annotation environment of the zoning experiment.

periment there. Consequently, we can neither measure how long it took the students to read the guidelines, nor control whether or to what extent they really studied the guidelines.

### 4.4.3 Results

Of the 50 students, five students cancelled their participation in the experiment, which is why group 3 and 7 only have four members and group 4 only has two members.

The **agreement** of the students and the experts is reported in Table 4.11. Each group marked two individual texts, and the common text that was annotated by all groups. The $\kappa$ values in the table are thus based on the annotation items of three texts. To allow a direct comparison with the expert annotators, the group-wise expert scores here are based on the very same items of the three texts a group annotated. The student in-group agreement

| | students | | | experts | | |
|---|---|---|---|---|---|---|
| group | $\kappa$ | AO | AE | $\kappa$ | AO | AE |
| 1 | .438 | .670 | .413 | .545 | .740 | .429 |
| 2 | .536 | .727 | .411 | .680 | .822 | .445 |
| 3 | .304 | .616 | .449 | .531 | .736 | .436 |
| 4 | .177 | .532 | .431 | .805 | .894 | .454 |
| 5 | .510 | .659 | .304 | .522 | .685 | .341 |
| 6 | .380 | .572 | .310 | .298 | .556 | .367 |
| 7 | .553 | .738 | .413 | .594 | .745 | .372 |
| 8 | .289 | .536 | .347 | .837 | .906 | .423 |
| 9 | .407 | .586 | .302 | .866 | .912 | .344 |
| 10 | .491 | .702 | .415 | .618 | .822 | .535 |

Table 4.11: Zoning results groups.

ranges from $\kappa$=0.177 up to 0.553, on average 0.409. The range for the expert agreement is from $\kappa$=0.298 to 0.866, with an average value of 0.630.

Since the common text has been annotated by every annotator, we can directly compare the student and the expert annotators for this specific instance. These results have to be read with caution, however, because the item size is very small with only 14 segments that are compared here. The students reach a considerable agreement of $\kappa$=0.519 (AO=0.711, AE=0.400), the experts an excellent agreement of $\kappa$=0.874 (AO=0.929; AE=0.431)

When assessing the expert agreement on all 21 texts, the corresponding value is $\kappa$=0.551 (AO=0.731; AE=0.400). The difference of this value to the average group-wise is due to the (above average) good expert agreement on the common text which positively influences every group-wise expert score. We can assume that the students' agreement over the whole corpus would be likewise smaller.

The large number of annotations for the common text also invites to apply the **clustering** of annotators: The dendrogram for all student annotators on the common text is shown in Figure 4.12. We observe three larger clusters, each of which has some annotators with a high cluster internal agreement. To understand the characteristics of these clusters, we investigate the differences in the labelling of the text, which is depicted right below the dendrogram for each of the annotators aligned with the ordering of the clustering. The common characteristic of all labellings in the leftmost cluster is that they identify the $7^{th}$ and / or $8^{th}$ segment as the central claim and leave the first segment the headline unlabelled. Nearly all other annotators label the headline as the central claim, instead. The difference between the middle and the larger cluster on the right is that the latter tends to classify the last segment as an upshot, while the middle one analyses it as a proponent's argument. The

| | central claim | proponent | opponent | background | upshot | unlabelled |
|---|---|---|---|---|---|---|
| central claim | **9** | 9 | | | 1 | 3 |
| proponent | 8 | **183** | 15 | 30 | 1 | 1 |
| opponent | | 2 | **42** | 4 | | |
| background | | 16 | 4 | **32** | 3 | |
| upshot | | | | | **1** | |
| unlabelled | 5 | | | | | **10** |

Table 4.12: Confusion matrix for the expert annotators.

presence of these three clusters might indicate a systematic ambiguity in the guidelines. Nevertheless, we cannot draw this conclusion here, since this result stems from only one single text and is rather to be attributed to the peculiarities of the common text than to general decision preferences of the annotators.

Let us now investigate the **confusions** between categories. We will focus on the expert annotators' result here. Table 4.12 shows the confusion matrix. The most frequent confusion is between background and proponent followed by central claim versus proponent and opponent versus proponent. While the first two confusions are to be expected, disagreements about the argumentative role are noteworthy. Some of them are due to different interpretations of contrastive markers, where one annotator identified a semantic, the other a pragmatic contrast (and only the latter would involve a role switch). Others are caused by perspective switches, where one annotator considers the new perspective to be that of the opponent, which is later rejected, while the other annotator takes it to be a deterrent example in the first place.

Another way to investigate confusions is Krippendorff's **category definition test**, the results of which are shown in Table 4.13: The best result with a gain of $\Delta\kappa$=0.181 is achieved by the opponent category, which has only few confusions with other categories. The unlabelled category is also very distinguished, as it is only used for headlines in text with explicit main claims. Future work on category definitions should focus on the background and the central claim categories. The strong drop for the upshot category is probably due to the low frequency of this category, causing a very high chance agreement against the rest.

While the confusion matrix could indicate to us which confusions were most frequent, Krippendorff's **category distinction test** can measure how much $\kappa$ we lost due to these confusions. For the result of this test, see Table 4.14. Most agreement is lost due to confusion between background and proponent, as also suggested by the confusion matrix. The second largest loss comes from confusions between central claim and the unlabelled headline, which was not expected from the frequency of this confusion. Finally it is worth to mention that eliminating the confusions between proponent and opponent segments do not lead to a higher agreement. To the contrary, the agreement is significantly decreased when both

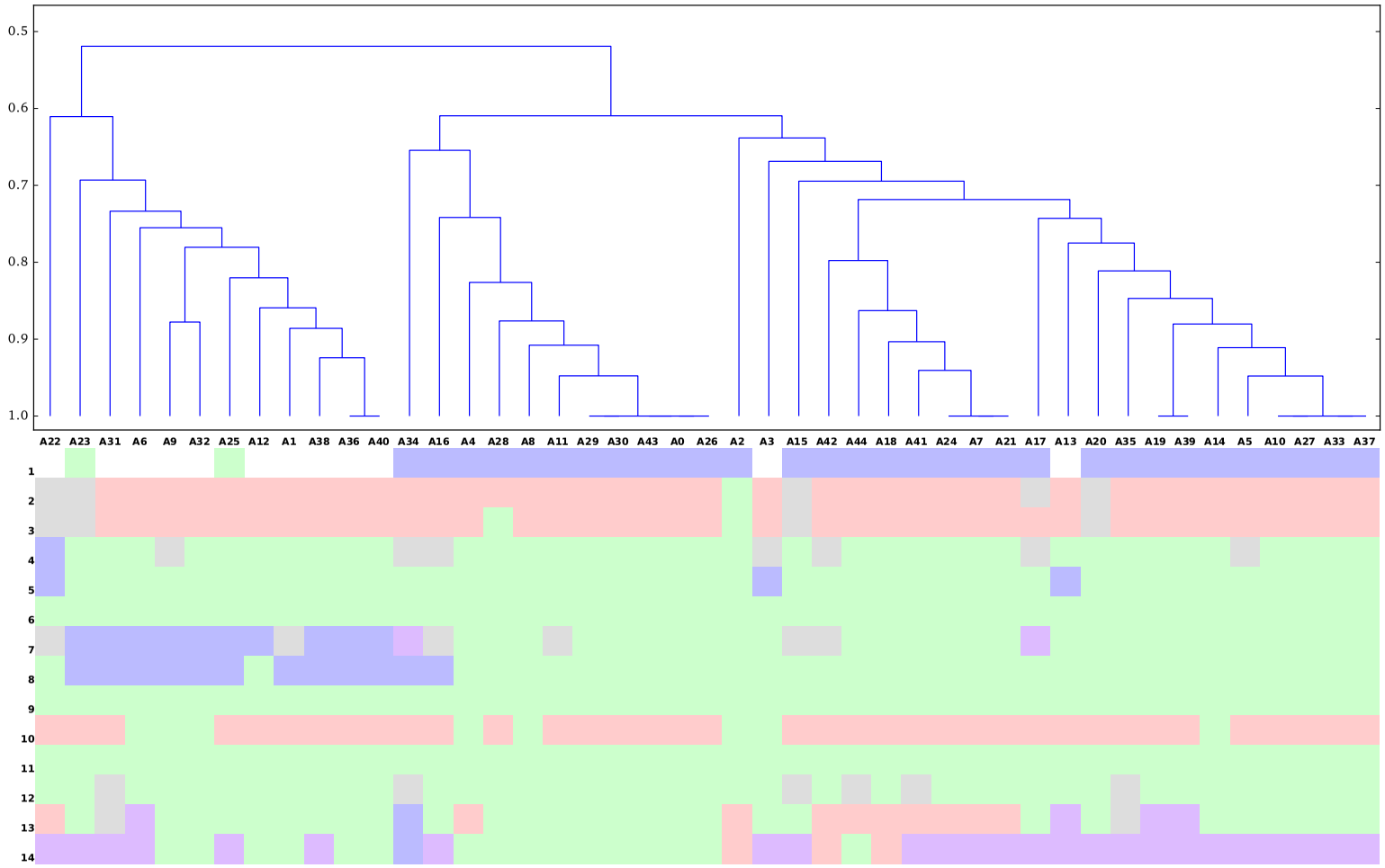



Figure 4.12: Clustering of the 45 student annotators IAA results for the zoning of the common text. The lower half shows the annotation of the annotators (central claims blue, proponent green, opponent red, background grey, upshot purple, and unlabelled white).

| category | $(\Delta)\kappa$ | AO | AE |
|---|---|---|---|
| *base* | .551 | .731 | .400 |
| opponent | +.181 | .934 | .754 |
| unlabelled | +.126 | .976 | .926 |
| proponent | +.001 | .784 | .517 |
| background | -.112 | .850 | .732 |
| central claim | -.179 | .931 | .891 |
| upshot | -.272 | .987 | .982 |

Table 4.13: Category definition test for the expert annotators.

| category pair | | $(\Delta)\kappa$ | AO | AE |
|---|---|---|---|---|
| *base* | | .551 | .731 | .400 |
| background | proponent | +.089 | .852 | .589 |
| central claim | unlabelled | +.076 | .778 | .406 |
| proponent | central claim | +.027 | .776 | .469 |
| proponent | unlabelled | +.015 | .760 | .447 |
| upshot | background | +.011 | .739 | .403 |
| upshot | central claim | +.004 | .734 | .401 |
| background | opponent | +.001 | .752 | .446 |
| upshot | proponent | -.004 | .734 | .411 |
| opponent | proponent | -.073 | .776 | .570 |

Table 4.14: Category distinction test for the expert annotators.

are collapsed. One possible explanation of this is the loss of distinction when collapsing a well defined category as opponent with a less defined category (see category definition results).

After the annotation experiments, the **gold standard** was created. The expert annotators compared their labellings of the texts. Several annotation disagreements could be fixed this way. Yet, it turned out that a considerable part of diverging labellings can be attributed to different but plausible readings of the argumentation in the text. We thus decided to create two gold standards, whenever the expert annotators could mutually agree that both readings are plausible interpretations of the text. From the 379 segments, 14 segments of the expert annotator 1 were revised to form gold standard 1, and 19 segments of expert annotator 2 to define gold standard 2. The gold standards agree on 299 of the segments. For only two of the 21 texts the gold standards are identical. They differ in one segment for

five texts, on two segments for three texts, on three segments for three texts, and on more segments in eight texts.

As an upper bound of the possible expert agreement, we calculate the $\kappa$-value between the two gold standards and observe an astonishingly low $\kappa$=0.646 (AO=0.789, AE=0.403). This means that two perfect expert annotators will at best achieve this agreement when they always pick a different but plausible reading of the text.

To get an understanding of the accuracy and bias of the students to these two gold standards, we calculate the agreement for each student to both gold standards. For each student, we only consider the texts that his group annotated. The results are listed in Table 4.15. We also collect the best of both agreement scores in a separate column, to allow a comparison of student annotators across their closeness to one or the other gold standard. The range of the maximum $\kappa$ values is between 0.229 and 0.743, with an average of 0.497. Two thirds of the student annotators rather tend towards gold standard 2, while one third tends towards gold standard 1. Note, though, that it would be more appropriate to tackle this evaluation in a multi-label setting, where the agreement with each item is tested individually against all possible labellings [see Ravenscroft et al., 2016, as an example]. If the difference between both gold standards is indeed based on multiple interpretations of the text and not on annotator preferences, then these interpretations could be distributed differently over the gold standards than here. Introducing the corresponding evaluation metrics and their methodology is, however, out of the scope of this section and we leave it for future research.

### 4.4.4 Conclusions

To summarise, we presented a scheme with five different functional zones for argumentative commentaries, such as pro and contra commentaries. The resulting representations of argumentation structure are shallow, but already cover important steps of argumentative analysis such as central claim identification, argumentative role classification and the determination of argumentative relevance. The annotation experiment showed that minimally trained students can reach a moderate agreement. Experts can achieve more reliable results. One special challenge is that even for this coarse-grained structure of argumentation, expert annotators can mutually agree on multiple readings of the text, which led us to define two gold standard annotations and allowed us to assess an upper bound for the expert agreement.

It remains to be investigated to what extent the agreement on ambiguous text could be enforced by defining preference rules that the annotators have to apply when facing multiple possible labellings. Furthermore, we consider it worthwhile to deepen our understanding of whether these ambiguities are restricted locally or whether they are global alternatives that depend on the whole reading of the text (e.g., is this conflict between background and proponent only depending on what is expressed in the segment, or does it depend on

| annotator | group | $\kappa$ gold 1 | $\kappa$ gold 2 | max | $\Delta$ |
|---|---|---|---|---|---|
| A0 | 1 | .295 | .598 | .598 | -.303 |
| A1 | 1 | .676 | .648 | .676 | +.027 |
| A2 | 1 | .266 | .440 | .440 | -.174 |
| A3 | 1 | .045 | .339 | .339 | -.294 |
| A4 | 1 | .198 | .524 | .524 | -.326 |
| A5 | 2 | .365 | .607 | .607 | -.242 |
| A6 | 2 | .534 | .407 | .534 | +.126 |
| A7 | 2 | .355 | .415 | .415 | -.061 |
| A8 | 2 | .526 | .563 | .563 | -.037 |
| A9 | 2 | .528 | .518 | .528 | +.010 |
| A10 | 3 | .374 | .548 | .548 | -.175 |
| A11 | 3 | .393 | .685 | .685 | -.293 |
| A12 | 3 | .349 | .362 | .362 | -.013 |
| A13 | 3 | .246 | .355 | .355 | -.108 |
| A14 | 4 | .399 | .462 | .462 | -.063 |
| A15 | 4 | .393 | .345 | .393 | +.048 |
| A16 | 5 | .583 | .663 | .663 | -.081 |
| A17 | 5 | .591 | .491 | .591 | +.099 |
| A18 | 5 | .556 | .472 | .556 | +.084 |
| A19 | 5 | .492 | .496 | .496 | -.004 |
| A20 | 5 | .456 | .565 | .565 | -.110 |
| A21 | 6 | .065 | .438 | .438 | -.373 |
| A22 | 6 | .141 | .229 | .229 | -.088 |
| A23 | 6 | .162 | .501 | .501 | -.340 |
| A24 | 6 | .093 | .505 | .505 | -.412 |
| A25 | 6 | .546 | .422 | .546 | +.124 |
| A26 | 7 | .450 | .472 | .472 | -.023 |
| A27 | 7 | .437 | .358 | .437 | +.078 |
| A28 | 7 | .677 | .447 | .677 | +.230 |
| A29 | 7 | .553 | .612 | .612 | -.059 |
| A30 | 8 | .547 | .564 | .564 | -.016 |
| A31 | 8 | .454 | .469 | .469 | -.015 |
| A32 | 8 | .224 | .183 | .224 | +.041 |
| A33 | 8 | .352 | .373 | .373 | -.021 |
| A34 | 8 | .355 | .367 | .367 | -.012 |
| A35 | 9 | .409 | .393 | .409 | +.016 |
| A36 | 9 | .743 | .672 | .743 | +.071 |
| A37 | 9 | .362 | .300 | .362 | +.062 |
| A38 | 9 | .483 | .466 | .483 | +.017 |
| A39 | 9 | .457 | .390 | .457 | +.067 |
| A40 | 10 | .678 | .687 | .687 | -.009 |
| A41 | 10 | .432 | .530 | .530 | -.098 |
| A42 | 10 | .199 | .307 | .307 | -.108 |
| A43 | 10 | .530 | .634 | .634 | -.105 |
| A44 | 10 | .274 | .455 | .455 | -.181 |
| *mean* | | .405 | .473 | .497 | -.068 |

Table 4.15: Agreement of the student annotators with the two gold standards measured in Fleiss' $\kappa$; the (max) column shows the best of both agreement values, the ($\Delta$) column the difference ($\kappa$ gold 1 - $\kappa$ gold 2). Positive values indicate a higher agreement with gold standard 1, negative values a higher agreement with gold standard 2.

whether another segment is considered to be an argument or not?). When investigating this further, the evaluation should act on the assumption of a multi-label setting.

One limitation of the study is the reduced power in terms of annotation items per student annotator – constraints which are evidently due to the nature of being a course exercise. Still, it allowed us to demonstrate the usefulness of the clustering techniques as a diagnostic tool for inspecting annotator characteristics. For future work, it will be interesting to apply the scheme to commentary text that is not strictly a pro and contra commentary, where central claims may be expressed differently and the proponent and opponent distinction is less evident.

## 4.5 Conclusions

In this chapter, we have presented three different annotation experiments, each investigating the question how we agree on argumentation structure from a different angle:

- The experiment in Section 4.2 evaluated the reliability of annotation of full argumentation structures on short texts. We could show that the scheme presented in Chapter 3 is stable and allows expert annotators to produce very reliable annotations with a $\kappa$=0.83. A group of minimally trained students reach only moderate agreement with $\kappa$=0.38, but we could identify subgroups of annotators by ranking and clustering analysis that obtain good agreement. A group of more experienced student annotators achieved a considerable result with $\kappa$=0.42.

- One of the more frequent confusions, the distinction between rebutting and undercutting attacks, was investigated in a follow up experiment, which we presented in Section 4.3. Here we used not only the microtexts but also the more complicated ProCon texts as source material. The experimental setup aimed at confronting the subjects with as little structural representation of argumentation as possible and forced them to decide mostly on the basis of pure text. Expert annotators performed very well here, with (again) $\kappa$=0.83, but student annotators did not even reach a fair level of agreement. Our conclusion is that the good performance of the experts is most likely due to their experience in the task of annotating argumentation structure in the first place, while the guidelines require a revision in order to be instructive also for the inexperienced annotators.

- The last experiment, presented in Section 4.4, then tested how reliably a coarse-grained argumentative zoning of the ProCon commentaries could be annotated. An existing scheme was refined, featuring six argumentative zones, which have been used to mark 21 commentaries. The student agreement was moderate. The experts yielded better results but not fully reliable agreement, with $\kappa$=0.55. While this figure

might be considered as unsatisfactory for expert annotators, we found that many of those disagreements were due to different interpretations of the argumentative structure of the text, which the experts could mutually agree on. As a result, we have a multi-labelling or two gold standards, the upper limit of agreement between both being $\kappa$=0.65.

In all our experiments we experienced the strong impact of training and commitment to the task on the result of non-expert annotators, which is not only relevant for experiments with undergraduate students but also with crowd-sourced annotation. This especially applies to a task such as argumentative analysis, where decisions are based on interpretation of complex meaning that are not easily made explicit and require thorough and often time-consuming consideration.

The agreement score our expert annotators reached for annotating the argumentation structure in microtexts ($\kappa$=0.83) compares favourably to related work. Stab and Gurevych [2014b] obtained $\alpha_U$=0.72 for argument components at the sentence level and $\alpha\approx$0.81 for argumentative relations in student essays. Stab and Gurevych [2016] reported a higher $\alpha_U$=0.77 for argument components and a lower $\alpha\approx$0.73 for argumentative relations in a follow up experiment. Kirschner et al. [2015] annotated two argumentative and two organisational relations in scientific text and reported $\kappa$=0.43.

After having shown that our scheme for annotating argumentation structure is stable and reliable for short argumentative texts, we started creating and annotating a corpus of such texts. On this we will report in the next chapter.

# 5 A corpus of texts annotated with argumentation structure

This chapter describes the creation, translation, annotation, extension, and delivery of a new and freely available resource – the corpus of argumentative 'microtexts'. It features short and dense authentic arguments, annotated according to our scheme for representing text-level argumentation structure. The corpus consists of 112 German texts plus professional English translations that preserve linearisation and argumentative structure. We hope this resource will foster research in the study of argumentation, of the relation of discourse and argumentation and of the automatic recognition of argumentation structure.

First, we will review related work and already available resources in Section 5.1. We then report in Section 5.2 on the creation, translation, and annotation of the microtext corpus and provide detailed statistics of the variety and the linguistic realisation of argumentation structure in the corpus. Section 5.3 presents two useful transformations of the annotated argumentation structures that will be used in our experiments of automizing the recognition of argumentative structure. An extension of the corpus by two additional annotation layers, one for RST and one for SDRT, is presented in Section 5.4. Finally, Section 5.5 concludes this chapter with a short summary and an overview of the corpus versions.

**Previously published material**

Section 5.2 is a slightly extended version of [Peldszus and Stede, 2016a]. The transformations presented in Section 5.3 have been first described, though only very briefly in [Peldszus and Stede, 2015a]. Section 5.4, finally, which describes the extension of the corpus, has been published before as [Stede et al., 2016]. It is joint work with our colleagues from Toulouse. My main responsibilities in this project, besides contributing to the final reconciliation of discourse segmentation and annotation, were (i) to apply the finer segmentation to the argumentation annotations, (ii) to implement and evaluate the alignment and correlation mechanisms, and (iii) to organise and maintain the distribution of the data. Note that the empirical study on aligning argumentation structure with RST or SDRT is not reproduced in this chapter, as it is only of subordinate interest for the purpose of this work. We refer the interested reader to the corresponding publication.

## 5.1 Related work

We focus here on resources of argumentation that mark up the argumentation structure in a piece of monologue text. We do not consider corpora of transcripts of dialogue or multi-party debates, mediation etc. Also not covered here are corpora that mark argumentative relations not in but *between* texts, for example between the entries of linear of threaded forums, below blog or article comments, product reviews, or Wikipedia articles / discussions. This also includes corpora that have been semi-automatically created from structured debate portals. Also, we do not consider corpora which contain only single argumentative relations without their context a larger text, or classify argumentative units without relating them.

It should be noted that many of the corpora that do not fit our criteria here are indeed useful for extending an existing model for a specific sub-task of argumentation mining, e.g. for identifying argumentative relevant units, or for classifying the type of ADUs, or even for learning certain argumentative relations. However, for a model of the argumentation structure that combines these tasks for structure prediction, the basis should be a corpus that is fully annotated with structures, so that at least some documents it is trained on are reasonably similar to the documents it is later applied to.

The first larger resource for argumentation was the AraucariaDB [Reed et al., 2008]. It contains roughly 650 documents of different types collected from several international newspapers, parliamentary records, legal judgements, weekly magazines, discussion forums etc., but also examples from classic argumentation literature. The documents in the corpus are not full texts but rather excerpts of the originally considered source that are argumentatively interesting. The length of these excerpts ranges from single sentences expressing one argumentative relation between a premise and a conclusion, to longer texts with more than 30 sentences and a deeply nested argumentation structure. The distribution of Araucaria that is currently available and downloadable [AIFdb, 2016] from the AIFdb [Lawrence et al., 2012] contains 661 documents (69 of which are text duplicates with different argumentative analyses). In total there are 3,991 text nodes, 1,745 supporting and 38 attacking argumentative relations (represented as RA and CA nodes). AraucariaDB also features reconstructed enthymemes. Unfortunately, they are not typed differently from the text nodes; matching the text spans of nodes against the source text indicates that about 1,600 text nodes are reconstructed premises. Also, about 40% of the relations are annotated with argumentation scheme types following different scheme-sets, e.g. that of Walton et al. [2008]. However, besides the detailed and analysis-intense annotations in this corpus and its important role for encouraging research in the field, it is not perfectly suited for the purposes of studying and automatically recognising the *structure* of argumentation. One reason is that the majority of the annotated argumentations are structurally not very complex: 270 of the 661 structures only have one argumentative relation, 485 have no more

than three relations, so only a quarter of the documents (but still more than 170) might be of interest. Secondly, the documents are excerpts of a larger text and available only in isolation. The context, such as immediately preceding or following text, could be important in the process of automatising structure recognition, as it might bear signals that set the stage for the reader to understand what the argument is, where it starts, and ends. The AraucariaDB has been prominently used in the experiments of Mochales and Moens [2011] for argument relevance determination and premise v. conclusion classification.

Another corpus that was used by Mochales and Moens [2011] is a collection of 47 judgments and decisions of the European Court of Human Rights (ECHR) in English [see also Mochales Palau and Ieven, 2009]. It includes 2,571 sentences of which 1,449 are non-argumentative. 764 premises and 304 conclusions as well as their relations are marked up in a tree-structure. Unfortunately, the annotated corpus is not freely available.

A small corpus of ten German Pro & Contra commentaries from the PCC has been annotated with argumentation structures according a scheme similar to ours, and was studied by Stede and Sauermann [2008]. The corpus is too small to be used on its own for machine learning. It could constitute an additional resource when compiling a corpus for a similar text genre, though. The annotations are not officially distributed with the PCC release, but may be available on request.

When we began compiling and annotating the microtext corpus in January 2014, these were the only available resources. Since then, the interest in argumentation mining greatly increased in general, and with it the awareness for the lack of available data. This brought forward other new corpora.

Stab and Gurevych [2014b] presented an annotated corpus of 90 student essays in English. The corpus comprises 1,673 sentences, with segments of clause or sentence size. There are 90 central claims annotated, 429 intermediary claims central to one paragraph, and 1,033 premises. Claims are marked with an attribute 'for' (365) or 'against' (64), but the authors do not report numbers on the stance of premises. Note, however, that the stance of premises could be inferred by the relation structure, i.e. the sequence of supposing and opposing relations. Of the 1,473 relations in the corpus, 161 are opposing.

Recently, a new version of the corpus has been released [Stab and Gurevych, 2016], which is significantly larger with 402 annotated student essays (covering 7,116 sentences of which 23% are non-argumentative). In contrast to the first release, restated central claims are now also annotated, amounting to 751 major claims. In total, 1,506 intermediary claims and 3,832 premises are marked up. Only 12% of the arguments have an attack relation. The corpus is currently the largest resource of argumentative *structures* available.

Kirschner et al. [2015] presented an annotation study of argumentation structure in scientific articles. For this, the introduction and the discussion sections of 24 German articles from the educational domain were annotated using a claim-premise scheme with supporting and attacking relations, as well as (RST-inspired) sequence and detail relations. The

basic unit of annotation were sentences, 2,742 have been marked. Until now, no finalised corpus has been compiled and made publicly available.

Finally, we want to mention one resource that may be useful for our purposes, even though structure is not explicitly coded as relations between units: Habernal and Gurevych [2017] created a corpus of 340 documents from user comments, forum posts, blog posts, and newswire articles and annotated it according to their modified Toulmin [1958] scheme (without the roles of warrant and qualifier, but including a role 'refutation' that marks a counter to the rebuttal). The corpus contains 3,899 sentences to which the labels of the scheme are assigned, where 2,214 sentences are non-argumentative. Although the nesting of these Toulmin-like structures is not allowed (which would violate our compositionality requirement), there are still relations between claim, premise, backing, rebuttal, and refutation that could be mapped to an argumentation structure following our scheme. However, the scheme also allows for implicit claims, and indeed 48% of the claims were implicit. Only those instances with an explicit central claim could be mapped effectively.

## 5.2 The microtext corpus

Argumentation can, for theoretical purposes, be studied on the basis of carefully constructed examples that illustrate specific phenomena, but for many researchers the link to authentic, human-authored text is highly desirable. Especially, since data-driven methods are increasingly applied to the problem of argumentation mining, the interest in argumentation-oriented corpora of monologue text as well as spoken dialog is increasing. In the work reported here, we address this need by making a resource publicly available that is designed to fill a particular gap: Short texts with explicit argumentation, little argumentatively irrelevant material, fewer rhetorical gimmicks (or even deception), in clean written language.

Let us set out why we think there is a need for this type of resource. Evidently, authentic texts from social media or newspapers are ultimately a target for automatic argumentation mining. These sources are, however, often not ideal for more qualitatively oriented research. In newswire text, the language can be quite complex. In our analysis of the Pro & Contra commentaries in the PCC (see e.g. the zoning experiments in Chapter 4.4), we found that although very argumentative, these texts often come with persuasive devices that are directing the focus of the reader rather than making an explicit argument. Another challenge are implicit main claims, which are understood by the reader but are required to be made explicit by the analyst. Social media and user-generated content 'in the wild' in general, adds its own challenges to the already demanding task of (automatic) argumentative analysis: The language is (not always but often) ill-formed, grammatically or orthographically flawed; the texts tend to be spontaneously produced rather than planned. Essays of language learners, finally, are the desired text source for developing applications

for argumentative writing support or essay scoring, but are not in general the perfect genre for studying the automatic prediction and the language of argumentation structure, as precisely the use of linguistic signals of argumentation, such as discourse connectives, of the respective writers is not yet proficient [see e.g. Stab and Gurevych, 2014a, p. 54]. All these factors have an impact on the underlying argumentation structure; in some cases it is trivial, and in other cases quite nontransparent.

Our contribution is a collection of 112 'microtexts' that have been written in response to trigger questions, mostly in the form of 'Should one do X'. The texts are short but at the same time 'complete' in that they provide a standpoint and a justification, by necessity in a fairly dense form. Hence, the underlying argumentation structure is relatively clear. We collected the texts in German and then had them translated to English; both versions are available to interested researchers.

In addition to the raw texts, we provide manually-created annotations of the argumentation structure, following the scheme presented in Chapter 4. Thus, argumentation researchers will find a resource of simple, authentic natural language texts together with suggestions of structural representations of the underlying argument. At the same time, the data can also be used for building models in automatic argumentation mining.

### 5.2.1 Data collection and cleaning

#### Collection

The microtext corpus consists of two parts. On the one hand, 23 texts were written by the author as a 'proof of concept' for the idea. These text were used in the annotation experiments in Chapter 4. They also were used as examples in teaching and testing argumentation analysis with students. An example text was shown in Figure 4.2.

On the other hand, 89 texts were collected in a controlled **text generation experiment**, where normal competent language users wrote short texts of controlled linguistic and rhetoric complexity.

To this end, 23 subjects were instructed to write a text on a topic that was to be chosen from a given set of trigger questions. All subjects were native speakers of German, of varying age, education, and profession. They received a short written instruction (about one page long) with a description of the task and three sample texts. The subjects were asked to first gather a list with the pros and cons of the trigger question, then take stance for one side and argue for it on the basis of their reflection in a short argumentative text. Each text was to fulfil three requirements: It was to be about five segments long; and all segments were to be argumentatively relevant, either formulating the main claim of the text, supporting the main claim or another segment, or attacking the main claim or another segment. Also, the subjects were asked to consider at least one possible objection to the claim in the text. Finally, the text was to be written in such a way that it would be comprehensible without

having its trigger question as a headline. The subjects were asked to write five texts, each about a different topic. They were allowed to select the topic according to their interest in order to maximise their engagement for the task.

### Cleaning

Since we aim for a corpus of texts featuring authentic argumentation but also regular language, all texts were corrected for spelling and grammar errors. Subsequently, the texts were segmented into elementary units of argumentation (ADUs). Most subjects already marked up in some way what they regarded as a segment. Their segmentation was corrected when necessary, e.g. when only complex noun phrase conjuncts or restrictive relative clauses had been marked, or when subordinate clauses had not been split off. All remaining texts were segmented from scratch. Due to this step of (re-)segmentation, not all of the final texts conform to the length restriction of five segments; they can be one segment longer or shorter.

Unfortunately, some subjects wrote relatively long texts. We decided to shorten these texts if possible by removing segments that appeared less relevant. This removal also required some modifications in the remaining segments to maintain text coherence, which we kept as minimal as possible.

Another source of problems were segments that did not meet our requirement of argumentative relevance. When writers did not concentrate on discussing the thesis but moved on to a different issue, we removed those segments, again with minimal changes in the remaining segments. Some texts containing several of such segments remained too short after the removal and thus were discarded from the dataset. After the cleanup steps, 89 of the original 100 texts remained for annotation of argumentation structure.

### Translation

To supplement the original German version of the collected texts, the whole corpus (including the constructed texts, as well as those from the text generation experiment) were professionally translated to English, in order to reach a wider audience of potential users. Our aim was to have a parallel corpus, where annotated argumentation structures could represent both the German and English version of a text. We thus constrained the translation to preserve the segmentation of the text on the one hand (effectively ruling out phrasal translations of clause-type segments), and to preserve its linearisation on the other hand (disallowing changes to the order of appearance of arguments). Beyond these constraints, the translation was free in any other respect. Note that the translator had only access to the segmented source text, but not to an argumentative analysis of the text.

| text length | number of texts |
|---|---|
| 3 | 3 |
| 4 | 11 |
| 5 | 71 |
| 6 | 26 |
| 7 | 2 |
| 8 | 0 |
| 9 | 0 |
| 10 | 1 |

Table 5.1: Length of the texts in segments

| position | number of central claims |
|---|---|
| 1/5 | 48 |
| 2/5 | 18 |
| 3/5 | 16 |
| 4/5 | 3 |
| 5/5 | 27 |

Table 5.2: Position of the central claim

### 5.2.2 Annotation process

All texts of the corpus were then marked up with argumentation structures by one expert annotator, according to the scheme presented in Chapter 3. All annotations were checked, with controversial instances discussed in a reconciliation phase by two or more expert annotators. The annotation of the corpus was originally done manually on paper. In follow-up annotations, we used GraPAT [Sonntag and Stede, 2014], a web-based annotation tool specifically dedicated to constructing graph structures.

Since the professional translation preserves linearisation and argumentation structures, all annotated graphs represent both the German original and the English translation of the argument.

The bilingual texts and the annotations are publicly available in a suitable XML format; see Section 5.2.4.

### 5.2.3 Corpus statistics

The corpus features a wide range of different argumentation patterns. In the following, we will present detailed statistics on these, including distribution of roles and argumentative moves, positioning of the central claim in the text, as well as forward (from premise to conclusion) and backward linearisations of arguments.

#### General statistics

In the corpus there are 112 texts, with 576 ADU segments in total. Table 5.1 shows the length of texts in the corpus measured in segments: The great majority of texts are four, five, or six segments long (the average being 5.1), with only a few exceptions.

### Topics and stances

The distribution of chosen topics and stances towards it is given in Table 5.3. It shows the variety of topics covered in the corpus, both on the dimension of a topic's attractiveness to the writers (measured in the amount of texts produced for a topic), as well as on the dimension of bias in stances taken (with very clear pro and con cases, but also more controversial topics with equal stance distribution).

### Central claim

In the English-speaking school of essay writing and debating, there is a tendency to state the central claim of a text or a paragraph in the very first sentence, followed by supporting arguments. To some extent, we can expect to find this pattern also in other languages. To investigate whether the tendency also holds in our corpus, we divide each text into five equal parts and count the occurrence of the central claim in this position. As Table 5.2 shows, the dominant position is indeed the beginning of a text, directly followed by the end of the text. Note, however, that the overall majority of central claims (57%) is at positions other than the beginning.

### Argumentative role

Of the 576 segments, 451 are proponent and 125 are opponent ones. While there are 15 texts where no opponent segment has been marked (either because the author did not conform to the requirement to consider at least one objection or because she phrased it indirectly in a non-clausal construction), the majority of texts (74) have exactly one opponent segment. Two opponent segments can be found in 18 texts, and three of them in five of the texts. Furthermore, Table 5.4 shows the position of opponent segments: It turns out that the dominant place to mention a potential objection is right before the end of the text, thus giving the author the possibility to conclude her text with a counter of the potential objection.

### Argumentative function

The frequency of argumentative functions annotated in our corpus is shown in Table 5.5: Most segments are normal support moves. Examples are used only rarely. About a third of the segments have an attacking function (either the opponent challenging the central claim or the proponent countering these objections), with overall more rebutters than undercutters.

It is noteworthy that rebutters and undercutters are not equally distributed over both argumentative roles. This is shown in Figure 5.1: The opponent typically rebuts, and the

| trigger question | pro | con | total |
|---|---|---|---|
| Should the fine for leaving dog excrements on sideways be increased? | 8 | 1 | 9 |
| Should Germany introduce the death penalty? | 1 | 7 | 8 |
| Should public health insurance cover treatments in complementary and alternative medicine? | 5 | 3 | 8 |
| Should shopping malls generally be allowed to open on holidays and Sundays? | 4 | 4 | 8 |
| Should only those viewers pay a TV licence fee who actually want to watch programs offered by public broadcasters? | 4 | 3 | 7 |
| Should the statutory retirement age remain at 63 years in the future? | 0 | 6 | 6 |
| Should all universities in Germany charge tuition fees? | 1 | 5 | 6 |
| Should there be a cap on rent increases for a change of tenant? | 6 | 1 | 6 |
| Should the the morning-after pill be sold over the counter at the pharmacy? | 2 | 4 | 6 |
| Should intelligence services be regulated more tightly by parliament? | 4 | 0 | 4 |
| Should the weight of the BA thesis in the final grade be increased? | 2 | 2 | 4 |
| Should school uniforms be worn again in our schools? | 2 | 1 | 3 |
| Should video games be made olympic? | 2 | 1 | 3 |
| Should the EU exert influence on the political events in Ukraine? | 2 | 1 | 3 |
| Should the Berlin Tegel airport remain operational after the opening of the Berlin Brandenburg airport? | 1 | 1 | † 3 |
| Should Germany buy CDs with tax evader data from dubious sources? | 0 | 2 | 2 |
| Should parts of the Tempelhofer Feld be made available for residential construction? | 2 | 0 | 2 |
| Should we continue to separate our waste for recycling? | 1 | 0 | 1 |

Table 5.3: Trigger questions (translated) and stances taken by the authors for those texts of the corpus that were written in the text generation experiment. († One of the three texts on the Berlin Tegel airport has an unclear / undecided stance.)

| position | number of objections |
|---|---|
| 1/5 | 20 |
| 2/5 | 29 |
| 3/5 | 22 |
| 4/5 | 36 |
| 5/5 | 18 |

Table 5.4: Position of opponent segments (objections)

| type | number | sub-type | number |
|---|---|---|---|
| support | 272 | normal | 263 |
| | | example | 9 |
| attack | 171 | rebut | 108 |
| | | undercut | 63 |
| central claim | 112 | | |

Table 5.5: Frequency of argumentative function

great majority of these rebuttals is directed against the central claim, while only a few work against supporting arguments. In contrast to that, the proponent usually undercuts. We attribute this to the common strategy of the authors to first concede a possible objection, thereby demonstrating that their presentation is not fully biased, and then render the objection irrelevant.

Also note that a possible objection (an attack of the opponent) does not necessarily need to be counter-attacked by the proponent: The total number of attacks by the proponent is significantly smaller than the total number of attacks by the opponent (63 vs. 108). This is not too surprising – an author might rather choose to present yet another good reason in favour of the central claim, and thereby outweigh the objection, or she might pose the possible objection in an unalluring manner, signalling that counter-attacking or outweighing is not even necessary.

Linked premises do not occur very frequently. In total, there are 21 linked premises found in corpus, which represents only 4.5% of the function bearing segments. With 12 instances linked premises are comparatively less frequent in the texts written in the text generation experiment than in the constructed text with nine instances. The great majority (16 of 21) of linked premises are part of a proponent's support (the most frequent structure), but we also find instances of all other combinations, i.e. linked supports or attacks of both the proponent or opponent role.

**Attachment distance**

One aspect of argumentation structure that makes its automatic recognition especially challenging is the possibility of long distance dependencies. Although segments are often connected locally, i.e. they are supporting or attacking adjacent segments, there may very well be direct argumentative relations across the entire text, even between the very first and the very last segment of a text. It is thus worthwhile to investigate the degree to which we find these non-local relations in our corpus.

To this end, we calculated the distance and the direction of attachment for every relation annotated in the corpus (464 in total, the remaining segments functioning as central

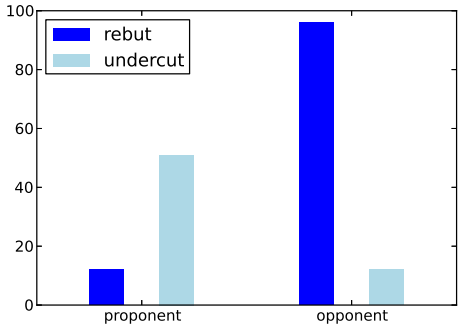



Figure 5.1: Attack moves against argumentative role.

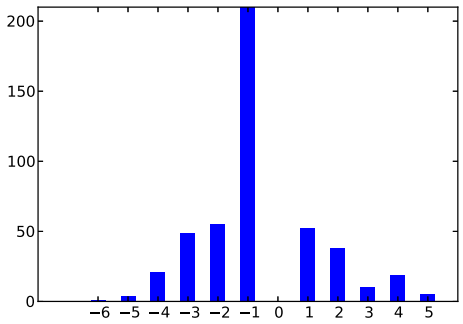



Figure 5.2: Attachment distance and direction (negative distances directed backwards and positive distances directed forwards).

claims). An attachment distance of -1 means that the target of the argumentative relation directly precedes the source, a distance of +1 corresponds to a target immediately following the source. For segments targeting a relation instead of another segment, as it is the case for undercutters and linked premises, we considered the position of the source of the targeted relation. For example, the undercutting segment 3 in the graph in Figure 1 has an attachment distance of -1, as it undercuts the relation of the previous segment 2.

The distribution of distances and directions of attachment found in the corpus is shown in Figure 5.2. The great majority (45%) of argumentative relations attach to the immediately preceding segment. Another 11% attach to the following segment. In total, 56% of the relations hold between adjacent segments, so conversely nearly half of the segments do not attach locally. Considering that our texts are relatively short, it is to be expected to find even more non-adjacent relations in longer texts. For instance, Stab and Gurevych [2014b] report a rate of 63% of non-adjacent relations in their corpus of student essays.

### Linearisation strategies

The final feature of the argumentation graphs we want to investigate is how authors linearise their arguments in the text. This has already been covered to some degree above when we studied at which positions in the text the central claim and objections are typically expressed. In the following, we combine this with the direction of attachment and distinguish four different simple linearisation strategies, which are summarised in Table 5.

The first strategy involves only backward relations, where the author opens her text with the central claim (c) and then presents a series of reasons, possible objections, and counters, all of them directed backwards (b), targeting propositions made in prior segments. The

| linearity strategy | pattern | frequency |
|---|---|---|
| backward | c b+ | 50% |
| forward | f+ c | 5% |
| forward-backward | f+ c b+ | 13% |
| other | other | 31% |

Table 5.6: Ratio of texts matching different linearisation strategies.

second linearisation strategy unfolds the argumentation the other way around, with only forward relations. The author first starts with premises and successively draws conclusions from them (f) until she finally reaches the central claim of the text. The third strategy combines these two patterns, presenting the central claim in the middle of the text. It naturally involves a switch of attachment direction after the central claim. All other texts not matching one of these three strategies involve a change in the direction of argumentation independent of the presentation of the central claim.

As shown in Table 5.6, the first strategy which opens with the central claim and argues for it with only backward relations, is the dominant one found in half of the texts. The reverse strategy is used only rarely, while the mixed strategy appears in at least 13% of the texts. Most interestingly, about 31% of the texts do not follow these strict patterns. As an example, see Figure 5.3: This text's linearisation pattern corresponds to 'fbfbcb', featuring multiple changes in direction before the central claim is stated.

### 5.2.4 Corpus delivery

The corpus has been published online[1] and freely distributed under a Creative Commons BY-NC-SA 4.0 International License.[2] The annotated graph structures are stored in the Potsdam Argumentation XML format (PAX), a both human- and machine-readable format, similar to GraphML [Brandes et al., 2002]. The corpus repository contains a well-documented specification of the format in form of a document type definition.

For both versions of the corpus, the German and the English one, we provide the raw source text, the annotated argumentation graph in PAX (primarily for machine reading), as well as a graphical argument diagram such as the one in Fig. 1, in order to facilitate human inspection of the structures. An importer for the PAX format has also recently been added to the Carneades Tools[3], allowing to map and evaluate the graphs of our corpus.

[1] `https://github.com/peldszus/arg-microtexts`
[2] `https://creativecommons.org/licenses/by-nc-sa/4.0/`
[3] `https://carneades.github.io/`

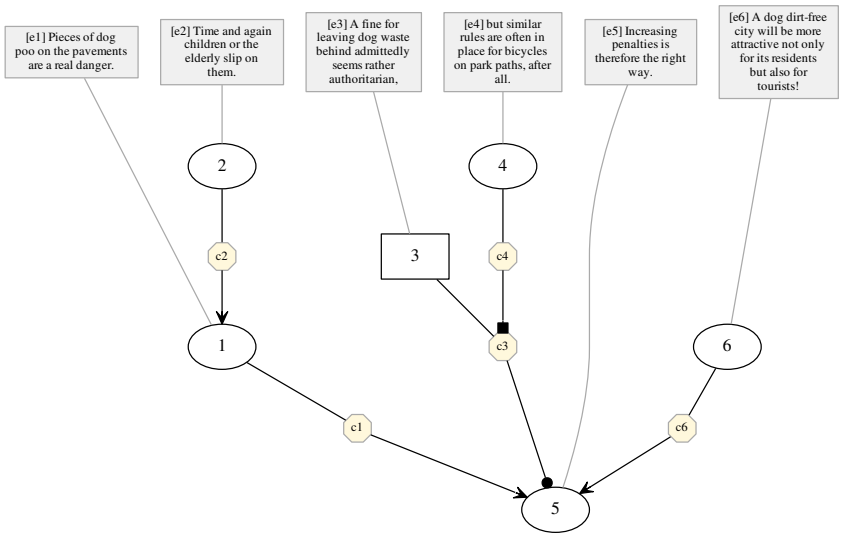



Figure 5.3: An example text (micro_k031) exhibiting multiple direction changes.

### 5.2.5 Conclusion

We presented a freely available parallel corpus of short argumentative texts. These micro-texts are 'authentic' in the sense that the vast majority was written by subjects not involved in the research, and the trigger questions concern issues of daily life and public interest. At the same time, they are 'constrained' because we provided the subjects with some instructions on target length and form. This was done in order to obtain a relatively homogeneous data set that allows for studying properties of the argumentation. For the same reason we decided to do a moderate 'cleaning' of the texts, which on the one hand reduces 'authenticity' but on the other contributes to uniformity and – for many purposes – usability.

Research in automatic argument mining typically targets social media contributions in their original form and often focuses on the task of argument identification and local relation identification. While the design of our corpus differs from this orientation, we still think that the data can be useful for purposes of feature engineering and as supplemental training data. Finally, we consider our data set to be a reasonable starting point for the task of automatic prediction of text-level argumentation structure, before tackling more complex text genres.

## 5.3 Useful transformations

We have argued that the graph structures defined by our scheme are adequate for representing the structure of argumentation (on the desired level of granularity). They meet our requirements and most of the desiderata we articulated in Chapter 2. However, this does not guarantee that these structures fit perfectly with the approaches and algorithms brought forward by NLP research. Some structural configurations might be incompatible or too complex; some distinctions might be too fine-grained to be modelled with already existing technologies.

In the following, we wish to present two transformations of the argumentation graphs which will help to tackle the challenging task of automizing the recognition of argumentation structure. The first is a structural transformation where the argumentation graphs are converted to dependency trees, which will enable us to apply well-researched graph algorithms for this type of structure. The second is a reduction of the set of relation labels, a useful simplification to start with basic and not too fine-grained distinctions before moving on to solve more complex problems.

### 5.3.1 Dependency conversion

One special challenge with the argumentation structures is that they presuppose three different types of nodes: EDUs representing elementary units of propositional content; ADUs representing argumentatively relevant claims, which can consist of multiple propositions; and relation nodes which serve as an attachment point for relations targeting other relations, such as in structures with linked premises or undercutting relations.

The fact that there are two possible basic units, ADUs or EDUs, represented in the graph is less problematic once the decision is made whether or not to tackle the argumentative relevance / segmentation task in the desired model. One simply has to decide which unit to take as the basic unit. In Chapter 6, we will present both models operating on ADUs and those that operate on EDUs.

The possibility to point relations to other relations, however, poses a real challenge. Many efficient algorithms have been proposed and established in NLP research for the prediction of tree structures, as well as of different members of the graph family. If these techniques could be used for the prediction of argumentation structures, this would be a valuable first step to understand and compare the complexity of this discourse processing task. However, these techniques require standard graph-theoretic structures without the special construct of relations pointing to relations. The Potsdam Argumentation XML format (PAX), in which the microtexts are serialised, is able to represent these relations to relations, because each relation is internally interpreted as a sub-graph: Instead of directly connecting source and target, this connection is split up into one incoming edge that connects the source with

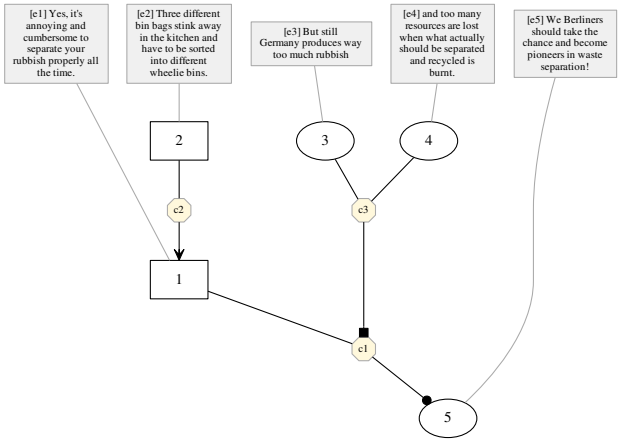
the relation node and one outgoing edge from the relation node to the target node. This way, the relations-to-relations can be represented in standard graph theoretic terms. Yet for automatic prediction, this is still impractical because it is not possible to simply reason over relations, but required to reason over subgraphs representing relations.

The solution we want to promote here is to convert the argumentation graph into a dependency tree. The first step to achieve this is to redirect all relations pointing to edges to point to the source node of the targeted edge. Undercutting relations, for instance, will target not the relation that is undercut, but the source node of the relation that is undercut. For linked relations, which have more than one source, the left-most source node is taken as the head, while all further sources attach to the head with a LINK relation. This way, we arrive at an argumentation graph that is free of relation nodes. The relations between ADUs can then be directly used as the edges of a dependency tree.

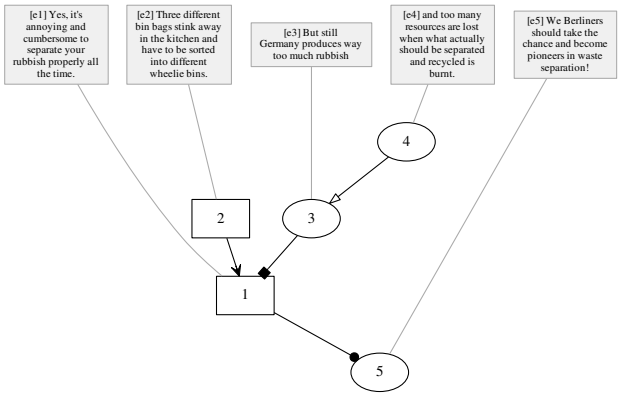
An example conversion is shown in Figure 5.4. The initial argumentation graph with relation nodes, an undercutter, and linked premises is shown in the upper Figure 5.4a. Redirecting those relations that point to other relations to the targeted relation's source node yields the structure in Figure 5.4b, which does not have any relation nodes. The relations there can than be directly used as the edges of a dependency tree, as displayed in the lower Figure 5.4c. In contrast to the argumentation graphs, relation types are visualised in the dependency tree through relation labels instead of arrow types. Note that the argumentative role is not explicitly represented in the dependency graph. However it can be inferred from the structure, assuming that the root of the structure (the central claim) bears the proponent role and that all attacking relations (i.e. rebut and undercut) invert the role.

This dependency conversion is loss-less and can easily be undone, if two conditions hold:

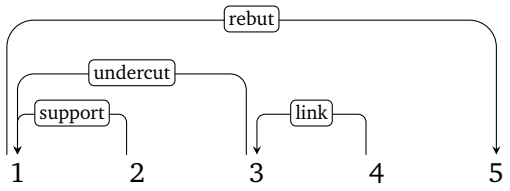
1. *Every relation type should either target only nodes or only relations, but not both.* If one relation type could point to both normal nodes and relations nodes, it would not be possible to determine from the dependency structure whether this relation targets the node or the relation the node is the source of. In our scheme this is not a problem, as undercut and link relations always target relations, while all other relation types always target ADUs. If a scheme does not meet this criterion, it could easily be adjusted by introducing further relation types to disambiguate their target preference.
2. *Every ADU has at most one function, i.e. one outgoing arc.* If one ADU has more than one outgoing relation, it would be impossible to determine from the dependency structure which of the outgoing relations was targeted. Note that this effectively rules out divergent argumentation, where multiple conclusion are drawn from one premise. This might be considered a severe restriction, and it is not hard to construct an example where divergent argumentation occurs. Indeed, it is technically possible to represent divergent argumentation in our scheme and in the XML serialisation.




(a) Full argumentation graph with relation nodes.




(b) Reduced argumentation graph without relation nodes.




(c) Dependency conversion of the ADU relations.

Figure 5.4: Example dependency conversion of micro_b001 in ADU segmentation.

Still, we consider it acceptable to restrict the structures this way for two reasons: First, we have not found instances of argumentation in our corpus that require divergent structures. Secondly, this restriction is also made in most other schemes for coding argumentation in natural language, as these typically assume tree structures and the restriction to one outgoing arc is an essential tree property.

It should be emphasised that the redirection of the edges is a representational trick. It does not imply a change of interpretation of the structure. Redirecting for example an undercutter to the source of the undercutted relations still means that the relation (the source of which is targeted) is undercut. The original argumentation graph and the dependency tree are supposed to represent the very same argumentation structure.

We will use this dependency conversion when predicting global argumentation structures and when comparing it to other discourse structures.

### 5.3.2 Label-set reduction

The dependency trees representing the argumentation structures could be further transformed by reducing the relation set to the simple, binary distinction between support and attack. The two attacking relation types (rebut and undercut) are subsumed under a general attack relation type, and all other relation types (support, example, and link) are subsumed under a general support relation type. Contrary to the dependency conversion, this transformation is of course a lossy one.

This reduction may, however, be useful for several reasons: First, such a coarse-grained binary distinction facilitates comparisons with other datasets. There are other datasets that use the support-vs-attack distinction, such as for example the corpus of student essays of Stab and Gurevych [2014b], or larger parts of the AIFdb [Reed et al., 2008]. Also, this basic distinction can be useful for comparison with datasets that also have a more fine-grained but different relation set than our corpus, such as for example the more elaborate set of argumentation schemes of Walton et al. [2008]. If a similar mapping from their fine-grained relation set to the support-vs-attack distinction can be defined, this would serve as a common base for comparison. Another motivation might be that the fine-grained distinctions are too hard to model automatically. This could be either because distinguishing between different sorts of support and attack is determined rather by the semantics of the related segments than by linguistic signals on the surface. Alternatively, a reduction of the relations set is advised when some of the fine-grained relation types are too infrequent in the corpus to be successfully learned.

In the following, we want to discuss the adequacy of this reduction and highlight some issues we find noteworthy: Reducing 'support' and support by 'example' to a general support class is regarded as uncontroversial. The situation is less clear for the reduction of 'link' to 'support'. Recall that linked premises are triggered by the question 'Why is that relevant?' in

the dialectical exchanges of Freeman [1991, 2011]. Providing justification for the relevance of a claim can certainly be regarded as support in a general sense, so this mapping seems reasonable. For linked attacks the effect might be surprising: The function of the first premise of the attack remains an attack, but all other premises of the linked attack then serve as supporting the first. The result of the reduction is a switch in argumentative function (but of course not in argumentative role). An alternative approach would be to treat linked premises simply as separate premises, when reducing the 'link' relation label. This would result in multiple supports in the case of a linked support, and multiple attacks in the case of a linked attack, thus preserving the overall argumentative function. But it would also require a more complex, custom interaction between dependency conversion and relation-set reduction.

The effect of reducing 'rebut' and 'undercut' to a general attack class also has to be investigated carefully. Collapsing these two relations labels is straightforwardly done as both are types of attack, but the reduction again interacts with the structural changes of the dependency conversion. Recall that the undercutting attack now targets the source, i.e. the premise of the undercutted relation, but still is interpreted as undercutting the relation. When unifying the two attacking relation types, it would of course not be possible to discern which attack was meant to be interpreted as targeting the relation. What was an undercutter is then simply understood as attacking the premise of the (formerly undercut) relation. In this case, when both dependency conversion and label-set reduction are applied and it is thus clear that this whole operation will be irreversible, it might actually be better to redirect the undercutters to the conclusion and not the premise of the undercut relation, as undercutters surely weaken the conclusion (by means of undercutting), but typically do not weaken the premise.

We will not go into further detail with this discussion. In our experiments we will apply both transformations if desired as it was described above, i.e in two separate steps, where the relation set reduction operates on the output of the dependency conversion. Still, we want to make the reader aware that there might be situations where we want to derive support-v-attack dependency trees directly using custom transformation functions in order to achieve a specific result that fits our needs.

These transformations can be applied when loading the corpus from the PAX XML serialisation. The (reduced) dependency structures themselves are not stored in a defined interchange format, but can easily be serialised in a table format such as tab separated values or as a JSON data structure. Note that we will no use these transformations in all of our experiments. We will report specifically for every experiment, whether and which transformation have been used to preprocess the argumentation structures.

## 5.4 Additional annotation layers on the microtext corpus

The microtext corpus has been extended with further annotation layers since it has first been made available. One very recent example is the annotation with Situation Entity Types [Becker et al., 2016]. What we want to present in this section are a more fine-grained segmentation into EDUs, as well as two additional layers of discourse structure annotation: one for Rhetorical Structure Theory (RST) [Mann and Thompson, 1988], the other for Segmented Discourse Representation Theory (SDRT) [Asher and Lascarides, 2003].

To date, it has been difficult to compare the two accounts, RST and SDRT, on empirical grounds, since there were no directly-comparable parallel annotations of the same texts. To improve this situation, we took the existing microtext corpus and added layers for RST and SDRT. To this end, we harmonized the underlying segmentation rules for minimal discourse units, so that the resulting structures can be compared straightforwardly. In addition to the opportunity to compare two discourse theories, RST and SDRT, with each other, we also see chances for the study of argumentation, as this resource will enable the study of the correlation between discourse structure and argumentation structure on empirical grounds, something which has not been undertaken in depth yet.

Furthermore, besides all theoretic interest, this extension of the microtext corpus enables us to progress application-wise. On the one hand, we will obtain argumentation structures based on a fine-grained segmentation into EDUs, which will allow us to train models working directly with EDUs, integrating the argumentation segmentation task. On the other hand, we can study practically how easy or hard it is to derive argumentation structures automatically from given discourse structures. Both problems will be approached in Chapter 6.7 and 6.8, respectively. The basis for this is the corpus presented in this section.

### 5.4.1 Data

The extensions of the corpus was done on the English version of the microtext corpus only, i.e. the finer EDU segmentation as well as the creation of the additional RST and SDRT annotation layers was done on the basis of the English text.

Mapping the new annotations back to the German version of the corpus must be set aside for future work as this cannot be done automatically: Recall that the translation of the microtext corpus was constrained in such a way that ADU segmentation and linearisation is preserved. It does not follow necessarily that this also holds for the inner EDU segmentation of an ADU. A manual analysis of all segments will thus be required.

### 5.4.2 Harmonized discourse segmentation into EDUs

In order to achieve comparable annotations on the three layers, we decided in the beginning of the project to aim at a common underlying discourse segmentation. For a start, the

argumentation layer already featured ADU segmentation; these units are relatively coarse, so it was clear that any ADU boundary would also be an EDU boundary in RST and SDRT. On the other hand, the discourse theories often use smaller segments. Our approach was to harmonize EDU segmentation in RST and SDRT, and then to introduce additional boundaries on the argumentation layer where required, using an 'argumentatively empty' JOIN relation.

As explained in the next two sections, RST and SDRT annotation start from slightly different assumptions regarding minimal units. After building the first versions of the structures (by the Toulouse and the Potsdam group, respectively), we discussed all cases of conflicting segmentations and changed both annotations so that eventually all EDUs were identical.

The critical cases fell into three groups:

- 'Rhetorical' prepositional phrases: Prepositions such as 'due to' or 'despite' can introduce segments that are rhetorically (and sometimes argumentatively) relevant, when for instance a justification is formulated as a nominalized eventuality. We decided to overwrite the syntactic segmentation criteria with a pragmatic one and split such PPs off their host clause in cases where they have an argumentative impact.
- VP conjunction: These notoriously difficult cases have to be judged for expressing either two separate eventualities or a single one. We worked with the criterion that conjoined VPs are split in separate EDUs if only the subject NP is elided in the second VP.
- Embedded EDUs: For technical reasons, the Potsdam Commentary Corpus annotation had not marked center-embedded discourse segments; and, in general, different RST projects treat them in different ways. In SDRT, however, they are routinely marked as separate EDUs. In the interest of compatibility with other projects, we decided to build two versions of RST trees for texts with embedded EDUs: One version ignores them, while the other splits them off and uses an artificial 'Same-Unit' relation to repair the structure (cf. Carlson et al. [2003]).

As a result of the finer segmentation, 83 ADUs not directly corresponding with an EDU have been split up, so that the final corpus contains 680 EDUs.

### 5.4.3 Annotation procedure

**ARG**

The argumentation structures based on ADUs are already annotated in the corpus. What remains to be done is to apply the new, more fine-grained segmentation into EDUs to the serialized argumentation structures. This amounts to introducing a 'joint'-node as the parent node of all adjacent EDUs that are supposed to constitute a single ADU, and then grounding

the ADU in this joint node. See Chapter 3.5 for details. This process was applied automatically.

An example argumentation structure is shown in Figure 5.6a, but note that for the sake of brevity of the discourse structure comparison it is a text where the ADU segmentation was already in accord with the EDU segmentation.

**RST**

The RST annotations have been created according to the guidelines [Stede, 2016b] that were developed for the Potsdam Commentary Corpus [Stede and Neumann, 2014, in German]. The relation set is quite close to the original proposal of Mann and Thompson [1988] and that of the RST website[4], but some relation definitions have been slightly modified to make the guidelines more amenable to argumentative text, as it is found in newspaper commentaries or in the short texts of the corpus we introduce here. Furthermore, the guidelines present the relation set in four different groups: primarily-semantic, primarily-pragmatic, textual, multinuclear. The assignment to 'semantic' and 'pragmatic' relations largely agrees with the subject-matter / presentational division made by Mann / Thompson and the RST website, but in some cases we made diverging decisions, again as a step to improve applicability to argumentative text. For example, we see EVALUATION as a pragmatic relation and not a semantic one. 'Textual' relations cover phenomena of text structuring; this group is motivated by the relation division proposed by Martin [1992], but the relations themselves are a subset of those of Mann / Thompson and the website (e.g., LIST, PREPARATION). Finally, the 'multinuclear' relations are taken from the original work, with only minor modifications to some definitions. The annotation procedure explained in the guidelines suggests to prefer pragmatic relations over semantic ones in cases of ambiguity or doubt, which is also intended as a genre-specific measure.

All RST annotations on the microtext corpus were done using the RSTTool[5]. An example annotation is shown in Figure 5.6c.

In the resulting corpus, there are 467 instances of RST relations, hence on average 4.13 per text. The most frequent relation is by a large margin REASON (178 instances), followed by CONCESSION (64), LIST (63), CONJUNCTION (44), ANTITHESIS (32), ELABORATION (27), and CAUSE/RESULT (22); other relations occur less than 20 times.

**SDRT**

The SDRT annotations were created following the ANNODIS annotation manual [Muller et al., 2012b] which was based upon Asher and Lascarides [2003]. The amount of information about discourse structure was intentionally restricted in this manual. Instead it focused

[4] http://www.sfu.ca/rst
[5] http://www.wagsoft.com/RSTTool/

essentially on two aspects of the discourse annotation process: segmentation and typology of relations. Concerning the first, annotators were provided with an intuitive introduction to discourse segments, including the fact that we allowed discourse segments to be embedded in one another as well as detailed instructions concerning simple phrases, conditional and correlative clauses, temporal, concessive or causal subordinate phrases, relative subordinate phrases, clefts, appositions, adverbials, coordinations, etc. Concerning discourse relations, the goal of the manual was to develop an intuition about the meaning of each relation. Occasional examples were provided, but we avoided an exhaustive listing of possible discourse markers that could trigger a particular relation, because many connectives are ambiguous and because the presence of a particular discourse connective is only one clue as to what the discourse relation linking two segments might be. The manual also did not provide any details concerning the structural postulates of the underlying theory, including constraints on attachment (the so-called 'right frontier' of discourse structure), crossed dependencies, and more theoretical postulates. The goal of omitting such structural guidelines was the examination of whether annotators respected the right-frontier constraint or not [Afantenos and Asher, 2010]. For the purposes of this annotation campaign we used the Glozz annotation tool.[6]

One structural feature that distinguishes SDRT graphs from RST trees is the presence of complex discourse units (CDUs). CDUs are needed in SDRT in order to give an explicit representation of the exact scope of a discourse relation in the discourse structure. If an argument of a discourse relation involves several EDUs and perhaps even a small discourse structure, we need to group them together to form a single argument for the relation. As an example, consider the SDRT structure shown in Figure 5.6e: Elementary Discourse Units are designated with numbers (1 through 5) while Complex Discourse Units are represented by $\pi_1$ and $\pi_2$, dashed lines indicating member-hood in a CDU.

The SDRT corpus contains 669 EDUs, 183 CDUs, and 556 relations. The most frequent relations are CONTRAST (144), ELABORATION (106), CONTINUATION (80), RESULT (76), EXPLANATION (55), PARALLEL (26), and CONDITIONAL (23), while the rest had fewer than 20 instances.

### 5.4.4 Corpus delivery

The corpus is published online[7] and freely distributed under a Creative Commons BY-NC-SA 4.0 International License.[8] The finer-segmented argumentation structures are stored as before in the Potsdam Argumentation XML format (PAX). RST structures are serialized in

[6] `http://www.glozz.org`
[7] `https://github.com/peldszus/arg-microtexts-multilayer`
[8] `https://creativecommons.org/licenses/by-nc-sa/4.0/`

the RSTTool's '.rs3' XML format. SDRT graphs are given in the XML representation of the Glozz annotation tool. In addition, the pure unsegmented text is made available.

### 5.4.5 A common dependency format

In the following, we propose a common dependency format which all three annotation levels are converted to. In our case, dependency structures are graphs whose nodes represent the EDUs and whose arcs represent the discourse relations between the EDUs. Given this representation, it will be easy to compare and map to each other the different discourse structures.

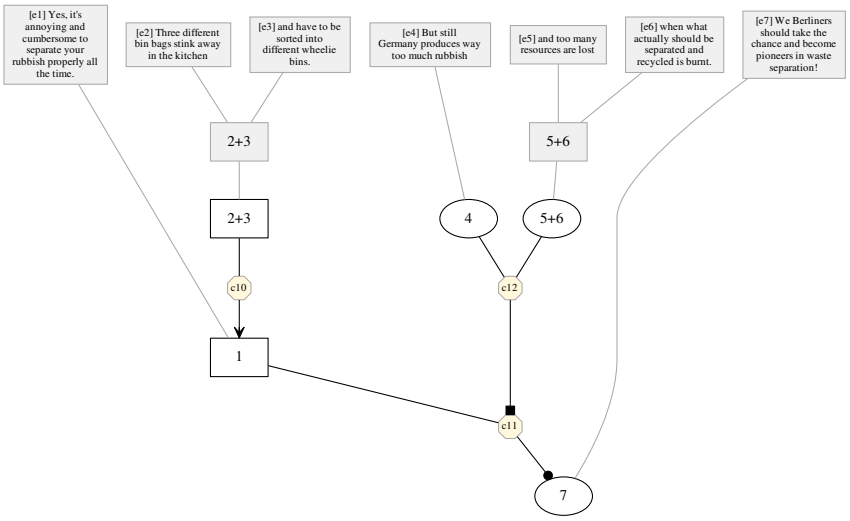
#### Motivation

Calculating correlations between argumentation and discourse as well as between the two discourse corpora themselves requires converting the annotations from their tool-specific XML formats (RSTTool, Glozz) into a common format. This is not an easy task since the two theories have fundamental differences at least as far as scoping of relations is concerned. We consider dependency structures as a reasonable candidate for a common format capturing the structures of RST and SDRT, as it had also been proposed earlier by Danlos [2005]. This is further facilitated by the fact that—with the exception of embedded EDUs in SDRT, for which we used the Same-Unit 'relation' in RST—both annotations use the same EDUs.

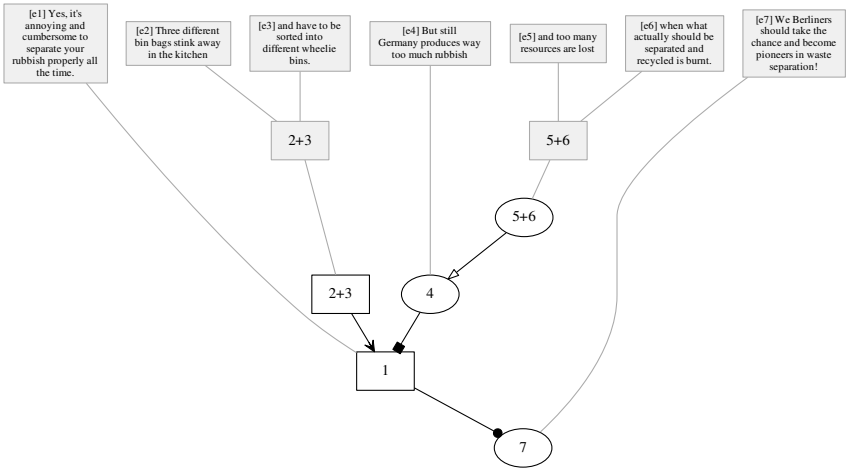
Furthermore, future experiments on discourse parsing and argumentation structure analysis can be facilitated by using a common format for all annotations; however, we need to be cautious when it comes to theory-specific discourse parsing, since the mapping between the theories is not one-to-one, as we will see.

#### From Discourse Structures to Dependency Structures

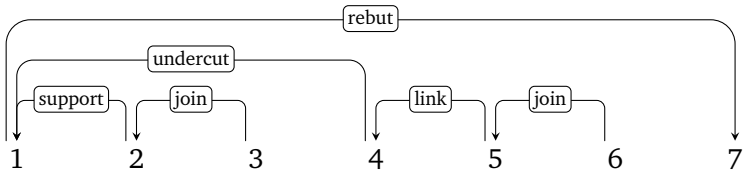
For the **argumentation structures**, a dependency conversion has already been presented in the previous section (5.3.1). In case every ADU corresponds to exactly one EDU, the conversion can be used as defined. This is the case in our example (see the dependency conversion of the argumentation graph in Figure 5.6b). If, however, there is an ADU in the graph that consists of multiple EDUs, the many-to-one relation between EDUs and ADUs has to be considered in the dependency conversion. For this, the procedure is augmented by a rule that translates ADUs spanning over multiple EDUs (represented through the joint-node) into flat left-to-right JOIN relations, with the leftmost EDU being the head of the original argumentative relation of the ADU. An example of this is shown in Figure 5.5, analogous to the initial ADU-based dependency conversion example. Note that in consequence of the procedure, the leaves of this dependency tree are not ADUs anymore, but of course EDUs.




(a) Full argumentation graph with relation nodes.




(b) Reduced argumentation graph without relation nodes.




(c) Dependency conversion of the relations base on EDU segmentation.

Figure 5.5: Example dependency conversion of micro_b001 in EDU segmentation.

For the **rhetoric structures** we follow the procedure that was initially proposed by Hirao et al. [2013] and later followed by Li et al. [2014b]. The first step in this approach includes binarizing the RST trees. In other words we transform all multi-nuclear relations into nested binary relations with the leftmost EDU being the head. Dependencies go from nucleus to satellite. For illustration, a dependency structure for the RST tree is shown in Figure 5.6d.

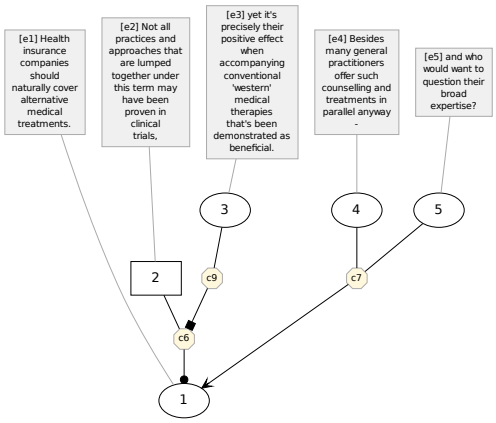
Concerning the **SDRT graphs**, predicting full SDRSs with CDUs has been to date impossible, because no reliable method has been identified in the literature for calculating the CDU membership relation. Instead, most approaches [for example Muller et al., 2012a, Afantenos et al., 2015, Perret et al., 2016] simplify the underlying structures by a *head replacement strategy* that removes nodes representing CDUs from the original hypergraphs and replacing any incoming or outgoing edges on these nodes on the *heads* of those CDUs, forming thus dependency structures and not hypergraphs. We adapted this strategy here. It should be pointed out that the expressive capacities of SDRT outrun those of theories that require tree-like discourse structures, and Afantenos et al. [2015] have shown that this expressive capacity is needed for multi-party dialogue. For the purposes of this study, nevertheless, with a corpus of short monologue texts, this restriction is acceptable. The result of the transformation for the example text is shown in Figure 5.6f.

It is important though to note that those transformations are not one-to-one, meaning that although transforming RST or SDRT structures into dependency structures always produces the same structure, going back to the initial RST or SDRT structure is ambiguous. This is different from the dependency conversion of the argumentation structures, where we argued that it is indeed reversible if some conditions are met and the relation set is not reduced (see Section 5.3).

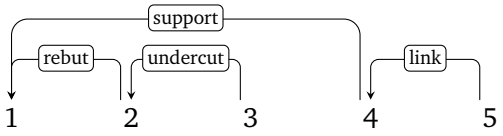
### 5.4.6 Conclusions

Our triply annotated corpus, with discourse annotations in the style of RST and SDRT and with an argumentation annotation, opens up several interesting lines of research that can now be pursued. Studying the connections between these theories is facilitated by the fact that we have transformed all structures into a common dependency format on the basis of a harmonized discourse segmentation. This way, it is possible to bring the study of the relation between RST and SDRT on empirical grounds, which has been advanced only on the theoretical side (for example by Venant et al. [2013] who provide a common logical representation and prove what correspondences are possible).

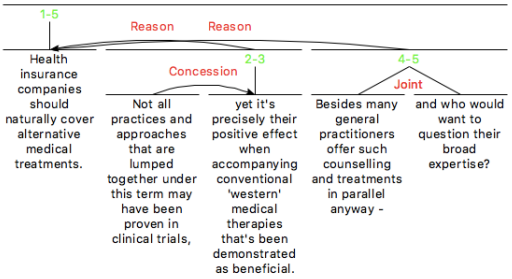
Moreover, we can now study how argumentation graphs map onto discourse structures and vice-versa. We believe this will be an important step to a better understanding of how various argumentation forms depend on discourse structure and, more generally, how argumentation is linguistically realized.



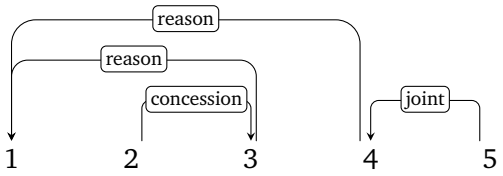

(a) ARG



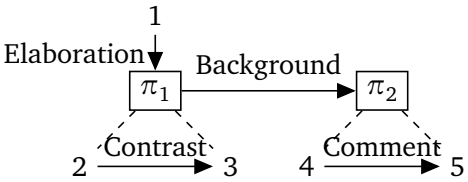

(b) ARG dependency tree



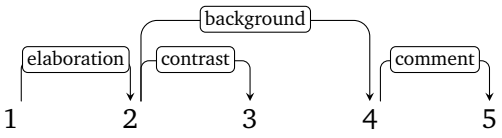

(c) RST




(d) RST dependency tree




(e) SDRT




(f) SDRT dependency tree

Figure 5.6: Example dependency conversions of ARG, RST and SDRT annotations of the text micro_b013. The original structure is shown in the left column, the dependency conversion in the right column.

Similarly, the extended corpus offers various possibilities for future work on the application side. With the EDU segmentation we can now define argumentation models that learn how to map EDUs to ADUs, or we can directly integrate the argumentative segmentation into the structure building process by considering the JOIN relation to be part of the argumentation structure. We will study this later in Chapter 6.7. Unfortunately, the fine-grained segmentation is still missing for German and we will have to leave it for future work. Furthermore, we can start to exploit discourse structure for argumentation mining, either by extending a model for argumentation with features that prove to be successful in discourse parsing of the input, or with features that represent aspects of predicted discourse structures. A third alternative is to even aim to learn argumentation structure directly from discourse structure, something we will attempt in Chapter 6.8. Out of the scope of this work, but nonetheless an interesting path for future research, is the question as to whether different discourse structures of the same text could be learnt jointly.

## 5.5 Conclusions

In this chapter, we have presented the 'microtext' corpus, which we will use in our experiments in automizing the recognition of argumentation structure. It is the first parallel corpus of argumentation (with a German and English version) and the first corpus annotated according to multiple theories of global discourse structure:

- In Section 5.2, we reported on the creation, translation, and annotation of the corpus and provided statistics of the variety and the linguistic realisation of argumentation structure in the corpus.

- We then showed in Section 5.3 how to transform the annotated argumentation structures to dependency structures. This was achieved through a mapping that is lossless under the given constraints and allows us to simplify the more complex properties of the argumentation graphs. This transformation enables us to use standard algorithms and tools for dependency trees. Furthermore, we discussed a transformation that reduces the set of relations used in the dependency trees and motivated in which situations it would be useful.

- In Section 5.4, we introduced further annotations layers of the microtext corpus, most importantly, two additional discourse structure annotations according to RST and SDRT. These parallel annotations and the common dependency format invite to study the interrelations of argumentation structure, RST and SDRT. Moreover, we are now enabled to test whether and to what extent argumentation structure can be learned from RST or SDRT structures. Finally, the new annotation layers are based on a more fine-grained segmentation of the texts into EDUs, which also has been applied to argumentation structures. This should help us to derive models that directly work on the output of EDU segmenters.

Finally, we want to briefly summarise and name the different transformed versions of the argumentation structure layer of the corpus we now have at hand:

- **microtext-graph-ADU:** the corpus in its original ADU-based annotation represented as argumentation graphs; available in German and English;
- **microtext-graph-EDU:** the corpus in its extended EDU-based annotation represented as argumentation graphs; available in German and English;
- **microtext-dep-ADU-full:** the corpus converted to ADU-based dependency trees, with the full label set; available in German and English;
- **microtext-dep-ADU-reduced:** the corpus converted to ADU-based dependency trees, with the reduced label set; available in German and English;

- **microtext-dep-EDU-full:** the corpus converted to EDU-based dependency trees, with the full label set; available only in English;
- **microtext-dep-EDU-reduced:** the corpus converted to EDU-based dependency trees, with the reduced label set; available only in English.

After having defined our scheme, validating it in annotation experiments, and using it to create a resource with argumentation structures, it is now time to investigate how these argumentation structures can be automatically recognised from text.

# 6 Automatic Recognition of Argumentation Structure

After first having identified a theory of the structure of argumentation, devising and evaluating an annotation scheme, and creating a corpus annotated with argumentation structure, we now face the aim of this work: automatically deriving argumentation structures from text.

This chapter will present different studies, approaching this problem from different views. The first study, presented in Section 6.3, will start with local models of different aspects of the argumentation structure. Section 6.4 elaborates on one of these aspects, the argumentative role of the opponent, in comparing the predictability of objections in our short microtexts as well as in ProCon commentaries. We will then combine the predictions of local models in a global model of the argumentation structure, the "Evidence Graph" model, which is the proposed approach of this thesis, in Section 6.5. An improved version of this model is shown in Section 6.6 and compared against other structure decoders proposed in the literature. The approach is then used in Section 6.7 to predict more fine-grained argumentation structures. Finally, in Section 6.8, we will demonstrate that this approach can also be applied to other problems of predicting argumentation structure, exemplified for the task of mapping rhetorical structures to argumentation structures. But before we report on these studies, we will again lay down the related work and our methodology.

**Previously published material**

Section 6.3 has been published as [Peldszus, 2014]. The study in Section 6.4 was first presented in [Peldszus and Stede, 2015b]. The Evidence Graph model covered in Section 6.5 has been published as Peldszus and Stede [2015a]. The improved version of the model and the comparison with other decoders, that is presented in Section 6.6, is joint work with the colleagues from Toulouse [Afantenos et al., under review]. My main contributions in this study were (i) improving the base models by exchanging and extending the features, (ii) assisting in the development of the ILP constraints, and improving the decoder interface and (iii) evaluating the results. The results in Section 6.7 have not been published yet. Section 6.8, finally, was published in [Peldszus and Stede, 2016b]. Here, my main contribution was the experiment on predicting argumentation structures from the annotated RST structures.

## 6.1 Related work

In this overview of related work, we will focus on tasks of argumentation mining that are closely related to the goal of this work: predicting the structure of argumentation. As presented in Chapter 1.2, this amounts to the tasks of ADU identification, ADU classification, relation identification, and relation classification. We will thus not consider work on e.g. the prerequisite step of identifying persuasive texts on a document basis, or ultimately on reconstructing enthymemes. Also not covered here are information retrieval oriented approaches such as retrieving supporting or opposing claims or opinions for a given claim from a larger database, or approaches focusing on the question as to whether a document or a single claim addresses or adheres to a given prompt or not.

### 6.1.1 Early work

One of the earlier studies on computational argument analysis with natural language understanding is that of Cohen [1983, 1987a,b]. She investigated strategies of serialising arguments in text and presented a model which incrementally processed the propositions of the argumentation, and consecutively built a tree structure representing the argumentation. The default construction procedure produced claim-first or claim-final argument chains. Cue phrases (connectives or phrases such as 'in addition' or 'as a result') were required to signal non-default moves that yield parallel or flipped structures. However, the mechanism relied on an 'evidence oracle', which hints to the model whether one proposition is evidence for another proposition. A prototypical implementation for this component was later sketched by Young and Cohen [1991]. It required extensive logic modelling. The prerequisite of robustly translating natural language to predicate logic was not addressed, however.

A cognitive modelling approach was taken by Alvarado et al. [1986], Alvarado [1990], who presented a model of text comprehension of editorial text in a political-economical domain. The system heavily relied on logical representations of domain knowledge, goals, plans, and strategies. Using the output of a semantic parser the model inferred a representation of the author's beliefs and integrated them in an argument map. On the basis of this representation, the comprehension system then derived answers to a posed question about the author's beliefs and reasoning. The model was exemplified on two short fragments of actual editorials. For applying it to new texts, an extensions of the conceptual and the thematic knowledge base would be necessary. It should be noted, though, that the main aim of this work is to understand the cognitive requirements of processing arguments (in terms of processes and knowledge structures), especially for refutations and accusations, rather than providing a model of the natural language of arguments.

The advent of the renewed interest in recognising argumentation structures from text is closely connected to the dissertation of Mochales Palau [2011] and the influential articles of Moens et al. [2007], Mochales Palau and Moens [2009], Mochales and Moens [2011]. They not only coined the term *argumentation mining*, and were the first to fully automatically derive argumentation structures from unseen text (to be precise from court decisions of the ECHR corpus), but also presented the first results on the AraucariaDB [Reed et al., 2008], which was the first larger and publicly available annotated resource for natural language argumentation. Since then, argumentation mining has become a very active line of research.

In the following, we will first review related work separately for the steps associated with argumentation mining, following the characterisation of the problem from Chapter 1.2, as far as they can be separated. We will then show related approaches that tackle the full problem.

### 6.1.2 ADU identification

Approaches to identify argumentative units can roughly be divided into those aiming for full sentence units, for clause-sized units or those allowing free token-spans. Some approaches, which integrate this with other classification tasks, are presented separately later on.

Moens et al. [2007] performed machine learning to distinguish between argumentative and non-argumentative sentences. As features they proposed n-grams, adverbs, verbs, modal, word-pairs, text statistics, punctuation, argumentative discourse markers, depth of parse tree, and number of subclauses. Multinominal naive Bayes and maximum entropy models were used for classification, with an accuracy of 73% on the AraucariaDB and later with an accuracy of 80% on the court decisions of the ECHR corpus [Mochales Palau and Moens, 2009].

Binary classification of sentences (or multiple sentences) into having an argumentative role or not was also the focus of Florou et al. [2013]. On a smaller corpus of 400 Greek sentences from an environmental rulemaking domain, they compared feature sets with discourse connectives and various morphological properties such as tense and mood; the latter turned out to be very important. Their decision tree model yielded best results when combining connectives with tense and mood.

Song et al. [2014] aimed to classify whether a sentences addresses a critical question of an argumentation scheme (in the sense of Walton) and can thus be considered argumentatively relevant, or does not allude to an aspect of an argumentation scheme and can thus be considered to be irrelevant for argumentation. They trained a logistic regression model on a dataset with 600 student essays about policy decisions, using n-grams, POS-tags, sentence length, and position and word overlap with the essay prompt as features. Their model

scored with $\kappa$=0.44, in comparison to the human annotator performance of $\kappa$=0.60 in this task.

Although sentence-based approaches are by far the most common, some work also uses clause-based segments. An example is the the experiment of Mochales Palau and Moens [2009] on the ECHR corpus. As a preprocessing step for distinguishing premises from conclusions, they constructed their classification items as clauses, obtained from a syntactical parser, and classified them as argumentative or not using the maximum entropy model described above. Although the performance of this model was assessed for sentence-length units, the segment-length units and the impact of automatic clause-splitting was not evaluated separately. Apart from this, we are not aware of any work focusing on argumentative segments as clauses or EDUs exclusively. There are, however, approaches that combine clause-based segmentation with other classification task (see below).

An interesting take on argumentative relevance based on free token-spans was presented by Lawrence et al. [2014], who determined argumentative relevance as the result of a structure prediction process. They first trained two token-based Naive Bayes models to predict opening and closing segment boundaries. As a second step they used semantic distance measures between a predicted segment and its predecessor of a LDA topic model to derive a non-directed hierarchical structure. Segments which do not exceed a specified similarity threshold to any preceding segment, are considered non-relevant for the argument and are not integrated into the structure.

### 6.1.3 ADU classification

We now turn to segment type classification, where segments of clause or sentence size, or similar token-spans are labelled with categories that are relevant to the analysis of argumentation. Different type-systems have been proposed for different purposes, including their role in argumentation structure (claims, premises), their role in the text (central claim/thesis or not), their rhetorical or argumentative function in the text (supporting, attacking), their verifiability, their dialectical relation to the main claim (the roles of proponent versus opponent), etc. We will consider examples for each of them.

For legal texts of the ECHR corpus, Mochales Palau and Moens [2009] demonstrated in their influential work how to classify the segment of a text into **premises and conclusions**. The SVM model takes as input clauses that have been predicted to be argumentative using the Maximum Entropy model described above. They obtain an F-score of 74% for conclusion and 68% for premise. As features they used among others: subject type, main verb tense, main verb argumentative types, rhetorical cue phrase classes, argumentative cue phrase classes, and a contextual feature with the prediction for the previous and next segment. They also report results of a CFG (used for predicting full argumentation structures; see

below) for only the segment type classification, reaching somewhat lower scores with 67% for conclusion and 64% for premises.

Eckle-Kohler et al. [2015] experimented with different discourse marker resources. Using the premise versus conclusion classification task on gold argumentative units in German newspaper text, they found that discourse markers alone help to beat a majority baseline, but that lists of non-open-class words and unigrams still significantly outperform discourse markers only.

We already stressed the importance of correctly identifying the **central claim** or conclusion of a text. In the domain of student essays, this overlaps with the task of identifying the thesis and also the possibly thesis-restating text-final conclusion. Burstein and Marcu [2003] proposed a decision-tree classifier using boosting, with positional features, cue words, and features extracted from automatic RST discourse parses. In a topic-independent evaluation, the approach reached an average F-score of 54% for thesis and 80% for conclusion segments. Kwon et al. [2007] identified subjective main claims in public comment on rule-making proposal in the environmental domain. Using n-grams, counts of positive and negative subjective words, position in text, and main verb predicate triples from a parser, their boosting model obtained an F1-score of 55%.

The **rhetorical function** of a segment is classified in the argumentative zoning approaches [Teufel and Moens, 2002], where certain coarse-grained patterns of argumentation in scholarly papers can be captured (see also Chapter 2.2.1). Teufel and Moens for example implemented classification using a naive-Bayes approach. The performance differs widely between the zones: F-measure ranges from 26% for Contrast to 86% for Own. One zone that is of relevance to argument mining is that of criticism/contrast sentences where a precision of 57% and recall of 42% was achieved. Various features have been used, which can be taken as inspiration also for the purposes of argumentation mining: These include position and length of sentences, cue phrases and formulaic expressions, morpho-syntactic features (voice, mood, tense, modality), verb classes, and contextual features. A maximum-entropy based model for end-to-end zoning was presented by Teufel and Kan [2011]. In a similar task, Liakata et al. [2012] used SVMs and CRF-based models. Guo et al. [2013] improved over such feature rich approaches further by including declarative expert knowledge as additional constraints on the prediction.

Park and Cardie [2014] focused on supporting segments in a corpus of user comments to policy proposals, and classified which **type of evidence** is presented in the segment. Using multiclass SVMs they achieved a macro-averaged F1-score of 69% for three classes. They contrast an n-gram baseline with a carefully selected feature set compromising POS-tags, sentiment counts, speech-verbs, imperative, emotional expressions, tense, and person. An extension of this work is [Park et al., 2015], where the problem is tackled as a sequence labelling task using CRFs. The results indicate that CRFs perform worse than the prior

SVM model, but can be extended with posterior constraints to yield comparable results in a semi-supervised setting with only 75% of the training data.

**Stance** classification might also be of interest for us, although most work typically focused on identifying the pro or contra stance of the whole document to a prompt [Hasan and Ng, 2013, Faulkner, 2014, Persing and Ng, 2016b]. Instead, the stance of a discourse segment towards the central claim of the text could be interesting in order to identify possible objections and thus the dialectical role of an argumentative unit.

Looking beyond the argumentation mining literature, elaborate approaches to **subjectivity** analysis are also relevant to us, as found in the *appraisal theory* of Martin and White [2005], whose multi-dimensional analysis also covers a speaker's consideration of conflicting standpoints. Appraisal is a very comprehensive scheme that is difficult to annotate [Read and Carroll, 2012a]; thus its automatic classification is hard, as experiments by Read and Carroll [2012b] showed. The task of identifying the dialectical role addressed here can be considered a subproblem of appraisal analysis.

A distinction between claims 'for' and 'against' the central claim could also be modelled on the corpus of student essays used by Stab and Gurevych [2014b]. Of the paragraph claims, 365 are marked as 'for' and 64 as 'against'. The authors do not report numbers on the stance of premises. These could be easily inferred by the sequence of supposing and attacking relations, though. To our knowledge, no work on classifying the dialectical role of the segments in this corpus has been published yet.

### 6.1.4 Argument identification and classification

In some research, argument identification and segment type classification have been married in a single classification task.

Rooney et al. [2012] addressed argument identification and segment type classification in one task, working on the AraucariaDB dataset. Their model predicts whether a sentence is non-argumentative, serves as a premise, as the final conclusion, or intermediary as both premise and conclusion. They proposed a SVM with a sequence kernel to model context. They related their results to [Mochales Palau and Moens, 2009], but the comparison is difficult to establish: For segment type classification Mochales and Moens worked on the ECHR corpus, not on AraucariaDB. For argument identification the results are not projected to the binary distinction of argumentative or not. Rooney et al. reported an overall accuracy of 65% for the whole task, but the class-wise accuracies are quite low for the non-argumentative class with only 40% (which is nevertheless balanced against argumentative instances). Scores for conclusions are even below this with around 30%, but here the label distribution is very skewed.

Ong et al. [2014] presented a simple rule-based algorithm developed for student essays. Their task included the recognition of sentence types (CurrentStudy, Hypothesis, Claim,

Citation, Support, and Oppose). Also, no type can be assigned, being somewhat equivalent to non-argumentative. The authors used eight hand-coded rules performing string matching using connective lexicons, word lists and citation patterns. It remains to be shown how well these hard-coded features will generalise to other text genres. In the lack of annotated data, the procedure was not intrinsically evaluated, but was instead used as an extra feature for essay scoring.

Stab and Gurevych [2014a] also combined ADU identification and classification in one multi-class task when modelling argumentation in student essays. They classified segments of clause or sentence size into non-argumentative, premises, claims (of the paragraph), and major claims (of the text). The macro average F-score for all classes is 73%, the F-score for the claim 54% and for the major claim 63%. Non-argumentative segments are quite well detected with an F-score of 88%. Nguyen and Litman [2015] improved over these results by replacing n-gram and unlexicalised grammar rule features with those restricted to a domain-specific and an argumentation-specific vocabulary, which has been extract from LDA topics over a larger collection of similar essays. Nguyen and Litman [2016], finally, proposed additional features: word overlap with preceding sentences and the essay title, indicators for comparative and superlative adverbs, plural first person pronouns, and lightweight discourse relations tags. They obtained further improvements in a domain-independent, cross-topic validation. Liebeck et al. [2016] had a similar multi-class task setup, but worked on a German dataset of comments from an online participation project.

Goudas et al. [2015] used a two-stage strategy to identify argumentative units in a corpus of web text on renewable energy sources in Greek language: Sentences are classified as being argumentative or not using (amongst others) logistic regression models. They proposed several new features, such as counts of adjectives, of relevant named entities in the previous sentences, as well as probabilities of words and POS-tags from only argumentative or only non-argumentative language models. A CRF model then predicted the token-spans of premises, expressed in sentences which have been marked as argumentative, while claims are reported to be typically left implicit in these texts, only signalled as positive or negative attitudes towards entities. In order to improve this fine-grained CRF model, Sardianos et al. [2015] proposed to use distributed word representations.

Habernal and Gurevych [2015] classified sentences into the **roles of a modified Toulmin scheme** (Backing, Claim, Premise, Rebuttal, Refutation) or non-argumentative. The token-span annotations were first projected to the sentence level. A sequence model then predicted begin- and inside-markers for all five roles as well as the outside marker on the sentence level. The predictions were then mapped back to the token level begin/inside/outside encoding for evaluation. In addition to a rich feature set, the authors proposed to use the distance between input sentences and clusters of vector space representations, which were derived from unlabelled texts from debate portals in an unsupervised fashion. These prove to be helpful across different genres and different domains.

### 6.1.5 Relation identification

The first approach for identifying relations between argumentative segments from natural language input was presented by Mochales Palau and Moens [2009]. Although they used machine-learning classifiers for the tasks of argument identification and segment type classification, their approach to connect the clauses of a text from the ECHR corpus in a full argumentation structure is inspired rather by a symbolic, grammar-oriented view. They defined a CFG that consumes important cue words and words of arbitrary clauses, identifies the non-terminal constituents as premises or conclusions, and connects them in a (recursive) tree-structure spanning over the whole document, which is intended to represent the argumentation structure. The authors reported an accuracy of 60% in detecting these structures.

Only recently, data-driven approaches have been applied. Lawrence et al. [2014] constructed tree structures on philosophical texts using unsupervised methods based on topical distance between the segments. The relations in the tree are neither labelled nor directed. Unfortunately, the method was evaluated only on a few annotated items.

Finally, [Stab and Gurevych, 2014a] presented a supervised data-driven approach for relation identification. They predicted attachment for pairs of argumentative segments of a student essay's paragraph in order to determine whether they should be connected in a potential argumentation structure. They obtained a macro F1 score of 72%, and a F1 score of 52% for positive attachment. Note that only pair-wise decisions are made, no overall structure is predicted, and no decoding is used to optimise global predictions per paragraph or text. Nguyen and Litman [2016] improved over these results, using (again) the LDA-based argumentative/domain vocabularies and through non-overlapping context windows over the source and the target segments.

### 6.1.6 Relation classification

One study explicitly on classifying argumentative relations was reported on by Feng and Hirst [2011]. They classified pairs of premise and conclusion from the newswire section of AraucariaDB text into a set of five frequently used argumentation schemes [Walton et al., 2008]: Argument from example, from cause to effect, from consequences, from verbal classification, and Practical Reasoning. Their model used some general features modelling the absolute and relative position of premises and conclusion, the ratio of length of premise and conclusion, and the number of premises. In addition, the type of argumentation structure – whether it is linked or convergent – needs to be supplied to the classifier. Also, scheme-specific features are added, which consist mainly of particular cue phrases, and a few other measures. In one-against-others classification, the system yielded best average accuracies of over 90% for two schemes, while for the other three schemes the results were between 63% and 70%.

Nguyen and Litman [2016] also modelled the more coarse-grained distinction between supporting and attacking relations for a pair of source and target segment on the student essay corpus [Stab and Gurevych, 2014b]. Using two feature sets (topic words and window context, described above) they achieved best results and improved over a baseline with word-pairs and syntactic production rules.

### 6.1.7 Approaches of argumentation structure recognition

We will now focus on related work that combines most if not all of the above presented tasks of argumentation mining, but at least ADU classification and relation identification. We will also briefly highlight our own contributions that have already been published, to outline the chronology in this fast moving field. They will be presented in more detail later in this chapter. A summary of the related work presented in this subsection is given in Table 6.1.

The first model that predicted full argumentation structures on text is certainly the context free grammar (CFG) model of Mochales and Moens [2008], Mochales Palau and Moens [2009]. Many aspects that have been described above as different subtasks of argumentation mining are tackled in their work by one grammar. The predicted trees have constituents labelled as decisions, premises and conclusion, or as non-argumentative. Constituents are of sentence-size, although determining the constituent type requires the recognition of cue words and phrases at token level. The tree structure represents argumentative relations between premises and conclusions, including serial and convergent structures. Relation types are not represented in the tree, although there is at least indirectly a distinction made between premises that come with a supporting cue word and those that exhibit a contrastive cue word.

A technically more powerful grammar-based approach was taken by Saint-Dizier [2012]: On the basis of a logic-based grammar language, which was designed to extend linguistic analysis from sentence to discourse level, the system can determine the boundaries of various types of spans and embed them to form nested discourse structures in a constraint-driven way and also handle dislocated structures. Experiments with different argumentation genres and phenomena have been conducted, yielding promising results. In [Saint-Dizier, 2012], warnings and their justifications, as well as advices and their justifications have been predicted in instructional text. Garcia Villalba and Saint-Dizier [2012] explored evaluative expressions in product reviews and their connections to coherence relations which could be interpreted as argumentative support. Kang and Saint-Dizier [2014] mined requirements in instructional text and developed the idea of an 'argument component', which is a cluster of argumentative active segments around one main claim in otherwise oftentimes largely non-argumentative text [see also Saint-Dizier and Kang, 2016]. While

| approach | text genre | input | tasks/architecture | | | | | full structure | end-to-end |
|---|---|---|---|---|---|---|---|---|---|
| | | | BD | AI | AC | RI | RC | | |
| Mochales Palau and Moens [2009] | court decisions | sentence | | | | | | ✓ | |
| Kang and Saint-Dizier [2014] | instructional | token | | | | | | ✓ | |
| Lawrence et al. [2014] | phil. text | token | | | | | | (✓) | |
| Peldszus [2014] | microtexts | ADU | | | | | | | |
| Stab and Gurevych [2014a] | student essays | ADU | | | | | | | |
| Lawrence and Reed [2015] | AIFdb | token | | | | | | (✓) | |
| Peldszus and Stede [2015a] | microtexts | ADU | | | | | | ✓ | |
| Stab and Gurevych [2016] | essays/micro | token | | | | | | ✓ | |
| Persing and Ng [2016a] | student essays | clause | | | | | | ✓ | ✓ |

Table 6.1: Overview of related work: For each approach, we list the text genre, the expected input, a summarising diagram of the architecture that shows which tasks have been tackled, whether the approach is able to predict full argumentation structure, and whether it has been evaluated end-to-end. The abbreviations of the tasks are: **B**oundary **D**etection, **A**DU **I**dentification, **A**DU **C**lassification, **R**elation **I**dentification, **R**elation **C**lassification. The architecture diagram depicts fully tackled tasks as • and tasks partially tackled as ∘. The following types of interaction are considered: •→• output of left is piped as input to right; •—• left and right are modelled as one task, e.g. in a joint classifier; ⟦•  •⟧ the output of left and right is combined according to some heuristic; [•  •] the output of left and right is globally optimised.

technically capable, this model has not yet been applied to derive full document argument structures (which is why the approach is shown grey in the table).

Our own earlier work [Peldszus, 2014], which will be presented in Section 6.3, included several types of ADU classification, relation identification, and relation classification. These were also combined to a single joint classifier. The approach assumed identified ADUs as input. [Stab and Gurevych, 2014a] worked on ADU identification and classification, as well as relation identification – relation type classification was not done as the authors considered only supporting structures. Both approaches have in common that they are data-driven and classify the units independently without combining them into a larger structure.

Although the approach of Lawrence et al. [2014] does not involve ADU classification, we still want to mention it here for its innovative way of identifying ADUs as a result of relation identification. Note, though, that the predicted structures are neither labelled nor directed, since the employed topic similarity criterion cannot decide about a direction. This is why we only depict the relation identification task as 'partially' tackled in the overview table. This approach was later extended in [Lawrence and Reed, 2015] to combine three strategies to predict argumentation structures. In a first step, argumentative discourse indicators were used to predict support and attack relations between adjacent segments with high precision but low recall. Second, argumentative schemes were identified: When all required segment-types involved in one of two of the considered schemes were predicted, they were connected according to the scheme. Finally, the similarity-based structure-building (this time using WordNet similarity, not LDA topics) was used to (undirectedly) connect segments which have not been connected in the first two steps. Evaluated on a small fragment of AIFdb, this combination of strategies was shown to yield improved results compared against each strategy alone. The evaluation was carried out on gold segments. This model can be seen as a combination of various overlapping approaches to partly predict segment types, relations, and relation types. The authors furthermore propose to integrate boundary detection and argument identification in the same manner as in their prior work.

Our subsequent work [Peldszus and Stede, 2015a], which will be presented in Section 6.5, introduced the "evidence graph" model, which jointly predicts aspects of argumentation structures for all tasks except argument identification and derives one globally optimal, full argumentation structure. This was achieved by combining the predictions of multiple local models into one graph and using the Minimum Spanning Tree (MST) algorithm to decode the optimal structure. The joint prediction also had a positive impact on the classification evaluated on separate levels.

Stab and Gurevych [2016] also proposed a model that jointly models multiple tasks. In their case, the predictions of a base classifier for segment type (major claim, claims and premises) and for relation identification were combined in an Integer Linear Program (ILP) together with constraints on possible solutions, which is then solved to yield the optimal structure. The classification of the relation types support and attack (in their terms 'stance')

was not covered in the joint modelling, contrary to our model, but can be regarded as a following step in a pipeline. The authors also offered a CRF model for ADU identification based on token-spans which yielded promising results. The structure prediction was evaluated on gold segments, though. Their approach has primarily been developed on the corpus of persuasive essays and evaluation results have been reported for the second release of this corpus. Furthermore, they report results on our microtext corpus, were they improved in relation classification, but not in ADU classification and relation identification.

An end-to-end argumentation mining model was presented by Persing and Ng [2016a]: Contrary to prior approaches, they evaluated the predictions of a full run through their models instead of evaluating each task independently with gold input. They worked with the first release of the student essay corpus [Stab and Gurevych, 2014b] and compared a pipeline system architecture with one using an ILP decoder for joint structure prediction. Their setup of base classifiers was as follows: A simple rule-based segmenter extracts clauses from automatic parse trees. Similar to several approaches above, the authors combine ADU identification and classification in one task (assigning either major claim, claim, premise or non-argumentative to each clause). Furthermore, relation identification and type classification is combined in one task in an innovative way: The classifier assigns for a pair of ordered segments one of five classes (forward-directed support, backward-directed support, forward-directed attack, backward-directed attack or no relation). Instead of generating all combinatorial possible candidate pairs, these are restricted heuristically, thereby reducing the task complexity in a reasonable way. The classifiers are maximum entropy models. While the pipeline system only used the 1-best prediction, the ILP decoder made use of the full probability distribution to satisfy the constraints. The authors showed that the joint decoding outperforms the pipeline model significantly.

### 6.1.8 Discourse parsing

Although the discourse structures of RST and SDRT are different from argumentation structure in important respects (see Chapter 2), it is worth revisiting the literature on automatically parsing these structures, as it is likely that techniques applied there might also be candidates for recognising argumentation structures automatically.

We will, in the interest of space, not elaborate on advances in shallow discourse parsing here. Although there is extensive work on recognising local coherence relations, e.g. in resources such as the PDTB, and especially since it has become a shared task in the CoNLL 2015 and 2016. We refer the interested reader to the results of these tasks [Xue et al., 2015, 2016]). Yet, as explained in Chapter 2.2.2, these models intentionally do not predict a global structure, contrary to our particular aim in this work.

**RST parsing**

An early approach to RST parsing was presented by Marcu [1999], who applied the technique of shift-reduce parsing and demonstrated that the sequence of parsing operations can be learned from an annotated corpus. Several variants of the reduction scheme have been explored since that early work, e.g., by Subba and Di Eugenio [2009] who did not learn the Shift operation but used it as default when no Reduce rule can be applied with a sufficiently high confidence.

Another bottom-up search strategy was proposed by Hernault et al. [2010]. They used two SVM classifiers: One pair-wise classifier to predict the probability of two EDUs being in relation, and one multiclass relation labeller. The structure is built in a greedy fashion, by first predicting most likely relations, then labelling it, and continuing to grow the existing partial tree by consecutive attachments to the partial tree until all EDUs are connected.

Feng and Hirst [2012] improved over these results by extending the features to furthermore include segment pair features of syntax to cover syntactic parallelism, contextual features of (gold) surrounding discourse relations, discourse production rules, WordNet-based similarity, and cue phrases. They evaluated a model trained only on intra-sentential instances, one trained exclusively on inter-sentential instances, as well as a hybrid model trained on full documents.

Two different ways of combining the predictions of inter- and intra-sentential models have been presented by Joty et al. [2013, 2015], in their case for CRF classifiers that jointly model structure and relation labels. The first approach simply assumes that every sentence corresponds to a well-formed discourse sub-tree, as it is the case mostly, and all subtrees are then combined in a full discourse tree. The other approach builds on the observation that when a sentence does not correspond to a well-formed sub-tree in its own right, they are usually part of the tree of the adjacent sentence. Consequently, the second approach tries to find the preferred attachment direction of a sentence by comparing the overlapping predictions between a sliding window of two sentences. The final structures are derived using an optimal (thus non-greedy) CKY-like bottom-up algorithm. Both approaches of combining the models were superior to previous work, with a preference for the second approach.

While optimal parsing algorithms such as the just mentioned are time-intensive for longer texts and are even considered intractable for very long input documents, a parser that runs in linear-time was proposed by Feng and Hirst [2014]. They used separate models for structure prediction and for relation classification. One set of these models is used for intra-, another for inter-sentential instances. Additionally, after a sentence- or document-level tree is predicted, a second pass model for post-editing was applied, which has additional features based on tree-properties that would not be available in the first pass. All models are linear-chain CRFs. This greedy approach was shown to yield improvements over afore-

mentioned work and is very fast. Furthermore, the post-editing step produced significantly better structures, though at the cost of doubling the processing time, which can be still regarded acceptable.

Ji and Eisenstein [2014] presented a shift-reduce parser that learns not only the parsing actions but also a low-dimensional vector-space feature representation of the EDUs, which replaces the sparse lexical features used in prior approaches. The authors stressed the need to learn these representations directly from the source data while training the target task. Using pre-trained word-embeddings instead is not advised, as these might not be tied to the application of interest and thus might impair the parsing process. They reported improvements in nuclearity detection and relation classification.

In a similar way, Li et al. [2014a] transform EDUs to abstract feature representations. They used Recurrent Neural Networks and employed two classifiers, one for structure prediction and one for relation classification. Besides vector representations for the leaf nodes of the trees, they also recursively represented subtrees that leaf nodes attach to. A distinction between intra- and inter-sentential segments was not made. The most likely tree was determined using an CKY-like bottom-up algorithm. Given that only a few extra features were used in addition to the learned representations, the approach yielded promising, comparable results, but it did not outperform previous work.

While most of the presented work focused on parsing RST trees as constituency trees, Li et al. [2014b] converted them to dependency structures and applied dependency parsing techniques [McDonald et al., 2005a]. They tested both MST based decoding, which can handle non-projective structures (but the converted dependencies are all projective), as well as the Eisner-algorithm, which will only predict projective structures. The difference between both approaches was insignificant. For comparison with related work, the predicted dependency trees were converted back to constituency trees. The authors reported improvements in nuclearity detection and relation classification, although to a limited extent. in comparison to [Ji and Eisenstein, 2014].

#### SDRT parsing

A first approach to parse SDRT structures in dialogue transcripts and newswire text was presented by Baldridge et al. [2007]. They converted the SDRT graphs to dependency trees and applied a dependency parser [McDonald et al., 2005a] that uses MST to decode the globally optimal structure. They showed that this approach is superior to a PCFG baseline.

Muller et al. [2012a] compare different decoding approaches, a greedy mechanism, MST, and A*-search. These are tested on a French corpus of newspaper articles and Wikipedia entries. The authors investigated how linguistically oriented constraints on discourse structure (such as the right frontier constraint) could be enforced in such a decoding process. Also, they evaluated a pipeline architecture (where relations are first identified, then labelled)

and a joint approach where the predictions of the identification and labelling process are combined. It was observed that MST and A*-decoders perform best, and that the pipeline approach worked significantly better than the joint prediction when predicting fully labelled structures.

Afantenos et al. [2015] used MST decoding for predicting the (dependency converted) SDRT structures of multiparty dialogues. Similar to the distinction between intra- and inter-sentential models in RST parsing, they proposed local models for intra- and inter-turn relations. The local model is a maximum entropy classifier that predicts relation labels for a given pair of EDUs. Since the structure inside a turn is rather sequential, the best scoring method for intra-turn was not the trained classifier but a heuristic baseline which always attaches to the previous segment. For inter-turn relations, the MST decoder produced the best results. Both are combined to a global, document-level model.

Working on the same corpus, Perret et al. [2016] improved over previous results by applying an ILP-based approach, which decodes a globally optimal structure from the local models for structure and relations classification, using amongst others also domain dependent constraints on the output structure. The authors also investigated different ways of distributing relations when converting the hyper-graph SDRT structures to (directed acyclic) dependency graphs. Their system can be considered the first discourse parser that is not restricted to tree-like structures.

Looking back over the related work from discourse parsing, we find several commonalities but also some differences with the field of argumentation structure parsing. Faced with the same problem of predicting a structure that is supposed to fulfil certain constraints, most of the work defines local models whose predictions are then combined to derive a globally valid structure. Consequently, there is a choice to be made concerning the decoders, e.g. between optimal and greedy ones. Other work employs structured learning. Also, although discourse parsing can be considered an established field, there are quite different strategies to operationalise the models. The distinction between intra- and inter-sentence, which is sometimes made for rhetoric parsing, is not yet an issue with argumentation mining, and it is not clear yet whether it would pay off to make this decision. The move from very rich and sparse feature sets to more general vector-space representations has certainly been put on the agenda in RST parsing, but is –with notable exceptions– still in its beginnings in the field of argumentation mining. We will have to keep these considerations in mind, when developing our approach to recover full argumentation structures.

## 6.2 Methodology

All approaches of automatic prediction proposed in this chapter are data-driven, supervised machine-learning models. We will train classifiers on different aspects of the argumentation

structure: Similar to the different label sets used for evaluating the annotation experiments (see Section 4.2.1), we will extract different label sets from the argumentation structure for different subtasks of argumentation mining.

When using the microtext corpus in machine learning experiments, we face two problems. First, the corpus is rather small. This not only restricts the possible selection of machine learning approaches: Some approaches (as for example deep learning architectures) require much larger datasets in order to converge, and are thus not applicable in our case. It also has implications for the experimental setup to reach conclusive results. Second, argumentative configurations and patterns are not equally distributed in our corpus: For example, there are more supporting than attacking relations or more proponent than opponent segments, etc. The distribution of classes a classifier has to learn will be very skewed. This requires a lot of care when setting up the classifier and evaluating the results.

For a large dataset, it is a reasonable practice to divide it into a fixed **training, development, and test set** before all experimentation, e.g. with a 70/10/20% proportion. The feature development as well as potential hyper-parameter tuning can be done on the development set, the training set is sufficiently large, and the test set can remain a true blind test set, whose instances are never looked at. However, this practice is not applicable in this work for two reasons: First, the experimenter was involved in the annotation of the corpus and can thus not be assumed not to know the instances of the test set. More pressing, though, is the small size of the corpus: A 20% sample for testing or a 10% sample for development is likely to miss less frequent but still characteristic phenomena of the corpus (e.g. relation types, structural configurations, or cues). In other words, the corpus size is too small to ensure a representative 20% sample. We therefore rely on **k-fold cross-validation** (CV) to get an estimate of the model's accuracy on all instances of the dataset [Stone, 1974, Hastie et al., 2013].

For imbalanced class distributions, the sampling of the instances can be **stratified**, i.e. the training and the test sample have a similar class distribution. This way, we can avoid the undesirable situation where less frequent classes are not represented in the test set and no score could be computed. Furthermore, cross-validation can be **repeated** with randomly different foldings of the data. This allows us to have a larger sample of testing scores, which in turn leads to better approximations when assessing statistical significance in comparing different experiment conditions.

In our setting, the ultimate goal is to predict the argumentative structure of a whole paragraph or text. A stratified folding is likely to split up the classification items of one paragraph or text across multiple folds in order to maximise the similarity of the individual fold's category distributions, with the caveat that we might not predict and evaluate a connected argumentation structure in one fold. To overcome this, we propose to use a **group-wise** stratification: Classification items of one group (in our case of one text) cannot be separated in stratification. Instead, the group-wise folding seeks to find sets of groups with similar

class distributions. In all experimental settings that aim for predicting full argumentation structures (Sections 6.5, 6.6, 6.7 and 6.8), we will use repeated, group-wise, stratified cross-validation. To allow reproducibility, we have made the resulting train-test-splits publicly available. In all other experiments, we use repeated, stratified cross-validation only.

For testing the **statistical significance** of a measured difference between two models for model selection, we use the Wilcoxon signed-rank test [Wilcoxon, 1945]. As advocated by Demšar [2006], it is less sensitive to outliers and does not assume a normal distribution as the standard paired t-test. We assume a significance level of $\alpha$=0.01.

Since we cannot **tune the hyperparameters** of the models on a designated development set, we determine the best parameter through an 'inner' cross-validation on the training set. This is also referred to as 'nested cross-validation' or 'double cross' [Stone, 1974]. Assume for example a nested 5×4 CV. The outer 5-fold CV, where the classifier is trained on 4/5 and tested on 1/5 of the dataset, is for evaluation only. The inner 4-fold CV on the training set of the outer CV is used to find the optimal choice of hyperparameters. Once the best hyperparameters are found, a model is trained on all training set items with these parameters and then tested on the test set. This procedure has the advantage that the performance evaluation is less biased by external factors such as sampling [see Cawley and Talbot, 2010], but it evidently comes with a higher computational cost. An experimental protocol that repeats these nested cross-validations has been presented by Filzmoser et al. [2009].

For measuring the performance of our classifiers, we use the following metrics:

- **Accuracy** is the proportion of correct decisions over all decisions. In a multi-class evaluation it is highly biased by the distribution of the class labels.
- The **F-measure** is the harmonic mean of precision and recall. It focuses only on one class and takes into account true and false positives and false negatives, but does not consider true negatives.[1] We will use this metric for reporting individual class-wise results.
- In a multi-class setting, even in a dichotomous setting where both classes are of interest, one way to derive a single score is to average the F-scores of the classes. While **micro-averages** give equal weight to each classification decision and are thus biased towards the performance of frequent classes, the **macro-averaged** F-scores weigh all classes equally and are thus helpful for determining the performance of the infrequent classes [Manning et al., 2008].
- Finally, we will again use the **kappa** metric, such as Cohen's $\kappa$ (see Section 4.1) as a chance-corrected metric. It will allow us to compare the system's performance against that of human annotators and opens up a useful continuum between 0.0 and 1.0 for

[1] This is considered a reasonable simplification in the field of information retrieval where this metric originates from, because only the relevant documents count for information retrieval.

interpreting the system's performance: A majority-baseline will score with $\kappa$=0.0, since it relies only on chance agreement, and thus represents the worst noteworthy baseline in a multiclass setting. Perfect agreement results in $\kappa$=1.0.

Since all reported scores are the result of repeated CVs, they are averages. Note that we average over the scores of the train-test splits of all repetitions of CV (not over the already averaged cross-validation scores). As an example, 10 iterations of 5-fold CV result in 50 different train-test splits and the reported score will be the average score of these 50 splits.

## 6.3 Study 1: Local models of aspects of argumentation structure

Deriving a full argumentation structure from text is a challenging task: All the subtasks involved have their own difficulties. Deriving the full structure requires sufficiently accurate predictions in these subtask in order to be successful, otherwise error propagation might derogate the results.

The goal of the study presented in this section is to start small: Instead of aiming for full structure prediction, we will investigate how well local models can capture certain aspects of the argumentation structure. This will provide us with an understanding of the difficulty of the various tasks involved in argumentation mining. For this purpose, we will extract various datasets from the corpus of argumentation structures, e.g. one for distinguishing proponent and opponent segments, one for the distinction between support and attack, and even for argumentative relation.

The models tested here are *local* models in the sense that they make their prediction without considering a larger context or even an optimal assignment given the whole text. They will consider only the target text segment itself and its next neighbours.

Our aim here is on the one hand to evaluate features that we found in the discourse parsing, argumentative zoning, and text mining literature in general, and on the other hand to test different machine learning algorithms in a rather out-of-the-box setting, without diving deeper into the individual algorithm's peculiarities of optimising hyper-parameters.

In the following we will first describe our experimental setup, the extraction of the datasets for each aspect of the argumentation structure, the features and the classifiers used, and then compare and discuss the results of the classifiers and the features.

### 6.3.1 Experimental setup

#### Data

As our source data we use the corpus of microtexts presented in Chapter 5, the German version of the original non-transformed **microtext-graph-ADU** form. As the experiment reported here was done shortly before the corpus was finalised and published, it uses a

prior, but near-final version of the corpus. In comparison to the final version, the corpus used here contains three more microtexts that have been discarded later for not fulfilling the required constraints (see Section 5.2.1), and misses only a few corrections of the annotated argumentation structures. We expect only minimal differences, had this experiment been done on the final version of the corpus, and thus regard the results serving the goals of this study.

#### Task

Similar to the evaluation of the annotation study in Section 4.2.1, we extract different segment-wise label sets from the argumentation graphs, each describing one or more aspects of the structure. This gives an overview over the basic label sets and their corresponding classes:

- **role**: Is the segment of the proponent (P) or opponent role (O)?
- **function***: Does the segment present the central claim/thesis (T) of the text, or does it support (S) or attack (A) another segment? This is the reduced set of argumentative functions.
- **function**: Does the segment present the central claim/thesis (T) of the text, does it support another segment normally (SN) or by example (SE), or does it attack another segment as a rebutter (AR) or as an undercutter (AU)?
- **comb**: Does the segment's function hold only in combination with that of another segment (C) or does is stand alone (S)?
- **target**: How far away is the target of the relation away, relative to the position of the source, encoded as an offset *-x...0...+x*. The prefix 'n' signals that the proposition of the node itself is the target, while the prefix 'r' signals that the relation coming from the node is the target.

We not only extract basic label sets such as argumentative role or function, but also complex ones that combine multiple basic label sets, such as e.g. role+function or even the full label set role+function+comb+target that covers all information encoded in an argument graph. Figure 6.1 shows the extraction of these full labels for an example text.

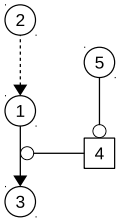
#### Features

All (unsegmented) texts have been automatically split into sentences and been tokenised by the OpenNLP-tools[2]. The mate-parser pipeline then processed the tokenised input, yielding lemmatisation, POS-tags, word-morphology and dependency parses [Bohnet, 2010]. The annotated gold-standard ADU segmentation in the dataset was then mapped to the

[2]Apache OpenNLP version 1.5.3, `https://opennlp.apache.org/`

[*Energy saving light bulbs contain a significant amount of toxins.*]$_1$ [*A commercially available bulb may contain for example up to five milligrams of mercury.*]$_2$ [*That's why they should be taken off the market,*]$_3$ [*unless they're unbreakable.*]$_4$ [*But precisely this is unfortunately not the case.*]$_5$

(a)




(b)

| node id | rel. id | full label | target |
|---|---|---|---|
| 1 | 1 | PSNS | (n+2) |
| 2 | 2 | PSES | (n-1) |
| 3 | 3 | PT | (0) |
| 4 | 4 | OAUS | (r-3) |
| 5 | 5 | PARS | (n-1) |

(c)

Figure 6.1: Extracted labels for an example text (micro_d21): the (here English) text segmented in ADUs given in (a), the argumentation structure graph in (b), the segment-based labelling representation in (c).

automatic sentence-splitting/tokenisation, in order to be able to extract exactly those linguistic features present in the gold-segments. Using this linguistic output and several other resources, we extracted the following features:

- **Lemma Unigrams**: We add a set of binary features for every lemma found in the present segment. Likewise, we extract these in the preceding and the subsequent segment respectively as a separate feature set, in order to represent the segment's context in a small window.
- **Lemma Bigrams**: We extracted lemma bigrams of the present segment.
- **POS Tags**: We add a set of binary features for every POS tag found in the present, preceding and subsequent segment.
- **Main verb morphology**: We added binary features for tempus and mood of the segment's main verb, as subjunctive mood might indicate anticipated objections and tempus might help to identify the main claim.
- **Dependency triples**: The dependency parses were used to extract features representing dependency triples (relation, head, dependent) for each token of the present segment. Two features sets were built, one with lemma representations, the other with POS tag representations of head and dependent.
- **Sentiment**: We calculate the sentiment value of the current segment by summing the values of all lemmata marked as positive or negative in SentiWS [Remus et al., 2010].[3]

[3]We are aware that this summation is a rather trivial and potentially error-prone way of deriving an overall sentiment value from the individual values of the tokens, but postpone the use of more sophisticated methods to future work.

- **Discourse markers**: For every lemma in the segment that is listed as potentially signalling a causal or contrastive discourse relation (cause, concession, contrast, asymmetric contrast) in a lexicon of German discourse markers [Stede, 2002] we add a binary feature representing the occurrence of the marker, and one representing the occurrence of the relation. Note that these are candidate markers, which are neither manually nor automatically disambiguated. Again, discourse marker / relations in the preceding and subsequent segment are registered in separate features.
- **First three lemmata**: In order to capture sentence-initial expressions that might indicate argumentative moves, but are not strictly defined as discourse markers, we add binary features representing the occurrence of the first three lemmata.
- **Negation marker presence**: We use a list of 76 German negation markers derived in Warzecha [2013] containing both closed class negation operators (negation particles, quantifiers, and adverbials etc.) and open class negation operators (nouns like "denial" or verbs like "refuse") to detect negation in the segment.
- **Segment position**: The (relative) position of the segment in the text might be helpful to identify typical linearisation strategies of argumentation.

In total a number of approx. 19,000 features has been extracted. The largest chunks are bigrams and lemma-based dependencies with ca. 6,000 features each. Each set of lemma unigrams (for the present, preceding, and subsequent segment) has around 2,000 features. These features are used for all classification tasks, i.e. for all the label sets extracted from the argumentation structures.

### 6.3.2 Models

We compare classifiers that have frequently been used in related work, particularly in the related field of argumentative zoning: Naïve Bayes (NB) approaches as in Teufel and Moens [2002], Support Vector Machines (SVM) and Conditional Random Fields (CRF) as in Liakata et al. [2012] and maximum entropy (MaxEnt) approaches as in Guo et al. [2013] or Teufel and Kan [2011]. We used the Weka data mining software, v.3.7.10, [Hall et al., 2009] for all approaches, except MaxEnt and CRF.

- **Majority**: This classifier assigns the most frequent class to each item. We use it as a lower bound of performance. The used implementation is Weka's ZeroR.
- **One Rule**: A simple but effective baseline is the one rule classification approach. It selects and uses the one feature whose values can describe the class majority with the smallest error rate. The used implementation is Weka's OneR with standard parameters.

- **Naïve Bayes**: We chose to apply a feature selected Naïve Bayes classifier to better cope with the large and partially redundant feature set.[4] When fitting the model, all features are first ranked according to their information gain observed on the training set. Features with information gain $\leq 0$ are excluded.
- **SVM**: For SVMs, we used Weka's wrapper to LibLinear [Fan et al., 2008] with the Crammer and Singer SVM type and standard wrapper parameters.
- **MaxEnt**: The maximum entropy classifiers are trained and tested with the MaxEnt toolkit [Zhang, 2004]. We used at maximum 50 iterations of L-BFGS parameter estimation without a Gaussian prior.
- **CRF**: For the implementation of CRFs we chose Mallet [McCallum, 2002]. We used the SimpleTagger interface with standard parameters.

Non-binary features have been binarized for the MaxEnt and CRF classifiers.

### 6.3.3 Results

All results presented in this section have been produced in 10 repetitions of 10-fold cross validation. No hyper-parameter optimisation was carried out. We report A(ccuracy), micro-averaged F(1-score) as a class-frequency weighted measure and Cohen's $\kappa$ as a measure focusing on less frequent classes. All scores are given in percentages.

#### Comparing classifiers

A comparison of the different classifiers is shown in Table 6.2; bold values indicate highest average.[5] To begin with, the majority and the one rule classifiers serve as our baseline. Due to the skewed label distribution, accuracy and the micro-averaged F1-scores are already at a high level, especially for the 'role' and 'comb'-level. Note that the agreement between predicted and gold for the majority classifier is equivalent to chance agreement and thus yields $\kappa$=0 on every level, even though there are F-scores near 0.70.

The Naïve Bayes classifier profits from the feature selection on levels with a small number of labels and gives best results for the 'function*' and 'role+function*' levels. On the most complex level with 48 possible labels, however, performance drops even below the OneR baseline, because features do not reach the information gain threshold. The MaxEnt classifier performs well on the 'role' and 'comb', as well as on the 'role+function' levels. It reaches

[4] With feature selection, we experienced better scores with the Naïve Bayes classifier, the only exception being the most complex level 'role+function+comb+target', where only very few features reached the information gain threshold.

[5] We also report standard deviations. The deviations can be considered as relatively high in comparison to the averages. We attribute this to the small size of the dataset, which may lead to infrequent classes being underrepresented in train or test folds. This has an even stronger effect on metrics that are sensitive to the performance of infrequent classes, which is why the $\kappa$-scores exhibit higher deviations than accuracy or F1-score values.

| level | Majority | | | OneR | | | CRF | | |
|---|---|---|---|---|---|---|---|---|---|
| | A | F | $\kappa$ | A | F | $\kappa$ | A | F | $\kappa$ |
| role | 78±1 | 69±1 | 0±0 | 83±3 | 79±4 | 33±13 | 86±5 | 84±6 | 49±16 |
| function* | 49±1 | 33±1 | 0±0 | 58±3 | 47±3 | 23±7 | 68±7 | 67±8 | 46±12 |
| function | 48±1 | 31±1 | 0±0 | 56±3 | 45±3 | 22±6 | 62±7 | 58±8 | 38±10 |
| comb | 74±1 | 62±1 | 0±0 | 81±4 | 77±4 | 44±12 | **84**±5 | **81**±7 | 55±13 |
| target | 24±1 | 9±1 | 0±0 | 37±5 | 29±4 | 24±6 | 47±11 | **45**±11 | **38**±12 |
| role+function* | 47±1 | 30±1 | 0±0 | 56±3 | 45±3 | 22±6 | 67±7 | 65±8 | 49±11 |
| role+function | 46±1 | 29±1 | 0±0 | 54±3 | 43±3 | 21±6 | 61±7 | 56±8 | 38±11 |
| role+function+comb | 41±1 | 24±1 | 0±0 | 50±4 | 38±3 | 19±6 | 56±7 | 51±8 | 36±9 |
| role+function+comb+target | 20±1 | 7±1 | 0±0 | 28±4 | 19±3 | 18±5 | 36±10 | 30±9 | 28±10 |
| level | Naïve Bayes | | | MaxEnt | | | LibLinear | | |
| | A | F | $\kappa$ | A | F | $\kappa$ | A | F | $\kappa$ |
| role | 84±5 | 84±5 | **52**±14 | **86**±4 | **85**±5 | **52**±15 | **86**±4 | 84±4 | 50±14 |
| function* | **74**±5 | **74**±5 | **57**±8 | 70±6 | 70±6 | 51±10 | 71±5 | 71±5 | 53±9 |
| function | **68**±5 | **67**±5 | **52**±8 | 63±6 | 62±6 | 43±9 | 65±6 | 62±6 | 44±9 |
| comb | 74±6 | 75±5 | 42±11 | **84**±5 | **81**±7 | **56**±12 | **84**±3 | **81**±4 | 54±10 |
| target | 38±6 | 38±6 | 29±6 | 47±8 | 44±8 | 37±9 | **48**±5 | 44±5 | **38**±6 |
| role+function* | **69**±6 | **69**±6 | **55**±9 | 68±7 | 67±7 | 51±10 | **69**±5 | 67±6 | 52±9 |
| role+function | 61±5 | **61**±5 | **45**±7 | 63±6 | **61**±6 | **45**±9 | **64**±5 | 60±5 | **45**±8 |
| role+function+comb | 53±6 | 51±6 | 36±8 | 58±6 | 54±7 | 41±8 | 61±5 | 56±5 | **44**±8 |
| role+function+comb+target | 22±4 | 19±4 | 16±4 | 36±6 | **33**±6 | 29±6 | **39**±5 | 32±4 | **31**±5 |

Table 6.2: Classifier performance comparison: Percent average and standard deviation in 10 repetitions of 10-fold cross-validation of A(ccuracy), micro averages of F1-scores, and Cohen's $\kappa$.

the highest F-score on the most complex level. The SVM generally performs well in terms of accuracy and specifically on the most interesting levels for future applications, namely in target identification and the complex 'role+function' and 'role+function+comb+target' levels. For the CRF classifier, we had hoped that approaching the dataset as a sequence labelling problem would be of advantage. However, applied out of the box as done here, it did not perform as well as the segment-based MaxEnt or SVM classifier. This might be attributed to our features, which already incooperated some sequential context. Also, it is in line with the results reported by Park et al. [2015], where CRFs did not improve over SVMs in a related task.

Overall, the SVM classifier scored best out of the box, followed by the MaxEnt classifier. In the following discussion of the difficulty of the task and of our experiment with the feature sets, we will focus on these two classifiers.

#### Difficulty of the levels

Our aim in this study is to obtain a first understanding of the difficulty to model certain aspects of argumentation structure in the microtexts. For the basic levels except target, the best classifiers achieve $\kappa$>0.50, which is already good start. On the 'role' and the 'function*' level all classes are covered in the predictions with class wise F1-score between 0.60 and 0.90.

On the more fine-grained 'function' level, we have the distinction between rebutting and undercutting attack, which is not easy to draw with an F1=0.42 for rebut and F1=0.36 for undercut. While the majority class of normal support is quite reliably predicted, example support is not predicted at all, probably because there are just not enough instances of this class in the corpus.

The 'comb' level, which represents whether a segment is effective only when linked with another segment, is very problematic. Although the best classifier yields a good score with $\kappa$=0.56, it turns out that it was not able to learn to predict segments whose relation needs to be combined with others. The high score is due to a peculiarity of segment-wise extraction: The label set distinguishes between three classes, segments with combined relations, segments with single relations, and segments without a relation (i.e. central claims). The classifier only learned to distinguish the latter two classes, but did not learn to identify the rather infrequent combined relations.

For the 'target' level, where the classifiers learn the offset of attachment, the score of the best model is $\kappa$=0.38. Looking at the predicted classes, we observe that the SVM for example correctly identifies 84% of the non-attaching central claims, as well as about 60% of the adjacent targets (with better results for preceding than for subsequent targets), but only 19% of the non-adjacent targets. This result is expected, provided the way we framed the task here, and we will present an approach that is capable to guide the prediction of whether

| label | Precision | Recall | F1-score |
|---|---|---|---|
| PT | 75±12 | 74±13 | 74±11 |
| PSN | 65±8 | 79±7 | 71±6 |
| PSE | 1±6 | 1±6 | 1±6 |
| PAR | 12±29 | 12±27 | 11±24 |
| PAU | 57±26 | 49±24 | 50±22 |
| OSN | 1±12 | 1±12 | 1±12 |
| OAR | 54±18 | 42±16 | 46±13 |
| OAU | 8±27 | 7±23 | 7±23 |

Table 6.3: MaxEnt class-wise results on the 'role+function' level: Percent average and standard deviation in 10 repetitions of 10-fold cross-validation of Precision, Recall and F1-score.

an argumentative relation holds and which segment is targeted by global constraints later in this chapter (see Section 6.5).

We now turn to the complex levels, where multiple basic label sets are combined in a single task. Our aim here is to investigate, firstly, how complex the task can be made while still achieving useful results. Secondly, some properties of the argumentation structure might be easier to predict in the context of another. Consider the 'role+function*' level as an example: The 'role' level seems slightly harder to predict than the 'function*' level. When learned together as one complex task, the performance is better than the original 'role' performance. Furthermore, we use the complex classifier to predict basic levels, by reading only one basic level from the complex tag. When assigning only the argumentative role from the prediction of the 'role+function*' classifier, we get a small improvement of two points $\kappa$ over the basic role classifier. We can conclude that role classification can profit from being integrated with the function* level. Conversely, if we predict only the function* level with the combined classifier, the result deteriorates by two points $\kappa$. In terms of confusions, about half of the proponent and half of the opponent attacks are confused with the majority class proponent support.

For the 'role+function' level, we present detailed class-specific results of the MaxEnt classifier in Table 6.3. Here we have the additional complexity of distinguishing rebutting from undercutting attacks and normal from example support. We observe that the recognition of the classes highly correlates with the number of instances found in the dataset. Rebuts are more frequent for the opponent and undercuts for the proponent (see Figure 5.1) and these types can be recognised to a certain degree, while infrequent types such as opponent's undercuts and proponent's rebuts, but also example support in general and opponent's support are not identified successfully.

Adding the 'comb' level does not change very much, at least for the best performing SVM here. As we know from above, combined relations could not be recognised, and since there are only a very few of them, the results decrease only by one point $\kappa$.

The full task is more interesting. Although it is obvious from the results so far and from the scores in Table 6.2 that a task of this complexity is infeasible given the amount of data and the simplicity of the local models, the SVM still correctly predicts 39% of the segments: 17% are central claims, 22% are supporting, rebutting, and undercutting relations mostly targeting adjacent segments. In general, too many central claims and proponent supports and undercuts have been predicted, which indicates that the classifier is biased towards the proponent role.

#### Feature ablation on 'role+function' level

We performed feature ablation tests with multiple classifiers on multiple levels. For the sake of brevity, we only present the results of the SVM and MaxEnt classifiers here on the 'role+function' level. The results are shown in Table 6.4. Bold values indicate greatest impact, i.e. strongest loss in the upper leave-one-feature-out half of the table and highest gain in the lower only-one-feature half of the table.

The greatest loss is produced by leaving out the discourse marker features. We assume that this impact can be attributed to the useful abstraction of introducing the signalled discourse relation as a feature, since the markers themselves are also present in other features (such as lemma unigrams, perhaps first three lemma or even lemma dependencies) that produce only minor losses.

For the single feature runs, lemma unigrams produce the best results, followed by discourse markers and other lemma features as bigrams, first three lemma, and lemma dependencies. Note that negation markers, segment position, and sentiment perform below or equal to the majority baseline. Whether the sentiment feature can prove more useful when we apply a more sophisticated calculation of a segment's sentiment value is something that may be investigated in future work. POS-tag based features are around the OneR baseline in terms of F-score and $\kappa$, but less accurate.

Interestingly, when using the LibLinear SVM, lemma bigrams have a larger impact on the overall performance than lemma based dependency triples in both ablation tests, even for a language with a relatively free word order as German. This indicates that the costly parsing of the sentences might not be required after all. However, this difference is not as clear for the MaxEnt classifier.

### 6.3.4 Conclusions

This concludes our first exploratory study in automizing argumentation analysis on the microtext corpus. We developed classifiers for predicting several properties of the argumenta-

| Features | LibLinear | | | MaxEnt | | |
|---|---|---|---|---|---|---|
| | A | F | $\kappa$ | A | F | $\kappa$ |
| all | 64±5 | 60±5 | 45±8 | 63±6 | 61±6 | 45±9 |
| all w/o dependencies lemma | 64±5 | 60±5 | 46±8 | 62±6 | 60±6 | 44±9 |
| all w/o dependencies pos | 65±5 | 61±5 | 46±8 | 63±6 | 61±7 | 45±9 |
| all w/o discourse markers | **62**±5 | **59**±5 | **43**±8 | **61**±7 | **58**±7 | **42**±9 |
| all w/o first three lemma | 64±5 | 60±5 | 44±8 | 63±6 | 60±7 | 44±9 |
| all w/o lemma unigrams | 63±5 | 60±5 | 45±8 | 62±6 | 60±7 | 44±9 |
| all w/o lemma bigrams | 63±5 | 60±5 | 44±8 | 62±6 | 60±6 | 44±9 |
| all w/o main verb morph | 64±5 | 60±5 | 45±8 | 62±6 | 60±6 | 43±9 |
| all w/o negation marker | 64±5 | 60±6 | 45±8 | 63±6 | 61±7 | 45±9 |
| all w/o pos | 64±5 | 61±5 | 45±8 | 63±6 | 60±7 | 44±8 |
| all w/o segment position | 64±5 | 60±5 | 45±8 | 63±6 | 61±6 | 45±9 |
| all w/o sentiment | 64±5 | 60±5 | 45±8 | 62±6 | 60±6 | 44±9 |
| only dependencies lemma | 56±4 | 47±4 | 27±6 | 56±6 | 49±7 | 30±8 |
| only dependencies pos | 42±6 | 41±6 | 18±8 | 41±7 | 40±7 | 16±9 |
| only discourse markers | 56±6 | 53±6 | 34±9 | 53±6 | 52±7 | 30±10 |
| only first three lemma | 54±6 | 52±6 | 33±9 | 50±6 | 48±6 | 26±8 |
| only lemma unigrams | **59**±5 | **55**±5 | **37**±8 | **59**±6 | **56**±7 | **38**±8 |
| only lemma bigrams | **59**±4 | 53±5 | 34±8 | 55±7 | 51±7 | 30±9 |
| only main verb morph | 49±6 | 39±4 | 16±7 | 52±5 | 41±6 | 20±6 |
| only negation marker | 25±14 | 19±8 | 00±4 | 46±5 | 29±5 | 00±0 |
| only pos | 45±6 | 45±6 | 24±9 | 46±8 | 45±7 | 23±10 |
| only segment position | 31±12 | 25±10 | 04±7 | 46±5 | 29±6 | 00±0 |
| only sentiment | 22±14 | 15±11 | -1±3 | 46±5 | 29±6 | 00±0 |

Table 6.4: Feature ablation tests on the 'role+function' level: Percent average and standard deviation in 10 repetitions of 10-fold cross-validation of A(ccuracy), micro averages of F1-scores, and Cohen's $\kappa$.

tion structures. We tested different machine learning algorithms with standard parameter settings and found maximum entropy models and linear SVMs to perform best. When it comes to features, not surprisingly, bag of words features already provide a very good baseline. Furthermore, discourse markers and the signalled discourse relations are very predictive. Parsing-based features turned out to be less important than expected. Sentiment scores or the presence of negation proved to be even irrelevant for the tasks.

For the basic levels such as argumentative role and argumentative function, the proposed models already yield promising the results. A fine-grained distinction between relation types is mainly impaired by the low frequency of certain argumentative functions. We observed that the segment-wise classification approach, even though offering the context of the adjacent segments, is not able to generalise for the identification of the target of the relation (60% of the adjacent and only about 20% of the non-adjacent relations could be identified). Finally, we found that there are interactions between the different levels of analysis, which could be used to guide the prediction. We will make a more elaborate proposal for the complex task of structure prediction that uses these interactions in Section 6.5.

But before we proceed to these global models, we will first focus on the aspect of argumentative dialectics in text, on the recognition of argumentative role both in microtexts and in the more complex ProCon commentaries.

## 6.4 Study 2: Finding the opponent

The exchange of argument and objection is obviously most typical for dialogue, but to a good extent it is also present in monologue text: Authors do not only provide justifications for their own position – they can also mention potential objections and then refute or outweigh them. In this way they demonstrate to have considered the position of "the other side", which altogether is designed to reinforce their own position. The term "counter-consideration" is used here in a general sense to cover all such moves of an author, no matter whether they are directed at the conclusion of the text, or at an intermediate argument, or at some support relation, and irrespective of whether they are explicitly refuted by the author or merely mentioned and left outweighed by the mass of arguments in favour of the main claim.[6]

For an author, presenting a counter-consideration involves a switch of perspective by temporarily adopting the opposing viewpoint and then moving back to one's own. This is a move that generally requires some form of explicit linguistic marking so that the reader can follow the line of argumentation. The kinds of marking include explicit belief attribution followed by a contrastive connective signalling the return ("Some people think that X.

[6] Govier [2011] discusses the role of such counter-considerations in 'pro and con' argumentation in more depth. Also, for a comprehensive overview of different notions of objections in argument analysis, see [Walton, 2009].

However, this ..."), and there can also be quite compact mentions of objections, as in "Even though the project is expensive, we need to pursue it, because..."

Detecting counter-considerations is thus a subtask of argumentation mining. It involves identifying the two points of perspective switching, a move from the *proponent* to the *opponent* role and back. We study this classification problem using two different corpora, on the microtext and on the ProCon corpus. Recall that the commentary texts are longer and more complex, and the opponent role can be encoded in quite subtle ways, which is why we expect the classification to be more difficult there.

### 6.4.1 Experimental setup

#### Task

We operationalise the task exactly as in the prior experiment: The goal is to assign the labels 'proponent' and 'opponent' to the individual segments. Those that are assigned the opponent label are considered to be counter-considerations.

#### Data

We use the German microtext corpus in the **microtext-graph-ADU** version and the ProCon corpus as our source data. While the microtexts are manually segmented into ADUs, we use an automatic segmentation module for German to split the ProCon texts. This is a statistical system trained on a similar corpus [Sidarenka et al., 2015], which aims at identifying clause-sized segments from the output of a dependency parser [Bohnet, 2010]. Segmentation leads to 2074 segments, which have then been annotated with the proponent/opponent label by two expert annotators.

Note that there is a considerable difference in segmentation: manual ADUs on the microtext corpus versus automatic EDUs on the ProCon corpus. We thus have to be careful in comparing the results of this experiment across corpora. This difference goes back to the fact that the EDU segmentation of the microtext corpus was not yet available when this experiment was conducted. A comparison between EDU and ADU could be problematic when multiple EDU are joined to form one complex ADU, but these structures are very infrequent in the microtext corpus for the opponent. On the proponent side these configuration do occur sometimes, but the main class of interest in this experiment is the opponent and we can refer to scores for that class in case of doubt. We thus consider this comparison to be meaningful.

Concerning the complexity of this annotation task for humans, we want to recapitulate the results found in Chapter 4: On the microtexts, naive and untrained student annotators reached an agreement of $\kappa$=0.52 in distinguishing proponent and opponent, more experienced students $\kappa$=0.60, while expert annotators achieved perfect agreement $\kappa$=1.0 (see

Table 4.1). For the ProCon texts we observe an expert agreement of $\kappa$=0.732 (see the category definition test for "opponent" in Table 4.13).

#### Genre comparison

To get a clearer picture of the distribution of opponent segments, we study their frequency and position in the individual texts: Table 6.5a summarises the corpus statistics. First of all, we observe that opponent segments are more frequent in the microtext corpus (with 21.7%) than in the ProCon corpus (with 8.3%). The difference between ADU and EDU segmentation might artificially increase this, but we can still assume objections to be significantly less frequent in ProCon commentaries. Furthermore, Table 6.5b shows the number of texts by the number (n) of included opponent segments. We observe that microtexts typically have one opponent segment as expected, and only some of the texts have no opponent segment, as e.g. in the case of implicit objections.[7] In contrast to that, 37% of the ProCon texts have no objection mentioned. When there is some opponent role present in a commentary, it is likely to cover more than one segment. Finally, Table 6.5c gives the percentage of opponent segments occurring in the first to fifth chunk of the text. There is clear tendency for opponent segments to appear in the opening of a ProCon text. In the microtexts, objections are more equally spread across the text, with a tendency towards the fourth and a smaller tendency towards the second chunk.

### 6.4.2 Models

In the previous study we compared various different classifier types. The maximum entropy model proved to be (together with the linear SVM) the most successful model. In this experiment we thus choose a maximum entropy model again: We trained a linear log-loss classifier using stochastic gradient descent learning as implemented in the Scikit learn library [Pedregosa et al., 2011]. The learning rate is set to optimal decrease, and the class weights are adjusted according to class distribution. Also, automatic feature selection is applied, i.e. the classifier selects the best $k$ features to be part of the model. The parameter $k$ is determined through hyper-parameter optimisation as one of 25, 50, 75, 100, 250, 500, 1000, or all.

Beside the two different corpora, we compare three models in this experiment, differing not in the underlying classifier but in the complexity of their feature set: two simple bag-of-word models as baselines and one model with additional features from automatic linguistic analysis. The first model (B) only extracts bag-of-words features in the target segment. The second model (B+C) additionally extracts these features from the preceding and the

[7] As an implicit objection we consider objections that the author expresses in a more compact way than a full proposition, e.g. through nominalisations or prepositional phrases. These are typically raised and countered in the very same proposition and thus do not constitute a segment on their own.

| | microtexts | ProCon |
|---|---|---|
| texts | 112 | 124 |
| segments | 576 | 2074 |
| segments (proponent) | 451 | 1902 |
| segments (opponent) | 125 | 172 |
| segments per text | 5.1±0.8 | 16.9±3.1 |
| opp. seg. per text | 1.1±0.7 | 1.4±1.5 |

(a) General statistics (averages with standard deviation).

| n | microtexts | ProCon |
|---|---|---|
| 0 | 15 | 46 |
| 1 | 74 | 32 |
| 2 | 18 | 16 |
| 3 | 5 | 17 |
| 4 | | 6 |
| 5 | | 3 |
| 6 | | 3 |

(b) Frequency of opponent segments: the number of texts with *n* opponent segments in the corpus.

| p | microtexts | ProCon |
|---|---|---|
| 1/5 | 16.0% | 35.5% |
| 2/5 | 23.2% | 18.6% |
| 3/5 | 17.6% | 19.1% |
| 4/5 | 28.8% | 12.8% |
| 5/5 | 14.4% | 11.6% |

(c) Position of the opponent segments across the text divided into fifths.

Table 6.5: Corpus statistics comparing opponent segment annotation in the microtext and the ProCon corpus.

subsequent segments, thus providing a small context window. The full model (B+C+L) adds parsing-based features for the whole context window, such as pos-tags, lemma- and pos-tag-based dependency-parse triples, the morphology of the main verb [Bohnet, 2010], as well as lemma-bigrams. Discourse connectives are taken from a list by Stede [2002] and used both as individual items and as indicating a coherence relation (Cause, Contrast, etc.). Furthermore, we use some positional statistics such as relative segment position, segment length, and punctuation count. Note that these features are extracted as in the previous study.

All results are reported as average and standard deviation over the 50 folds resulting from 10 iterations of 5x3 nested cross validation, using a group-wise, stratified folding. The models are optimised for macro averaged F1-score. Besides Cohen's Kappa $\kappa$ and macro averaged F1 scores, we also report results for the class of interest, the opponent class, in terms of Precision, Recall, and F1.

### 6.4.3 Results

The performance of the classifiers is shown in Table 6.6. Comparing the results for the two datasets confirms our assumption that the task is much harder on the ProCon texts. When comparing the different models, we observe that the simple baseline model without context performs poorly; adding context improves the results significantly. The full feature-set (B+C+L) always yields best results, except for a small drop of precision on the ProCon texts. The improvement of the full model over B+C is significant for the microtexts ($p < 0.003$ for $\kappa$, F1 macro and opponent F1), but not significant for the ProCon texts.

Feature selection mostly supports the classification of the ProCon texts, where the mass of extracted features impairs the generalisation. Typically only 25 features were chosen. For the microtexts, reducing the features to the 50 best-performing ones still yields good but not the best results. One reason for the difference in feature selection behaviour between the datasets might be that the proportion of proponent and opponent labels is more skewed for the ProCons than for the microtexts. Another reason might be the richer set of expressions marking the role switch in the ProCon texts.

A common observation for both corpora is that the connective *aber* ('but') in the subsequent segment is the best predictor for an opponent role. Other important lexical items (also as part of dependency triples) are the modal particles *natürlich* ('of course', 'naturally') and *ja* (here in the reading: 'as is well-known'), and the auxiliary verb *mögen* (here: 'may'). All of these occur in the opponent role segment itself, and they have in common that they "colour" a statement as something that the author concedes (but will overturn in the next step), which corresponds to the temporary change of perspective. As for differences between the corpora, we find that the connective *zwar*, which introduces a concessive minor clause, is very important in the microtexts but less prominent in ProCon. We attribute this

| | microtexts | | | ProCon | | |
|---|---|---|---|---|---|---|
| | B | B+C | B+C+R | B | B+C | B+C+R |
| $\kappa$ | .375±.109 | .503±.080 | .545±.098 | .187±.064 | .320±.078 | .323±.091 |
| F1 macro | .685±.056 | .751±.040 | .772±.049 | .588±.033 | .659±.040 | .660±.047 |
| opponent P. | .548±.097 | .647±.081 | .668±.096 | .428±.165 | .370±.063 | .361±.074 |
| opponent R. | .474±.146 | .575±.084 | .626±.108 | .163±.054 | .400±.109 | .422±.117 |
| opponent F1 | .497±.101 | .604±.065 | .640±.081 | .225±.064 | .378±.073 | .382±.083 |

Table 6.6: Results for role-identification, reported as average and standard deviation

to the microtext instruction of writing rather short texts, which supposedly leads the writers to often formulating their counter-considerations as compact minor clauses, for which *zwar* ('granted that') is the perfect marker. Presumably for the same reason we observe that the concessive subordinator *obwohl* ('although') is among the top-10 features for microtexts but not even among the top-50 for ProCon. In ProCon, the group of connectives indicating the Contrast coherence relation is a very good feature, and it is absent from the microtext top-50; recall, though, that the single connective *aber* ('but') is their strongest predictor, and the very similar *doch* is also highly predictive.

### Error analysis

Argumentative role identification is not an easy classification task. For microtexts, the results can be considered fairly satisfactory. For ProCons, there is a significant drop in F1 macro, and even more so for opponent precision, recall, and F1. This was in principle to be expected, but is worth to be investigated more deeply in a qualitative error analysis.

**Segmentation** As pointed out, ProCon texts have been automatically segmented, which leads to a number of errors that generate some of the classification problems; we found, however, that this is only a small factor.

There are other points to remark on segmentation, though. First, we find 37 cases where more than one opponent role segment appear in a sequence (mostly two of them, but ranging up to six), as compared to 68 cases of individual segments. The sequences pose problems for segment-wise classification focusing on perspective *change* signals, especially when the context window is small. Many of the sequences occur right at the beginning of the text, where the author provides an extended description from the opponent's view, and then switches to his own perspective. Correctly identifying complete sequences would require a deeper analysis of cohesive devices for finding continuation or break of topic/perspective/argumentative orientation. Also, note that many of the sequences actually contain

argumentative sub-structure, where, for example, the possible objection is first backed up with purported evidence and then refuted.

Furthermore, the question of segmentation grain-size arises. In the present annotation, we do not label segments as 'opponent role' when they include not only the opponent's objection but also the author's refutation or dismissal. This is because on the whole, the segment conveys the author's (proponent's) position. A translated example from the corpus is: "Not convincing at all is the argument that to the government, teachers should be worth more than a one-Euro-job." Besides such cases of explicit dismissal, we find, for instance, concessive PPs that include an opposing argument: "Despite the high cost, the building must be constructed now." We leave it to future work to dissect such complex segments and split them into an opponent and a proponent part.

**Connectives** Contrastive connectives are very good indicators for changing back from the opponent role to the proponent role, but unfortunately they occur quite frequently also with other functions. Consider for example the ProCon corpus: There are 105 opponent segments or sequences thereof in the corpus, but 195 instances of the words *aber* and *doch*, which are the most frequent contrastive connectives. Therefore, their presence needs to be correlated with other features in order to serve as reliable indicators.

**Language** While our focus in this paper was on the performance difference between the German microtexts and the ProCon texts, we want to remark that the overall classification results for microtexts do hardly differ between the German and English version (this will be demonstrated in the following section where we classify both languages). This leads us to expect that for English pro/contra commentaries, we would likewise obtain results similar to those for German, and that the results of comparing genre, as done in this experiment, are not specific to German only.

### 6.4.4 Conclusion

Our prior exploratory study showed that argumentative role classification is, though challenging, one of the tasks that are feasible to tackle on the microtext corpus. In this section, we investigated this more deeply by comparing the results on the microtext corpus with those on the more complex ProCon commentaries.

We compared the frequency and position of opponent segments in both corpora and found less opponents segment in the commentary corpus, relative to its size. Although these counter-considerations are not as frequent as supporting arguments, they still constitute rhetorical moves that authors regularly advance to strengthen their points. After all, refuting a potential objection is in itself an argument in support of the conclusion. Almost two thirds of the newspaper pro/contra texts in our corpus have counter-considerations,

and so we think these devices are definitely worth studying in order to arrive at complete argumentation analyses.

We obtained good results on our corpus of microtexts, whereas we see room for improvement for the longer and more complex pro/contra newspaper texts. In an error analysis, we found that contrastive discourse markers are very important for the classification, but that they are often indicative not only for pragmatic but also for semantic contrasts. Furthermore, the frequency of certain connectives is different across the corpora.

One of the shortcomings of the experimental setup was the difference in the underlying segmentation between the microtext and the ProCon corpus. However, in our error analysis we found this to have only a small impact. Finally, it remains for future work to provide a more elaborate, quantitative analysis of the linguistic signals of argumentative role and role switches.

## 6.5 Study 3: The Evidence Graph - A global model of argumentation structure

Identifying the structure of argumentation according to our scheme involves choosing one segment as the central claim of the text, deciding how the other segments are related to the central claim and to each other, identifying the argumentative role of each segment, and finally the argumentative function of each relation.

Several of these task have been already tackled in our prior experiments on automating the recognition of argumentation structure. There, we approached the problem as a segment-wise classification task (see Section 6.3). Formulating the task this way was successful for the recognition of argumentative role and function of a segment. For the automation of the structure building however, the segment-wise classification of attachment with only a small context window around the target segment proved to be a very hard task: The identification of the target of an argumentative relation was especially challenging for relations between non-adjacent segments. These long-distance dependencies are frequently found in argumentation graphs. For example, 46% of the relations marked in the corpus used for this study involve non-adjacent segments. For longer texts this number might increase further: Stab and Gurevych [2014a] report a rate of 63% of non-adjacent relations in their corpus.

Another problem is, that the predictions of the classifiers presented above cannot necessarily be united to a valid argumentation structure: The predictions might contradict, or introduce cycles or disconnected segments when interpreting them as a whole structure. While the predictions might by locally optimal, i.e. in the context of the single classification instance, they might be dispreferred globally, i.e. in the context of all other classification instances brought together to determine an overall valid argumentation structure.

In this study, we therefore frame the task of structure building differently: The attachment classification is considered a binary decision, where the classifier, when given a pair of a source and a target segment, chooses whether or not to establish a relation from the source to the target. Since these relations can hold not only between adjacent but between arbitrary segments of the text, all possible combinations of segments are required to be tested. Secondly, we will use a *decoding* mechanism to derive globally optimal predictions. Thirdly, we will show that jointly predicting different aspects of the argumentation graph through such a decoding mechanism further improves the classification results. Consequently, we will first focus solely on the task of relation identification and then consider also ADU type and relation type classification.

The feature set that we will use in this study is very similar to that of the previous ones. We want to forestall, however, that a more elaborate feature set leading to better classifiers will be used in the next section 6.6. This also involved a technical change of the underlying linguistic pipeline, which will be elaborated on later. While all results and comparison to be presented now are nonetheless conclusive, we want to make the reader aware that the final and best base classifier results will be shown only after this study.

### 6.5.1 Experimental setup

#### Data

In this experiment, we use the **microtext-dep-ADU-reduced** version the microtext corpus, where the argumentation structures have been converted to ADU-based dependency trees, and the relation labels have been reduced (see Chapter 5.5). Results will be shown for both the German and the English version of the corpus.

#### Task

We distinguish the following classifications tasks in this experiment:

- **attachment (at):** Is there an argumentative connection between the source and the target segment? In the corpus, a relation has been annotated for 464 segment pairs, no relation has been annotated for the combinatorially remaining 2,000 pairs of segments.
- **central claim (cc):** Is the current segment the central claim of the text? In our data 112 of the 576 segments are central claims.
- **role (ro):** Does the current segment present a claim of the proponent or the opponent? In our data 451 of the 576 segments are proponent segments and 125 are opponent segments.

- **function (fu):** Has the current segment a supporting or an attacking function? In our data, 290 segments are supports, 174 are attacks, and 112 are the central claim and thus have no own function.

Note that role and function classification tasks are framed as in our previous study (in Section 6.3). The 'comp' level is obsolete due to the reduction of relation labels. The segment-wise 'target' classification is replaced by pair-wise attachment classification.

### 6.5.2 Models

We compare two heuristic baseline models and different data-driven models that we developed, each of them trained and evaluated separately on both language versions of the corpus. All models are evaluated on the basis of 10 iterations of 5x3-fold nested cross validation, with a text-wise stratified folding (see Section 6.2 on methodology).

#### Baseline: attach to first

In the English-speaking school of essay writing and debating, there is the tendency to state the central claim of a text or a paragraph in the very first sentence, followed by supporting arguments. It is therefore a reasonable baseline to assume that all segments attach to the first segment. In our corpus, the first segment is the central claim in 50 of the 112 texts (44.6%).

This baseline (**BL-first**) will not be able to capture serial argumentation, where one more general argument is supported or attacked by a more specific one. However, it will cover convergent argumentation, where separate arguments are put forward in favour of the central claim (given that it is expressed in the first segment). It will always produce flat trees. In our corpus, 176 of the 464 relations (37.9%) attach to the first segment.

#### Baseline: attach to preceding

A typically very strong baseline in discourse parsing is attaching to the immediately preceding segment [Muller et al., 2012a]. This is certainly true for rhetorical structure trees, where most of the relations are between adjacent segments. Since argumentation structures often exhibit non-adjacent relations, this baseline might be considered slightly weaker in our scenario, but it is still a challenging heuristic.

This baseline (**BL-preced.**) will always produce chain trees and thus cover serial argumentation, but not convergent argumentation. In our corpus, 210 of all 464 relations (45.3%) attach to the preceding segment.

#### Learned attachment without decoding

We train a linear log-loss model (**simple**) using stochastic gradient descent (SGD) learning, with elastic net regularisation, the learning rate set to optimal decrease and class weight adjusted according to class distribution [Pedregosa et al., 2011]. The following hyper parameters are tuned in the inner CV: the regularisation parameter alpha, the elastic net mixing parameter, and the number of iterations. We optimise for macro averaged F1-score.

For each text segment, we extract binary features for lemma, pos-tags, lemma- and pos-tag- based dependency-parse triples, and the main verb morphology [Bohnet, 2010], and discourse connectives [Stede, 2002], furthermore simple statistics like relative segment position, segment length, and punctuation count. These features are equivalent to those in the prior studies (see Section 6.3.1). Furthermore, we extract for each pair of text segments the relative distance between the segments and their linear order (is the source before or after the target). The feature vector for each pair then contains both pair features and segment features for source and target segment and their adjacent segments.

Note that we experimented with several features, some of which were dismissed from the final evaluation runs due to lack of impact. This included sentiment values and the presence of negation for segments. Also, similarity measures had been proposed as useful features. However, in our experiments all following distance measures between pairs of segments did not affect the results: word-overlap, tf-idf, and LDA distributions.

#### Learned attachment with MST decoding

The simple model just described might be able to learn which segment pairs actually attach, i.e., correspond to some argumentative relation in the corpus. However, it is not guaranteed to yield predictions that can be combined to a tree structure again. A more appropriate model would enforce global constraints on its predictions. In the **simple+MST** model, this is achieved by a *minimum spanning tree* (MST) decoding, which has first been applied for syntactic dependency parsing [McDonald et al., 2005a,b] and later for discourse parsing [Baldridge et al., 2007, Muller et al., 2012a]. First, we build a fully-connected directed graph, with one node for each text segment. The weight of each edge is the attachment probability predicted by the learned classifier for the corresponding pair of source and target segment. We then apply the Chu-Liu-Edmonds algorithm [Chu and Liu, 1965, Edmonds, 1967] to determine the minimum spanning tree, i.e., the subgraph connecting all nodes with minimal total edge cost (in our case highest total edge probability). This resulting tree then represents the best global attachment structure for a text given the predicted probabilities.

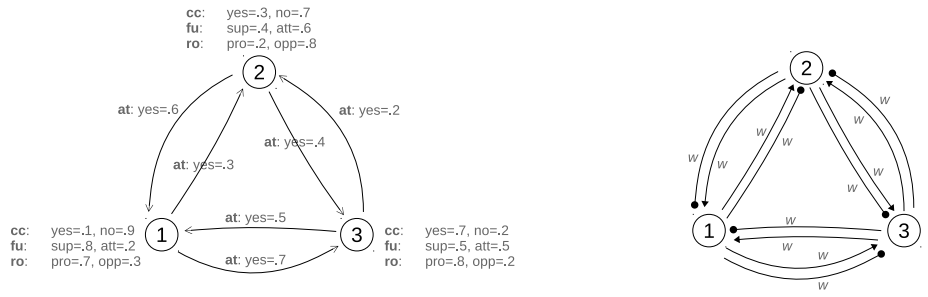



Figure 6.2: An example evidence graph before (left) and after (right) the predicted probabilities of the different levels have been combined in a single edge score.

### The Evidence Graph: Joint prediction with MST decoding

All models presented in the previous subsections have in common that they do not rely on other features of the argumentation graph. However, it is fair to assume that knowledge about the argumentative role and function of a segment or its likeliness to be the central claim might improve the attachment classification. Consequently, our next model considers not only the predicted probability of attachment for a segment pair, but also the predicted probabilities of argumentative role, function and of being the central claim for each segment. The predictions of all levels are combined in one *evidence graph*.

**Additional segment-wise base classifiers:** We train base classifiers for the role, function and central claim level using the same learning regime as used in the simple model. Contrary to the attachment classification, the items are not segment pairs but single segments. We thus extract all segment-based features as described above for the target segment and its adjacent segments.

**Combining segment and segment-pair predictions:** Our goal in this model is to combine the predicted probabilities of all levels in one edge score, so that the MST decoding can be applied as before. Figure 6.2 depicts the situation before and after the combination, first with separate prediction for segments and segment pairs and then with the combined edge scores.

The evidence graph is constructed as follows: First, we build a fully connected multigraph over all segments with as many edges per segment-pair as there are edge types. In our scenario there are two edge types, supporting and attacking edges. Then we translate the segment-wise predictions into level-specific edge scores.

The edge-score for the central claim level $\overline{\text{cc}}_{i,j}$ is equal to the probability of the edge's source not to be the central claim. This captures the intuition that central claims are unlikely to have outgoing edges:

$$\overline{\text{cc}}_{i,j} = \quad p(\text{cc}_i = \text{no}) \tag{6.1}$$

The edge-score for the argumentative function level $\overline{\text{fu}}_{i,j}$ is equal to the probability that the source segment has the corresponding function:

$$\overline{\text{fu}}_{i,j} = \begin{cases} p(\text{fu}_i = \text{sup}) & \text{for sup. edges} \\ p(\text{fu}_i = \text{att}) & \text{for att. edges} \end{cases} \tag{6.2}$$

The edge-score for the argumentative role level $\overline{\text{ro}}_{i,j}$ is also determined by the edge type. Attacking edges involve a role switch (proponent or opponent would not attack their own claims respectively), while supporting edges preserve the role (proponent or opponent will only support their own claims respectively):

$$\overline{\text{ro}}_{i,j} = \begin{cases} p(\text{ro}_i = \text{pro}) \times p(\text{ro}_j = \text{pro}) + \\ p(\text{ro}_i = \text{opp}) \times p(\text{ro}_j = \text{opp}) & \text{for sup. edges} \\ \\ p(\text{ro}_i = \text{pro}) \times p(\text{ro}_j = \text{opp}) + \\ p(\text{ro}_i = \text{opp}) \times p(\text{ro}_j = \text{pro}) & \text{for att. edges} \end{cases} \tag{6.3}$$

Finally, of course the edge-score for the attachment level $\overline{\text{at}}_{i,j}$ is equal to the probability of attachment between the segment pair:

$$\overline{\text{at}}_{i,j} = \quad p(\text{at}_{i,j} = \text{yes}) \tag{6.4}$$

The combined score of an edge $w_{i,j}$ is then defined as the weighted sum of the level-specific edge score:

$$w_{i,j} = \frac{\phi_1 \overline{\text{ro}}_{i,j} + \phi_2 \overline{\text{fu}}_{i,j} + \phi_3 \overline{\text{cc}}_{i,j} + \phi_4 \overline{\text{at}}_{i,j}}{\sum \phi_n} \tag{6.5}$$

In our implementation, the combined evidence graphs can be constructed without a weighting, and then be instantiated with a specific weighting to yield the combined edge scores $w_{i,j}$.

**Procedure:** As before, we first tune the hyperparameters in the inner CV, train the model on the whole training data, and predict probabilities on all items of the test set. Also, we predict all items in the training data "as unseen" in a second inner CV using the best hyperparameters. This procedure is executed for every level. Using the predictions of all four levels, we then build the evidence graphs for training and test set.

**Finding the right weighting:** We evaluate two versions of the evidence graph model. The first version (**EG equal**) gives equal weight to each level-specific edge score. The second version (**EG best**) optimises the weighting of the base classifiers with a simple evolutionary search on all evidence graphs of the training set, i.e. it searches for a weighting that maximises the average level evaluation score of the decoded argumentation structures in the training set. Finally, all evidence graphs of the test set are instantiated with the selected weighting (the equal one or optimised one) and evaluated.

#### Comparison: MST parser

Finally, we compare our models to the well-known mstparser[8], which was also used in the discourse parsing experiments of Baldridge et al. [2007]. The mstparser applies 1-best MIRA structured learning, a learning regime that we expect to be superior to the simple training in the previous models. In all experiments in this paper, we use ten iterations for training, non-projective 1-best MST decoding, and no second order features. The base mstparser model (**MP**) evaluated here uses the same features as above, as well as its own features extracted from the dependency structure. Second, we evaluate a pre-classification scenario (**MP+p**), where the predictions of the base classifiers trained in the above models for central claim, role, and function are added as additional features. We expect this to improve the central claim identification as well as the edge labelling.

For the full task involving all levels, we combine the mstparser with an external edge labeller, as the internal one is reported to be weak. In this setting (**MP+r**), we replace the edge labels predicted by the mstparser with the predictions of the base classifier for argumentative function. Furthermore, the combination of pre-classification, mstparser and external relation labeller (**MP+p+r**) is evaluated. Finally, we evaluate a scenario (**MP$\epsilon$+p+r**) where the mstparser has access only to its own features and to those of the pre-classification, but not to the extracted features used by the simple model, and the external relation labeller is used. In this scenario, the mstparser exclusively serves as a meta-model on the base classifier's predictions.

### 6.5.3 Results

As in the previous studies, all results are reported as average and standard deviation over the train-test splits, i.e. over 50 split resulting from ten iterations of (the outer) 5-fold cross validation.

#### Attachment task

Table 6.7 shows the results in the attachment task. The rule-based baseline scores are equal for both languages, since they rely only on the annotated structure of the parallel corpus. Here, attach-to-first is the lower bound, attach-to-preceding is a more competitive baseline, as we had hypothesised in Section 6.5.2.

The learned classifier (simple) beats both baselines in both languages, although the improvement is much smaller for English than for German. In general, the classifier lacks precision compared to recall: It predicts too many edges. As a result, the graph constructed from the predicted edges for one text very often does not form a tree. In Table 6.8, we give a summary of how often tree constraints are fulfilled, showing that without decoding, valid

[8] http://sourceforge.net/projects/mstparser/

| | BL-first | BL-preced. | simple | simple+MST | EG equal | EG best | MP | MP+p |
|---|---|---|---|---|---|---|---|---|
| F1 macro | .618±.041 | .662±.025 | .679±.025 | .688±.032 | .712±.026 | .710±.028 | .724±.030 | **.728**±.033 |
| attach F1 | .380±.067 | .452±.039 | .504±.038 | .494±.053 | .533±.042 | .530±.044 | .553±.048 | **.559**±.053 |
| $\kappa$ | .236±.081 | .325±.050 | .365±.048 | .377±.064 | .424±.052 | .421±.055 | .449±.060 | **.456**±.066 |
| trees | 100% | 100% | 15.4% | 100% | 100% | 100% | 100% | 100% |

(a) German

| | BL-first | BL-preced. | simple | simple+MST | EG equal | EG best | MP | MP+p |
|---|---|---|---|---|---|---|---|---|
| F1 macro | .618±.041 | .662±.025 | .663±.030 | .674±.036 | .692±.034 | .693±.031 | .707±.035 | **.720**±.034 |
| attach F1 | .380±.067 | .452±.039 | .478±.049 | .470±.058 | .501±.056 | .502±.052 | .524±.056 | **.546**±.056 |
| $\kappa$ | .236±.081 | .325±.050 | .333±.059 | .347±.071 | .384±.068 | .386±.063 | .414±.070 | **.440**±.069 |
| trees | 100% | 100% | 11.6% | 100% | 100% | 100% | 100% | 100% |

(b) English

Table 6.7: Results for the attachment task: for German (a) and English (b), best values highlighted.

| | German | | English | |
|---|---|---|---|---|
| total graphs | 1120 | 100.0% | 1120 | 100.0% |
| rooted | 1091 | 97.4% | 1088 | 97.1% |
| cycle free | 1059 | 94.6% | 995 | 88.8% |
| full span | 908 | 81.1% | 864 | 77.1% |
| out degree | 298 | 26.6% | 283 | 25.3% |
| trees | 173 | 15.4% | 120 | 10.7% |

Table 6.8: Number and percentage of valid trees for the "simple" attachment model

trees can only be predicted for 15.4% of texts in German and for 10.7% of texts in English. The most frequently violated constraint is "out degree", stating that every node in the graph should have at most one outgoing edge. Note that all other models, the baselines as well as MST decoding models, are guaranteed to predict tree structures.

The simple+MST model yields slightly lower F1-scores for positive attachment than without decoding, trading off a loss of ten points in recall of the over-optimistic base classifier against a gain of five in precision. However, the output graphs are constrained to be trees now, which is rewarded by a slight increase in the summarising metrics macro F1 and $\kappa$.

The evidence graph models (EG equal & EG best) clearly outperform the simple and simple+MST model, indicating that the attachment classification can benefit from jointly predicting the four different levels. Note that the EG model with equal weighting scores slightly better than the one with optimised weighting for German but not for English. However, this difference is not significant ($p>0.5$) for both languages, which indicates that the search for an optimal weighting is not necessary for the attachment task.

The overall best result is achieved by the mstparser model. We attribute this to the superior structured learning regime. The improvement of MP over EG equal and best is significant in both languages ($p<0.008$). Using pre-classification further improves the results, although difference is neither significant for German ($p=0.4$) nor for English ($p=0.016$).

**Full task**

Until now, we only focused on the attachment task. In this subsection we will present results on the impact of joint prediction for all levels.

The results in Table 6.9 show significant improvements of the EG models over the base-classifiers on the central claim, the function and the attachment levels ($p<0.0001$). This demonstrates the positive impact of jointly predicting all levels. The EG models achieve the best scores in central claim identification and function classification, and the second best result in role identification. The differences between EG equal and EG best are not

| | | simple | EG equal | EG best | MP | MP+p | MP+r | MP+p+r | MP$\epsilon$+p+r |
|---|---|---|---|---|---|---|---|---|---|
| cc | maF1 | .849±.035 | .879±.042 | **.890**±.037 | .825±.055 | .855±.055 | .825±.055 | .855±.055 | .854±.053 |
| | $\kappa$ | .698±.071 | .759±.085 | **.780**±.073 | .650±.111 | .710±.110 | .650±.111 | .710±.110 | .707±.105 |
| ro | maF1 | **.755**±.049 | .737±.052 | .734±.046 | .464±.042 | .477±.047 | .656±.054 | .669±.062 | .664±.053 |
| | $\kappa$ | **.511**±.097 | .477±.103 | .472±.092 | .014±.049 | .022±.063 | .315±.106 | .340±.122 | .330±.105 |
| fu | maF1 | .703±.046 | .735±.045 | **.736**±.043 | .499±.054 | .527±.047 | .698±.054 | .723±.052 | .723±.050 |
| | $\kappa$ | .528±.068 | .573±.066 | **.570**±.063 | .293±.056 | .326±.056 | .522±.076 | .557±.075 | .560±.073 |
| at | maF1 | .679±.025 | .712±.026 | .710±.028 | .724±.030 | **.728**±.033 | .724±.030 | **.728**±.033 | .724±.029 |
| | $\kappa$ | .365±.048 | .424±.052 | .421±.055 | .449±.060 | **.456**±.066 | .449±.060 | **.456**±.066 | .448±.059 |

(a) German

| | | simple | EG equal | EG best | MP | MP+p | MP+r | MP+p+r | MP$\epsilon$+p+r |
|---|---|---|---|---|---|---|---|---|---|
| cc | maF1 | .817±.045 | .860±.051 | **.869**±.053 | .780±.063 | .831±.059 | .780±.063 | .831±.059 | .823±.063 |
| | $\kappa$ | .634±.090 | .720±.103 | **.737**±.107 | .559±.126 | .661±.118 | .559±.126 | .661±.118 | .647±.122 |
| ro | maF1 | **.750**±.045 | .721±.051 | .720±.047 | .482±.053 | .475±.047 | .620±.064 | .638±.057 | .641±.062 |
| | $\kappa$ | **.502**±.090 | .445±.098 | .442±.092 | .024±.068 | .015±.060 | .243±.126 | .280±.114 | .285±.122 |
| fu | maF1 | .671±.049 | .707±.048 | **.710**±.050 | .489±.062 | .514±.059 | .642±.057 | .681±.057 | .677±.059 |
| | $\kappa$ | .475±.074 | .529±.070 | **.530**±.072 | .254±.058 | .296±.063 | .440±.081 | .491±.083 | .486±.083 |
| at | maF1 | .663±.030 | .692±.034 | .693±.031 | .707±.035 | **.720**±.034 | .707±.035 | **.720**±.034 | .713±.033 |
| | $\kappa$ | .333±.095 | .384±.068 | .386±.063 | .414±.070 | **.440**±.069 | .414±.070 | **.440**±.069 | .427±.066 |

(b) English

Table 6.9: Results for the full task: for German (a) and English (b), best values highlighted.

significant on any level, which again indicates that we can dispense with the extra step of optimising the weighting and use the simple equal weighting. These result are consistent across both languages.

The pure labelled mstparser model (MP) performs worse than the base classifiers on all levels except for the attachment task. Adding pre-classification yields significant improvements on all levels but role identification. Using the external relation labeller drastically improves function classification and indirectly also role identification. The combined model (MP+p+r) yields best results for all mstparser models, but is still significantly outperformed by EG equal in all tasks except attachment classification. There, the mstparser models achieve best results, the improvement of MP+p+r over EG equal is significant for English ($p<0.0001$) and for German ($p=0.001$). Interestingly, the meta-model (MP$\epsilon$+p+r) which has access to its own features and to those of the pre-classification, but not to the features used in the simple model, performs nearly as well as the combined model (MP+p+r).

The only level not benefiting from any MST model in comparison with the base classifier is the role classification: In the final MST, the role of each segment is only implicitly represented, and can be determined by following the series of the role-switches of each argumentative function from the proponent root to the segment. The loss of accuracy for predicting the argumentative role is much smaller for German than for English, probably due to the better attachment classification in the first place.

Finally, note that the EG best model gives the highest total score when summed over all levels, followed by EG equal and then MP+p+r.

**Projecting further improvements:** We have shown that joint prediction of all levels in the evidence graph models helps to improve the classification on single levels. To measure exactly how much a level contributes to the predictions of other levels, we simulate better base classifiers and study their impact. To achieve this, we artificially improved the classification of one target level by overwriting a percentage of its predictions with ground truth. The overwritten predictions where drawn randomly, corresponding to the label distribution of the target level. E.g. for a 20% improvement on the argumentative function level, the predictions of 20% of the true "attack"-items were set to attack and the predictions of 20% of the true "support"-items were set to support, irrespective of whether the classifier already chose the correct label.

The results of the simulations are presented in Figure 6.3 for English; the results for German exhibit the same trends. The figure plots the $\kappa$-score on the y-axis against the percentage of improvement on the x-axis. Artificially improved levels are drawn as a dashed line. As the first plot shows, function classification is greatly improved by a better role classification (due to the logical connection between them), whereas the other levels are unaffected. In contrast, all levels would benefit from a better function classification, most importantly even the attachment classification. Potential improvements in the central claim identification mostly affect function classification (as these classification tasks partly over-

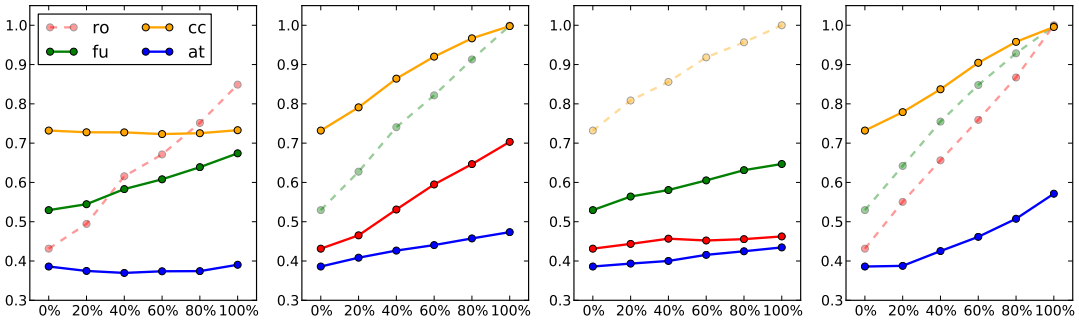



Figure 6.3: Simulations of the effect of better base classifiers in the EG equal model for English: dashed levels artificially improved, x = number of predictions overwritten with ground truth; y = average $\kappa$ score in 10 iterations of 5-fold CV.

lap: central claims will not be assigned a function they cannot have). Finally, a combined improvement on the logically coupled task of role and function identification, would even more help the attachment classification. It might thus be useful to work on a better joint role and function classifier in the near future.

**Evidence combination:** Combining the evidence in an edge score as a weighted sum of probabilities (see equation 6.5), instead of a product of probabilities might be considered inadequate, and result in a model that optimises the highest scored but not the most probable structure.[9] In order to investigate this question, we compared the EG equal against an EG model with a product of probability. The model scores are nearly identical and do not show a significant difference.

### 6.5.4 Conclusions

In this section, we introduced a new approach to argumentation mining. Our *evidence graph* model jointly predicts different aspects of the structure by combining the predictions of different subtasks in the edge weights of an evidence graph; we then apply a standard MST decoding algorithm. This model not only outperforms two reasonable baselines and two simple models for the difficult subtask of attachment/relation identification, but also improves the results for central claim identification and relation classification, and it compares favourably to a 3-pass mstparser pipeline.

To the best of our knowledge, this model, originally published as [Peldszus and Stede, 2015a], is the first data-driven model that optimises argumentation structure globally for the complete sequence of input segments. Furthermore, it is the first model jointly tackling segment type classification, relation identification and relation type classification.

[9] This question was raised by one of the reviewer of the published paper describing this experiment [Peldszus and Stede, 2015a].

In our experiment, we used the ADU segmented corpus with the reduced relation set, which was a reasonable simplification for an initial investigation. We will come back to this in Section 6.7, where we not only consider the fine-grained relation set but also start from the EDU segmented arguments. But before we do this, we compare it with other decoding approaches that have been recently proposed in the literature, and additionally present an improved set of base classifiers.

## 6.6 Study 4: Comparing decoding mechanisms

In the previous section, we have presented a model to predict globally optimal text-level argumentation structures, given the output of four local models. Since the evidence graph model has been published [Peldszus and Stede, 2015a], other approaches for predicting argumentation structures have been proposed: Like our approach and similar to discourse parsing, these approaches follow the decoding paradigm [Smith, 2011], i.e. they opt for learning a local model and then perform global decoding over the local probability distributions, often imposing constraints that are specific to the dataset used. While our approach was based on a Minimum Spanning Trees (MST) decoding, other approaches choose what Smith [2011] categorises under polytope decoding, and more specifically a decoding based on Integer Linear Programming (ILP). Two examples are [Stab and Gurevych, 2016] and [Persing and Ng, 2016a]. The approach of Stab and Gurevych [2016] was also evaluated on the English microtext corpus, which allows us a direct comparison.

Our goals in this study are thus twofold: Our main focus is to investigate the impact of the decoding strategy. For this purpose, we test different decoders on the same corpus, using the same local models. We replicate characteristic features of the ILP decoders of Stab and Gurevych [2016] and Persing and Ng [2016a]. Also, we propose a novel ILP decoder with additional constraints, that make use of the interaction between different aspects of argumentation structure. We compare all decoders to the MST-based evidence graph model. Furthermore, we aim to improve the results of our local models, most importantly that for function classification, where Stab and Gurevych [2016] improved upon our earlier results. We will add new features we consider to be helpful and also incorporate features inspired by their feature set.

### 6.6.1 Experimental setup

#### Data and Task

As in the previous study (in Section 6.5.1), we use the **microtext-dep-ADU-reduced** version the microtext corpus, where the argumentation structures have been converted to ADU-

based dependency trees, and the relation labels have been reduced. Results will be shown for both the German and the English version of the corpus.

The task setting is likewise exactly that of the previous study: We predict and evaluate the four levels of attachment (at), central claim (cc), role (ro), and function (fu).

### 6.6.2 Local models

One aim of this study is to improve the base classifiers by extending the feature set. This involved exchanging the component for syntactic analysis: In the previous study in Section 6.5 we used the Mate parser [Bohnet, 2010], which offered excellent parsing accuracy as well as morpho-syntactic tagging (for German). In this experiment we instead tested the *spacy* parser [Honnibal and Johnson, 2015]. Both parsers provide pretrained models for English and German. The spacy parser is slightly less accurate and does not yet offer morphological tagging, but it is very fast and due to its Python interface allowed us to integrate it into the model more elegantly.[10] Most importantly, it comes with Brown-clusters and vector-space representations, which we want to test in addition to the features proposed in the earlier studies.

In comparison to the feature set used in the previous study, we implemented the following extra features: We added Brown cluster unigrams (BC) and bigrams (BC2) of words occurring in the segment. We completed the discourse relations features (DR): While the lexicon of discourse connectives for German used in our previous experiments was annotated with potentially signalled discourse relations, the English lexicon was lacking this information. We thus extended the English connective lexicon by those collected in the EDUCE project[11], which also have been annotated with signalled discourse relations. In addition, a feature representing the main verb of the segment was added (VS); the already existing verb features either focused on the verb of the whole sentence which might be too restrictive, or on all possible verbal forms in the segment which might not be restrictive enough.

In order to investigate the impact of word embeddings for this task, we added the 300-dimensional word-vector representations, averaged over all content words of the segment, as a feature for segment wise classifiers (VEC). Stab and Gurevych [2016] gained small improvements – around 1 point F1-score on their dataset – by adding word-embeddings as features to their argumentative stance classifier. Moreover, we derived scores of semantic distance between two segments using these vectors: We measured the cosine distance between the average word vector representations of the segment and its left and right antecedents (VLR). Also, for the attachment classifier, we measured the cosine distance between the average word vectors of the source and target segment (VST).

[10] The Mate parser, which we used before, was called as a preprocessing step before model-building. This involved temporary files and additional logic in the evaluation procedure. The spacy parser on the other hand can be directly integrated into the feature function, making it very suitable for end-to-end modelling.

[11] `https://github.com/irit-melodi/educe`

Finally, we added features for better capturing the inter-sentential structure, i.e. for relations with subordinate clauses: one feature representing that the source and target segments are part of the same sentence (SS) and one representing that the target is the matrix clause of the source (MC).

As in our previous experiment, all classifiers for the segment-wise classification tasks (cc, ro, and fu) will use the same feature set. The segment-pair-wise classification task of attachment (at) will use all segment-wise features for the source and target segment, as well as pair-wise features (such as the VST for semantic distance between source and target).

### 6.6.3 Decoders

The most important goal of this study is to compare various decoding mechanism on the same data using the same local models. In the following, we present the decoders that we want to compare. We will first propose a novel ILP decoder, which contrary to the approaches of Stab and Gurevych [2016] and Persing and Ng [2016a] integrates predictions on the role level and is thus able to make use of interactions between the role and other levels. We will then describe our replication of characteristic features of the two related decoders and compare this against our MST-based evidence graph model. All ILP-based decoders in this study use the SCIP suite for solving [Gamrath et al., 2016].

#### Novel ILP decoder

As an alternative to the evidence graph model, we now present a new decoder based on ILP. The goal is to build a directed acyclic graph $G =< V, E, R >$. Vertices (ADUs) are referred by their position in textual order, indexed starting from $i$ or $j = 1$ to the total number $n = |V|$. The argumentative functions $central\_claim, attack, support$ are referred by their respective indexes $\nu_{cc} = 1$, $\nu_a = 2$, $\nu_s = 3$. We create four sets of core variables corresponding to the levels of prediction:

$$cc_i = 1 \equiv adu_i \text{ is a central claim}$$

$$ro_i = \begin{cases} 1 & \text{if } adu_i \text{ is a proponent node} \\ 0 & \text{if } adu_i \text{ is an opponent node} \end{cases}$$

$$fu_{ik} = 1 \equiv adu_i \text{ has function label } k$$

$$at_{ij} = 1 \equiv (i, j) \in E$$

The local models described above provide us with four real-valued functions:

$$s_{cc} : \{1, \ldots, n\} \mapsto \mathbb{R}$$
$$s_{ro} : \{1, \ldots, n\} \mapsto \mathbb{R}$$
$$s_{fu} : \{1, \ldots, n\} \times \{v_{cc}, v_a, v_s\} \mapsto \mathbb{R}$$
$$s_{at} : \{1, \ldots, n\}^2 \mapsto \mathbb{R}$$

Then, the objective function that we try to maximise is a linear combination of the four functions:

$$S = \sum_{i=1}^{n} s_{cc}(i) cc_i + \sum_{i=1}^{n} s_{ro}(i) ro_i + \sum_{i=1}^{n} \sum_{k=1}^{3} s_{fu}(ik) fu_{ik} + \sum_{i=1}^{n} \sum_{j=1}^{n} s_{at}(ij) at_{ij}$$

We define different sets of constraints. Each set is identified with a short name, given in parentheses in the header. The novel ILP decoder will include all constraint sets. In addition, we will investigate different combinations of constraint sets, in order evaluate the individual impact of each set.

**Tree constraints on the attachment predictions (tree):** We require that the predicted graphs are trees. This amounts to all nodes having one or no outgoing arc (6.6) and as many arcs as nodes except the root node (6.7). Also, we require that our graphs are acyclic. For this we introduce an auxiliary set of integer variables $c_i$ and impose constraints that enforce acyclicity (6.8 and 6.9).

$$\forall i \quad 0 \leq \sum_{j}^{n} at_{ij} \leq 1 \tag{6.6}$$

$$\sum_{i,j}^{n \times n} at_{ij} \quad = n - 1 \tag{6.7}$$

$$\forall i \quad 1 \leq c_i \leq n \tag{6.8}$$

$$\forall i, j \quad c_j \leq c_i - 1 + n(1 - at_{ij}) \tag{6.9}$$

**Relation labelling through function predictions (label)**: The constraints above only yield an unlabelled tree. We want to use the prediction of function classifier and assign one argumentative function to every node (6.10). Furthermore, we only want to predict support or attack as an relation label, since the central claim function is not a relation label and the decoded tree might have another root than the segment predicted to have the function central claim (6.11).

$$\forall i \quad \sum_{k=1}^{3} fu_{ik} = 1 \tag{6.10}$$

$$\forall i \quad fu_{iv_{cc}} = 0 \tag{6.11}$$

**Interaction between central claim and attachment (cc-at):** Until now, we have only combined the attachment and the function predictions. The other levels' predictions have

not been integrated and no interaction constraints between them have yet been defined. In the following, we integrate the predictions of the designated central claim classifier and also describe the relation between the identified central claim and the root of the attachment tree. First, we require that only one central claim is chosen (6.12). Secondly, all vertices have exactly one outgoing edge with the exception of the central claim, which is a sink node (6.13): If $adu_i$ is the central claim, all $at_{ij}$ will be set to 0. If not, there will be only one $at_{ij}$ set to 1. Note that once these constraints are added, the root constraints for attachment (6.6 and 6.7) are redundant.

$$\sum_{i=1}^{n} cc_i = 1 \tag{6.12}$$

$$\forall i \quad \left( cc_i + \sum_{j=1}^{n} at_{ij} \right) = 1 \tag{6.13}$$

**Interaction between central claim and role (ro-cc):** Integrating the predictions of the role classifier allows us to define the simple requirement that the central claim must be a proponent node (6.14). This bans the case $cc_i = 1, ro_i = 0$, where the central claim is an opponent node. All other cases are allowed.

$$\forall i \quad cc_i \leq ro_i \tag{6.14}$$

**Interaction between role and function (ro-fu):** More importantly, we can now describe in detail the relationship between argumentative functions and roles. The aim of the following constraints is to represent the intuition that every argumentative role (proponent or opponent) will only support itself, not the other, and only attack the other, not itself. This means that supporting relations are role-preserving, and attacking relations are role-inverting. Consider an edge from $adu_i$ to $adu_j$. We build the following table that represents which role and function configurations are valid:

| $at_{ij}$ | $ro_i$ | $fu_{iv_s}$ | $ro_j$ | valid? | Comments |
|---|---|---|---|---|---|
| 0 | * | * | * | yes | No attachment, no restrictions |
| 1 | 0 | 0 | 0 | no | OPP attacks OPP |
| 1 | 0 | 0 | 1 | yes | OPP attacks PRO |
| 1 | 0 | 1 | 0 | yes | OPP supports OPP |
| 1 | 0 | 1 | 1 | no | OPP supports PRO |
| 1 | 1 | 0 | 0 | yes | PRO attacks OPP |
| 1 | 1 | 0 | 1 | no | PRO attacks PRO |
| 1 | 1 | 1 | 0 | no | PRO supports OPP |
| 1 | 1 | 1 | 1 | yes | PRO supports PRO |

We now define $S_{ij} = ro_i + fu_{iv_s} + ro_j$. The table can be reduced to:

| $at_{ij}$ | $S_{ij}$ | valid? |
|---|---|---|
| 0 | * | yes |
| 1 | 0 | no |
| 1 | 1 | yes |
| 1 | 2 | no |
| 1 | 3 | yes |

We introduce a set of auxiliary variables $psp_{ij}$, which are set to 1 if and only if $adu_i$ and $adu_j$ form a "PRO supports PRO" pattern. In this case the ADUs need not be attached and the defining constraint is as follows:

$$\forall i,j \quad 0 \leq S_{ij} - 3psp_{ij} \leq 2 \tag{6.15}$$

If $0 \leq S_{ij} \leq 2$, then $psp_{ij}$ must be 0, or the sum will be negative. If $S_{ij} = 3$, then $psp_{ij}$ must be 1, or the sum will be greater than 2. We now define $K_{ij} = S_{ij} - 2psp_{ij}$. The table can be completed:

| $at_{ij}$ | $S_{ij}$ | $psp_{ij}$ | $K_{ij}$ | valid? |
|---|---|---|---|---|
| 0 | * | * | * | yes |
| 1 | 0 | 0 | 0 | no |
| 1 | 1 | 0 | 1 | yes |
| 1 | 2 | 0 | 2 | no |
| 1 | 3 | 1 | 1 | yes |

If $at_{ij} = 1$, then the case is valid iff $K_{ij} = 1$. If $at_{ij} = 0$, then $K_{ij}$ can take any value between 0 and 2. Therefore, we build the following constraint:

$$\forall i,j \quad at_{ij} \leq K_{ij} \leq 2 - at_{ij} \tag{6.16}$$

We will refer to this decoder as **new ILP** in the presentation of results. Different variations of the constraint sets will be tested. All of them share the basic (tree) and the (label) constraint set, in order to decode labelled trees. We will test the impact of adding the interaction constraints respectively, and also experiment with combinations of them.

#### ILP approach by Stab and Gurevych

We replicate the approach of Stab and Gurevych [2016], who proposed an ILP decoder for deriving argumentation structures. The model has been primarily developed on the corpus of persuasive essays and results have been reported for the second release of the corpus [Stab and Gurevych, 2016], but they also report results on the English microtext corpus.

Their base classifiers include one for distinguishing between claims and premises (the distinction between major claims and claims is not necessary for the microtexts), and one for identifying argumentative relations, i.e. for attachment. In order to replicate this setup, we

can use our base classifiers for central claim and attachment. The classification of argumentative function (in their terms 'stance') is independent in their model, i.e. the distinction between support and attack is not part of the joint modelling and could also be regarded as a following step in a pipeline, contrary to both the evidence graph model and the novel ILP decoder. We can use the output of the function base classifier for this. Role classification is not used in their approach.

In contrast with us, they do not use probability distributions of the local model in order to perform global decoding over them. Instead, they take classification decisions to create matrices depicting these results. They linearly combine these matrices with another matrix derived from a combination of incoming and outgoing links on the non-decoded graph. The intention of this procedure is to promote central claims as relation targets and to degrade central claims as relation sources, in order to model the interaction between central claim and attachment. This linear combination provides a new matrix which they use in order to maximise their objective function. ILP constraints are merely used to guarantee a rooted tree without cycles. We will refer to this decoder as **repl. ILP S&G** in the presentation of results.

#### ILP approach by Persing and Ng

We replicate the decoding approach of Persing and Ng [2016a], who worked on the first release of the student essay corpus [Stab and Gurevych, 2014b]. This replication amounts only to the objective function and the general (corpus-independent) decoding of argumentation structures using ILP. It does not aim to cover the various other aspects of their end-to-end model.

Their setup of base classifiers is different from all other models presented here: They combine attachment and function classification in one classifier for relation identification that assigns for a pair of ordered segments one of five classes: forward-directed support, backward-directed support, forward-directed attack, backward-directed attack, or no relation. In addition, they have one base classifier for assigning ADU types (major claim, claim, premise or none) to segments.

In contrast to the ILP approach of [Stab and Gurevych, 2016] and to our novel ILP decoder, they aim to maximise the average F-score of these two base classifiers, rather than maximising the average classification accuracy. They achieve this by estimating the expected values of TP, FP, and FN values from the results of their two classifiers.

The constraints that they use include constraints on major claims (exactly one major claim, in the first paragraph, no parents), premises (at least one parent from the same paragraph, only in paragraph relations), claims (at most one parent which is a major claim) as well as some other constraints (at most two argumentative components per sentence, etc).

Although their base classifier setup is different, we can use our attachment and function classifier to emulate the results of their combined attachment and function classifier, and use it together with the central claim classifier for distinguishing between claims and premises to replicate their objective function. We include only the tree constraints and the interaction between central claim and attachment. All other constraints they propose are domain-specific for the annotated student essays (paragraph structure and the major-claim/claim distinction). We will refer to this decoder as **repl. ILP P&N** in the presentation of results.

#### Novel approach with objective from Persing & Ng

In this model, we use the aforementioned objective function by Persing and Ng [2016a] in combination with the set of constraints of the novel ILP decoder. We will refer to this decoder as **new ILP objective 2** in the presentation of results.

#### Evidence graph

Our baseline in this experiment is the evidence graph model, see Section 6.5.2. It uses the predictions of all four (improved) base classifiers, combines them in one evidence graph and decodes the globally optimal structure using the MST algorithm. Compared to the constraints described above, the EG model is theoretically equally powerful: It enforces a tree structure via the MST algorithm. It models all three interactions between the levels by the combination of level-specific edge scores. In this experiment, we again leave the weighting of the base classifiers equal. We will refer to this decoder as **new EG equal** in the presentation of results.

### 6.6.4 Results

#### Evaluation procedure

In this experiments, we follow exactly the evaluation procedure of the previous study. The correctness of our local model's predictions as well of those of the predicted structures is assessed separately for the four subtasks, reported as macro averaged F1.

While these four scores cover important aspects of the structures, it would be advantageous to have a unified, summarising metric for evaluating the decoded argumentation structures. To our knowledge, no such metric has yet been proposed; prior work just averaged over the different evaluation levels. Here, we will additionally report labelled attachment score (LAS) as a measure that combines attachment and argumentative function labelling as it is commonly used in dependency parsing. Note however, that this metric is not specifically sensitive for the importance of selecting the right central claim and also

| | English | | | | German | | | |
|---|---|---|---|---|---|---|---|---|
| model | cc | ro | fu | at | cc | ro | fu | at |
| Section 6.5 | .817 | .750 | .671 | .663 | **.849** | .755 | .703 | .679 |
| Stab & Gurevych [2016] | .830 | | **.745** | .650 | | | | |
| base | .832 | .762 | .710 | .690 | .827 | .757 | .709 | .696 |
| base + BC | +.008 | -.005 | +.001 | +.004 | +.008 | +.005 | -.001 | -.003 |
| base + BC2 | | -.003 | -.002 | +.001 | -.001 | +.003 | | -.001 |
| base + DR | +.005 | +.018 | +.019 | +.003 | +.002 | -.002 | | -.001 |
| base + VS | -.001 | -.002 | -.001 | +.002 | +.001 | | +.001 | -.001 |
| base + VEC | -.002 | -.002 | -.002 | +.001 | +.004 | -.003 | +.002 | +.002 |
| base + VLR | | -.002 | | +.001 | -.001 | | +.001 | -.002 |
| base + VST | | | | | | | | -.001 |
| base + SS | | | | +.009 | | | | +.009 |
| base + MC | | | | +.012 | | | | +.016 |
| all - VEC | **.840** | **.782** | .723 | **.711** | .837 | **.765** | .709 | **.711** |
| all | .840 | .780 | .724 | .710 | .836 | .762 | **.712** | **.711** |

Table 6.10: Evaluation scores for the base classifiers reported as macro avg. F1: The first two rows report on earlier results. Against this we compare the new classifiers using the new linguistic pipeline (base), followed by a feature study showing the impact of adding the new features (described in Section 6.6.2). Finally, we show the results of the final classifiers combining these features.

not sensitive for the dialectical dimension (choosing just one incorrect argumentative function might render the argumentative role assignment for the whole argumentative thread wrong).

**Local models**

The results of the experiment with local models are shown in Table 6.10. We first repeat the reported results of the previous study from Section 6.5, which have been published as [Peldszus and Stede, 2015a], and those of [Stab and Gurevych, 2016] for comparison. Their base classifiers improved over ours in central claim identification by 1.3 point and considerably in function classification by 7.4 points. Role classification was not tackled by them. For attachment, they scored 1.3 below our results. They evaluated their classifiers on the English, but not on the German version of the corpus.

We then report the results of the new base classifiers of this study: First of our local models without any new features added, but exchanging the linguistic pipeline of the Mate parser by that of the spacy parser (base). This already provides a substantial improvement on all levels for the English version of the dataset. We attribute this mainly to the better

performance of spacy in parsing English. For German, the results are competitive. Only for central claim identification our new local models does not fully match the original model, which might be due to the fact that the spacy parser does not offer a morphological analysis as deep as the mate parser and thus does not derive predictions for sentence mood.

What follows is a feature analysis, where we report on the impact of adding each new feature to the base model (base + x). Scores are reported as the delta. Empty cells indicate no change. The highest gain is achieved by adding the features for subordinate clauses (SS and MC) to the attachment classifier. Brown cluster unigrams (BC) give a moderate boost for central claim identification. Interestingly, the word-vector representation did not have a significant impact. The averaged word embeddings themselves (VEC) lowered the scores minimally for English and improved the results minimally for German, but increased the training time dramatically.[12] The distance measures based on word vectors (VST and VLR) yielded no improvement likewise.

Taking all features together provides us with local models that achieve state-of-the-art performance on all levels but fu for English and cc for German. Note that we exclude from our final model only the feature that adds the raw word embeddings (VEC), since it does not yield improvements across the levels when combined with all other features and also substantially slows down the model training. The resulting models (all - VEC) are used as the base classifiers in all decoding experiments.

#### Global model

The results of the experiments with the decoders are shown in Table 6.11. We again first repeat previously reported scores, i.e. the results of the evidence graph model from Section 6.5, which has been published as [Peldszus and Stede, 2015a], as well as the results of Stab and Gurevych [2016].

We then present the results for decoders developed in this study. Recall that these decoders all use the output of the same, improved base classifiers (the best performing all - VEC models), but that not all decoders make use of all task levels: The replicated ILP approaches for example make no use of the predictions of argumentative role. Nevertheless, we are able to derive and evaluate the predicted argumentative role from the predicted argumentation structures of these approaches.

Let us first focus on the novel ILP decoder and the impact of adding the different constraint sets to the baseline, which just predicts labelled trees without exploiting any interaction. Adding the cc-at interaction constraints yields an improvement on the central claim and function level. The ro-cc interaction does not increase the scores on its own, but it gives

[12] One explanation for the missing impact of the raw word embeddings could be that we used pre-trained word embeddings and did not learn representations specific for our task, as advised by Ji and Eisenstein [2014] in the context of RST parsing.

| | English | | | | | German | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| model | cc | ro | fu | at | LAS | cc | ro | fu | at | LAS |
| old EG equal (Section 6.5) | .860 | .721 | .707 | .692 | .481 | **.879** | .737 | .735 | .712 | .508 |
| [Stab and Gurevych, 2016] | .857 | | .745 | .683 | | | | | | |
| new ILP (no interaction, just labelled trees) | .844 | .689 | .733 | .715 | .494 | .858 | .656 | .719 | .722 | .490 |
| new ILP (cc-at) | .870 | .699 | .752 | .716 | .502 | .865 | .651 | .725 | .722 | .493 |
| new ILP (ro-cc) | .844 | .689 | .733 | .715 | .494 | .858 | .656 | .719 | .722 | .490 |
| new ILP (ro-fu) | .846 | .770 | .742 | .718 | .516 | .852 | .745 | .726 | .729 | .517 |
| new ILP (cc-at + ro-cc) | .870 | .701 | .752 | .716 | .503 | .872 | .654 | .730 | .724 | .497 |
| new ILP (cc-at + ro-cc + ro-fu) | .862 | **.783** | .750 | .720 | .524 | .870 | .753 | **.740** | **.733** | .528 |
| new ILP (cc-at + ro-cc + ro-fu) objective 2 | .866 | .782 | .753 | **.722** | .526 | .867 | **.756** | .739 | **.733** | **.529** |
| repl. ILP S&G | .837 | .673 | .727 | .687 | .456 | .834 | .654 | .704 | .690 | .451 |
| repl. ILP P&N | .869 | .699 | .751 | .716 | .502 | .866 | .653 | .726 | .723 | .494 |
| new EG equal | **.876** | .766 | **.757** | **.722** | **.529** | .861 | .730 | .725 | .731 | .523 |

Table 6.11: Evaluation scores for the decoders reported as macro avg. F1 for the cc, ro, fu, and at levels, and as labelled attachment score (LAS). The first two rows report on earlier results. In the following block of rows, we present the novel ILP decoder model with different sets of constraints used: The first only produces labelled trees without exploiting interactions. We then report on the impact of adding the interaction constraints. In the final row block, we report on our replication of related approaches and on the evidence graph model serving as a baseline.

a little extra when combined with the cc-at interaction constraint sets. A strong improvement in role classification and a smaller one in function classification is achieved by adding the ro-fu interaction. The full constraint set with all three interactions included yields the best novel ILP decoder model. Changing our objective function against that of Persing and Ng [2016a] (objective 2) does not significantly affect the results.

When we compare the replicated decoders (repl ILP S&G) and (repl ILP P&N) against our novel ILP decoder, we observe that they perform worse by nearly 10 points F1 in role classification. This is to some degree expected, as these approaches do not involve a role classifier designated to learn this dialectical view on argumentation structure and thus no interactions involving the role level can be exploited. The novel ILP decoder, however, also yields better results in attachment and (for German) in function classification. The results of (repl ILP P&N) are nearly equal to that of new ILP (cc-at): This is expected, as firstly their constraints are very similar, and secondly the special objective function of P&N has been shown to not have a significant effect here. To our surprise, the (repl ILP S&G) performs worse that the labelled tree baseline, although it adds a variant of the cc-at interaction that the labelled tree baseline does not have. Bear in mind, though, that our replications only amount to characteristic features of their decoders and constraints that are applicable on the microtext corpus. The result we obtained here do not represent what their whole system (including their local models and domain-specific constraints) might predict.

We observed that the novel ILP decoder with the full constraint set gives the best result under all ILP decoders. But how does it compare to the MST-based evidence graph model? Our results show that the EG model and the new ILP decoder are generally on par, but have different strengths. The EG model is better in central claim identification, the new ILP decoder is better in role classification. Both models perform equally on the attachment level. Function classification, on the other hand, varies depending on the language. This is also supported when looking at the LAS metric, where (new EG equal) scores best for English, and the new ILP decoder for German. The differences in cc, ro, and fu are statistically significant. However, they are spread across different levels and partly depend the modelled language. We therefore cannot conclude that one approach is superior to the other, but only that on a level playing field (in terms of local models and structural interactions exploited) the MLT-based and the ILP-based decoders can yield equivalent results.

Finally, it is worth pointing out that the improved evidence graph model for English and the novel ILP decoder for German represent the new state of the art for automatically recognising the argumentation structure in the microtext corpus.

### 6.6.5 Conclusions

We presented a comparative study of various structured prediction methods for the extraction of argumentation graphs. For this we compared the MST-based evidence graph model

that we presented in the previous section against ILP decoding approaches. In particular we replicated the decoders of Stab and Gurevych [2016] and Persing and Ng [2016a] and proposed a novel ILP decoder with several constraint sets, some of them novel, each of which we could demonstrate the impact of. In order to be able to compare between the various decoding mechanisms, we used the same underlying corpus, the microtext corpus, and the same local models, which are an improved version of the base classifiers presented in the previous section.

Our observation is that the novel ILP decoder outperforms existing ILP approaches, but that it is on par with the MST-based evidence graph model. We therefore argue that when it comes to predicting argumentation structures in general and as long as these structures are represented as tree structures, MST-based methods suffice to obtain the globally optimal structure. ILP-based decoding approaches may pay off when *extra* constraints on the output structure can be defined, e.g. when the texts follow certain domain-specific regularities, as it might be the case for student essays or juridical documents. Whether these extra constraints make a significant difference in comparison to a general MST-based decoding of argumentation structure remains to be shown in each specific case.

The improved local models derived in this section, in combination with the evidence graph model and the novel ILP decoder, improve upon previous results from Section 6.5 and upon those of [Stab and Gurevych, 2016] and thus represent state of the art results on the microtext corpus.

## 6.7 Study 5: Fine-grained argumentation structures

Until now, we have investigated the predictability of argumentation structures in a (to a certain degree) simplified way. This concerns on the one hand the segmentation – we used the coarse-grained ADU segmentation – and on the other hand the granularity of the relation set – we reduced the relation labels to the binary distinction between support and attack. These simplifications had their justification: They allowed us to assess the complexity of structure prediction in abstraction of the problem of segmentation and without tackling the oftentimes challenging distinction between various fine-grained relation types, such as e.g. rebutting versus undercutting attacks.

In this study, we want to increase the complexity of the task and report on results of the proposed evidence graph model on less simplified versions of the corpus. We will not only predict structures with the full, fine-grained relation set, but also use the EDU segmentation of the corpus. The goal is to arrive at a model that is able to predict fine-grained argumentation structures from EDUs, such as those derived by an automatic discourse segmenter.[13] We will proceed step by step and present the results for the intermediary scenarios (ADU

[13] Note that we still train our models on a corpus where all segments (even EDUs) are argumentatively relevant. We will not tackle the task to determine argumentative relevance here. Nevertheless, starting from an EDU

| corpus version | relation set | | | | |
|---|---|---|---|---|---|
| | root | role-preserving | | | role-inverting |
| ADU reduced | central claim | | | support | attack |
| EDU reduced | central claim | | | support | attack |
| ADU full | central claim | | link | support, example | rebut, undercut |
| EDU full | central claim | join | link | support, example | rebut, undercut |

Table 6.12: Relation sets on the argumentative function level for the different corpus versions.

segmentation with the full set of relations, as well as EDU segmentation with the reduced set of relations), and finally for the target scenario with EDU segmentation and fine-grained relations.

### 6.7.1 Experimental setup

#### Data and Task

Contrary to the previous studies, we will therefore not use the **microtext-dep-ADU-reduced** version of the microtext corpus as the data in our experiment, but all other three conversions: **microtext-dep-ADU-full**, **microtext-dep-EDU-reduced**, and finally **microtext-dep-EDU-full**. Results will be shown for both the German and the English version of the corpus for the scenario with ADU segmentation, and only for English in the EDU-based scenarios.[14]

The task setting is similar to that of the previous studies: We predict and evaluate the four levels of attachment (at), central claim (cc), role (ro), and function (fu). The function level, though, has different labels depending on the corpus version. An overview of the relation sets of the different corpus versions is given in Table 6.12. Instead of the simple binary distinction between support and attack, the full relations set offers multiple supporting (normal support, support by example) and multiple attacking relations (rebut and undercut). Also, there is the link relation, expressing that the source segment and the targeted segment have their argumentative function only when taken together. Finally, with EDU segmentation, we have the join relation, indicating that multiple adjacent EDUs form an ADU. As all role-preserving functions, both link and join are mapped to support in the reduced label set. See Chapter 5.3.2 for details about the label set reduction.

segmentation involves determining whether multiple adjacent EDUs form an ADU together or whether each of them constitutes an ADU in its own right.

[14] An EDU-segmented version of the German corpus is not yet available, see Chapter 5.4.

### 6.7.2 Models

As a baseline, we will use the two heuristic baseline models presented in Section 6.5.2: **BL-first**, which takes the first segment as the central claim and all other segments as direct support to it; and **BL-preced.**, which attaches each segment to its preceding segment as a simple support relation. Note that the purpose of these baselines in this case is not to seriously challenge the proposed method. It is rather the other MST- or ILP-based methods presented earlier who have proven to be competitors of the evidence graph model. The purpose of the baselines here is to determine the lower bound using exactly the same methods as in previous experiments. This will enable us to put these base results in relation to those derived on different corpus versions in order to quantify the increase in task complexity.

We will not reproduce all other competitors spawned in the previous sections: For the mstparser-based models from Section 6.5, we have already shown that, despite their advantage in predicting attachment, the overall performance was not on par with that of the evidence graph model. Furthermore, they prove to be significantly worse in predicting the argumentative function – one of the task that will be even more challenging in this study's scenarios. The ILP-based decoders from Section 6.6 would be interesting to compare to. However, the more complex relation set would require customisation of the existing constraints and maybe additional constraints for the new relation types, which is out of the scope of the present study.

The evidence graph model does not require any adaption to the new scenarios. It only needs to be aware of those relations that are role-inverting. The graph is automatically populated with edges of all existing relation types, be it only support and attack (as in previous experiments) or the six different relations of the full EDU-based relation set. We will, again, only report the results of the **EG-equal** model, which does not optimise the weighting of the four base classifier scores but simply assumes an equal weighting.

### 6.7.3 Results

For evaluating the results of our experiment, we follow the same procedure of the previous studies and assess the correctness of the predicted structures separately for each of the four subtasks as macro averaged F1 and also report labelled attachment scores (LAS).

#### ADU reduced

Let us first, for the ease of comparison, recapitulate the results on the ADU segmented corpus with the reduced relation-set. They are shown in Table 6.13a and consist of the baselines as well as the (feature-wise improved) EG-equal model from the previous section. Remember that baseline results only differ from each other in their attachment strategy and yield equal results for central claim (always choosing the first segment), role (all pro-

| | English | | | | | German | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| model | cc | ro | fu | at | LAS | cc | ro | fu | at | LAS |
| BL-first | .712 | .439 | .407 | .618 | .313 | .712 | .439 | .407 | .618 | .313 |
| BL-preced. | .712 | .439 | .407 | .662 | .300 | .712 | .439 | .407 | .662 | .300 |
| EG-equal | .876 | .766 | .757 | .722 | .529 | .861 | .730 | .725 | .731 | .523 |

(a) Evaluation results for all levels (recapitulation of results from Section 6.6).

| | English | | | German | | |
|---|---|---|---|---|---|---|
| relation | Precision | Recall | F1 | Precision | Recall | F1 |
| central claim | .801 | .801 | .801 | .777 | .777 | .777 |
| support | .762 | .817 | .788 | .728 | .819 | .770 |
| attack | .736 | .641 | .683 | .730 | .563 | .630 |

(b) Individual relations on the function level for the EG-equal model

Table 6.13: Evaluation results for the *ADU reduced* scenario.

ponent), and function (all support). Also, following a structural heuristic, their results are consistent across language versions of the corpus. Finally, note that the evidence graph model produces slightly better predictions for the English version of the corpus than for the German version, as we already discussed in Section 6.6.

We will additionally report detailed scores for the different relations on the argumentative function level and later compare them with the results using the more fine-grained relations set. The results for the reduced relation set are presented in Table 6.13b as precision, recall, and F1 score for each relation class.

**ADU full**

We will now turn to the first scenario with more fine-grained structures, using the full relation set on the ADU segmented corpus. The level-wise scores are reported in Table 6.14a. Comparing these scores against the those with the reduced relation set confirms our expectation that only the level of argumentative function is affected. On all other levels the results are stable. The decrease in the scores for argumentative function is observed for both the baseline models as well as for the evidence graph model. For the baselines, the drop is quite dramatic, but remember that they only assign support relations and the function score is a *macro* average over all classes. With more classes the macro average will drop, even though roughly the same number of items are correctly predicted. This also has to be kept in mind when interpreting the function score for the EG-equal model.

| | English | | | | | German | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| model | cc | ro | fu | at | LAS | cc | ro | fu | at | LAS |
| BL-first | .712 | .439 | .194 | .618 | .301 | .712 | .439 | .194 | .618 | .301 |
| BL-preced. | .712 | .439 | .194 | .662 | .250 | .712 | .439 | .194 | .662 | .250 |
| EG-equal | .873 | .770 | .456 | .718 | .487 | .860 | .731 | .444 | .733 | .473 |

(a) Evaluation results for all levels

| | English | | | German | | |
|---|---|---|---|---|---|---|
| relation | Precision | Recall | F1 | Precision | Recall | F1 |
| central claim | .796 | .796 | .796 | .774 | .774 | .774 |
| support | .687 | .783 | .731 | .649 | .752 | .695 |
| example | .060 | .027 | .037 | .050 | .030 | .037 |
| link | .170 | .135 | .134 | .263 | .230 | .224 |
| rebut | .607 | .524 | .558 | .590 | .489 | .528 |
| undercut | .538 | .449 | .482 | .525 | .348 | .407 |

(b) Individual relations on the function level for the EG-equal model

Table 6.14: Evaluation results for the *ADU full* scenario.

To get a better picture, let us see the individual scores for each relations, presented in Table 6.14b. In general, we observe the highest F-scores for central claim and normal support. The fine-grained attack types, i.e. rebutter and undercutter, can be predicted to a good extent, but there are still improvements to be made here. Linked relations and support by example are not sufficiently predicted. We can assume that there are just not enough instances in the training data to adequately cover these classes. Note that without these two infrequent classes, the macro average over the remaining relation types is 0.633 for English and 0.601 for German.

When comparing these figures against those obtained with the reduced relation set, we observe that supports are harder to identify with the full than with the reduced set. One reason for this may be those classes that were subsumed under support in the reduced scenario (examples and links) and which are not sufficiently covered in the corpus. Another reason may be found on the attack side: Among the common structural confusions is the discrimination of reactions towards attacks: One choice is a counter-attack (typically by an undercutter) the other a new pro-argument, which is not challenging the acceptability of the attack's inference, but rather outweighing it with a "more important" reason. With the full relation set the model can distinguish between rebutting and undercutting attacks, and thus confuse undercuts with supports in this configurations. This is also supported by the

fact that undercuts score lower than rebuts. Compared to the reduced relation set, where attack relations scored with 0.68 F1, we achieve only an F1 of 0.56 for rebutting and of 0.48 for undercutting relations (in the case of English).

Interestingly, the relations results also differ between the English and the German version of the corpus. The quite infrequent linked relations are predicted more accurately in German, but all other relations achieve better scores in English, even with a difference of 7 points F1 for the undercutting relation. Without further analysis, we can only speculate here about the possible reasons. We leave it for future work to investigate whether this language dependence is due to differences in coverage of the lexical features (from parsing model and connective lexicon), or an artefact of translation, or more generally due to a difference in the usage, variety and markedness of the discourse connectives across both languages.

Summing up this scenario, we found that some distinctions from the more fine-grained relation set, such as between rebutting and undercutting attacks, can be successfully predicted to some extent. Other relations, such as link and example, are just not covered sufficiently in the training corpus to be reliably predicted. Also, we want to stress that increasing the complexity of the relation set (even to a degree that certain relations cannot be predicted) did only affect the function level and thus did not impair the quality of joint prediction in the evidence graph model for other levels.

#### EDU reduced

The next scenario is based on the EDU segmentation, but with the reduced relation set. Recall that only the English version of the corpus offers EDU segmentation. The overall results are shown in Table 6.15a. In order to quantify the effect of using a more fine-grained segmentation, we can compare these results with that of the ADU reduced scenario. The differences for the baseline are relative small: Attachment classification improves by three points for BL-preced., which is due to the serialisation of join relations during the dependency conversion. Furthermore, there is a small gain of two points in function classification, as these join relations are mapped to supports by relation set reduction.

The differences of the EG-equal model's prediction are also rather small. The tasks of central claim identification and role classification become a little bit more difficult (a decrease by two to three points F1) and there is a minor impact on function classification. We thus focus on individual scores for the relations, which are shown in Table 6.15b: The F-score for predicting the central claim function is 5 points lower; supporting relations are better recognised (+3 points), due to gains in precision; attacking relations finally are predicted less accurately (-2 points), due to a loss in precision.

Overall, moving from an ADU segmentation to an EDU segmentation does not make the task of argumentation mining significantly more complex, as long as we work with a reduced

| | English | | | | |
|---|---|---|---|---|---|
| model | cc | ro | fu | at | LAS |
| BL-first | .722 | .443 | .429 | .613 | .293 |
| BL-preced. | .722 | .433 | .429 | .692 | .358 |
| EG-equal | .849 | .732 | .743 | .721 | .505 |

(a) Evaluation results for all levels

| | English | | |
|---|---|---|---|
| relation | Precision | Recall | F1 |
| central claim | .747 | .747 | .747 |
| support | .807 | .832 | .819 |
| attack | .695 | .639 | .663 |

(b) Individual relations on the function level for the EG-equal model

Table 6.15: Evaluation results for the *EDU reduced* scenario.

relation set that treats EDU-joints as supporting relations. This surely is a simplification – though one that might still be a practical starting point when predicting structures right from automatically segmented text. The theoretically more adequate treatment of EDU segmentations with explicit join-relations will be handled in the next and final scenario.

**EDU full**

We will now evaluate the most elaborate setting, where both segmentation and the relation set are more complex than in the studies of the previous sections: We will predict structures with the full relation set on fine-grained EDU-segmented text. Consequently, we will be able to compare this scenario into both directions, with the ADU-full scenario in order to measure the impact of EDU segmentation, and with EDU-simple scenario to measure the impact of the full relation set.

We will start with the general evaluation results of the models, see Table 6.16a. Similar to what we had already found when assessing the impact of EDU segmentation for the reduced relation set, we observe here with the full relation set a small improvement of 3 points for the attachment results of the BL-preced. model, since joins are represented as serial structures. In function classification, however, there is a loss of 4 points: The additional join relation is not predicted by the baseline models and thus reduces the overall macro average (see above). The evidence graph model also exhibits some minor losses with -3 points on the

| | English | | | | |
|---|---|---|---|---|---|
| model | cc | ro | fu | at | LAS |
| BL-first | .722 | .443 | .155 | .613 | .255 |
| BL-preced. | .722 | .443 | .155 | .692 | .197 |
| EG-equal | .849 | .741 | .446 | .722 | .424 |

(a) Evaluation results for all levels

| | English | | |
|---|---|---|---|
| relation | Precision | Recall | F1 |
| central claim | .749 | .749 | .749 |
| support | .659 | .650 | .652 |
| example | .000 | .000 | .000 |
| join | .533 | .669 | .589 |
| link | .165 | .120 | .129 |
| rebut | .553 | .492 | .518 |
| undercut | .504 | .485 | .488 |

(b) Individual relations on the function level for the EG-equal model

Table 6.16: Evaluation results for the *EDU full* scenario.

role level, -2 for central claims, and -1 for the argumentative function. Attachment is not affected.

When we compare the scores for the levels with the EDU reduced results, the effect of making the relation set more complex is similar to what we have already seen in ADU segmentation: The results on all levels but function remain stable. For function classification, the macro-average drops significantly.

A look on the individual relation's scores of the EG-equal model will help us to investigate this in more detail (see Table 6.16b). As in the ADU full scenario, we observe that the example and the link relations are not successfully predicted, most likely because they are underrepresented in the data. Without them, the macro-average would be 0.598. The join-relation, which expresses membership of more than one EDU to an ADU, is predicted with an F-score of nearly 60% and thus recognised with greater success than attack relations.

Adding the join relation has an influence on the other relation types, though, which might have been underestimated when only looking at the loss of one point on the averaged function level. In the individual scores of relations we observe significant losses compared with the ADU full scenario: -5 points for central claim, -8 point for support (mainly a loss

in recall), and -4 points for rebuts. This means that join relations are also a source for confusions, and better features might be required to make these distinctions clearer.

Finally, we compare the results with those of the reduced relation set in the EDU reduced scenario. Simply treating all supports, examples, joins, and links as a support yields a 16 points higher F-score for this class than when the model is forced to distinguish proper supports from these other classes. On the attack side where an F-score of 0.66 was reached with the reduced relation set, rebuts score with 0.52 and undercuts with 0.48.

To summarise the results obtained in this scenario: The impact of a more fine-grained segmentation is stronger in combination with the full relation set, because introducing the join relations brings in another source of confusion. This merely affects the function level, all other levels remain relatively stable in the evidence graphs model's predictions. The fine-grained relations are not equally easy to predict, but the attacks as well as the join relation can be recognised with some success, though at the expense of accuracy for other classes.

### 6.7.4 Conclusions

Let us conclude what has been achieved in this study. We have applied the evidence graph model to more complex versions of the microtext corpus. Complexity was increased on two dimensions: Moving from ADU to EDU segmentation allows us to directly use the output of a discourse segmenter to predict argumentation structures. Using the full, and not the reduced relation set that is available in the annotated data, allows us to learn about interesting aspects of argumentation structures that go beyond the simple support versus attack distinction, such as the difference between rebutting and undercutting attacks or linked structures.

We found that using the EDU segmentation is not critical when using a reduced relation set, but affects function classification when using a fine-grained relation set. For the fine-grained relation set, we observed that we did not have enough instances of support by example and of linked relations to be able to recognise them reliably. Linked relations might even be problematic when enough data is available, as this class expresses a deep understanding of the inferences underlying the argument. The join relation for EDU-ADU-membership, as well as the distinction between rebutter and undercutter can be predicted to some extent.

Generally, fine-grained relations come at a price: It might be worth considering to be very selective when making a relation set more fine-grained. As long as most of these relations are not predicted with very high accuracy, one should not introduce additional sources of confusion with relations that are not highly desirable for the intended application.

Finally, we want to stress that this is the maximally complex scenario. The development of features was driven towards capturing the distinction between supports and attacks, but

not for more fine-grained relations. We have not, for example, studied the data explicitly for signals of rebuttal or undercutting, nor for those of EDU-ADU membership. The results nevertheless achieved by the evidence graph model in these complex scenarios thus demonstrate the versatility and stability of this approach.

## 6.8 Study 6: Predicting argumentation structures from rhetorical structure

In the previous studies, we have presented the evidence graph model and achieved state-of-the-art results on the microtext corpus comparing against various baselines and different decoding models. Furthermore, we have shown that the model is flexible to derive argumentation structures of different granularity by presenting results for the different versions of the corpus, based on ADU or EDU-segmentation and with a coarse- or fine-grained relation set. In this last study, we will demonstrate that the proposed way of deriving argumentation structures is also useful when we have something completely different than natural language text as input: We will use the rhetorical structures that have been annotated on the microtext corpus to predict argumentation structures.

This scenario is of theoretical interest, since we can deepen our understanding of the correlations between RST and argumentation. We have already discussed in Chapter 2.2.5 why argumentation structures cannot be adequately represented by RST trees. In Chapter 5.4.3, we introduced the additional annotation layer for RST in the microtext corpus. It enabled us to ground claims about the correlation between RST and argumentation in systematic empirical evidence, while related studies were based on the experiences that their authors had made with manually applying RST and with analysing argumentation. Examples of such empirical studies are the *quantitative* analysis of these alignments given in [Stede et al., 2016] and the *qualitative* analysis later provided in [Peldszus and Stede, 2016b]. These studies found a large proportion of "canonical" correspondences between RST subtrees and the central notions of argumentation. But they also identified systematic mismatches, which are either due to an inherent ambiguity of RST analysis (informational versus intentional) or caused by more technical aspects of granularity (multinuclear relations). Even if these mismatches would be resolved by using annotation guidelines that "drive" the annotator toward capturing underlying argumentation, there would still remain problems with non-adjacency in the argumentation structure, which are likely to increase when texts are larger than our microtexts.

More importantly for this section, this scenario is of practical interest: We want to investigate whether those aspects of argumentation that are represented in RST trees can be feasibly decoded in an automated fashion to predict argumentation structures. To this end, we will compare three approaches, a tree transformation approach, a graph aligner, and the

evidence graph model. If RST structures are argumentatively informative enough, either due to direct correspondences, or due to cooccurrences that can be learnt, this opens up the door to exploit existing RST parsers for argumentation mining, such as those presented in Section 6.1.8. Hence, a potentially useful architecture for argumentation mining could involve an RST parser as an early step that accomplishes a good share of the overall task. How feasible this is has so far not been determined, though. In the following, we report on experiments in automatically mapping RST trees to argumentation structures, for now on the basis of the manually-annotated "gold" RST trees.

### 6.8.1 Experimental setup

#### Data

In this study, we will use the EDU segmented microtext corpus, which is only available in English: on the one hand the RST annotation layer in its dependency conversion as the input structure, and on the other hand the argumentation structure layer in its dependency conversion with the full label set (**microtext-dep-EDU-full**) as the gold output structure. For an example of the original structures and their dependency conversion, we point the interested reader to Chapter 5.4.5, especially to Figure 5.6.

#### Task

Our aim is to predict argumentation structures as in our previous studies, but instead from a segmented natural language text, we predict them from the gold RST dependencies in our corpus here. We follow the very same experimental setup of our previous studies, i.e. the same train-test splits, and the same evaluation procedure, where the correctness of predicted structures is assessed in four subtasks at, cc, ro, and fu.

### 6.8.2 Models

We have implemented three different models: A simple heuristic tree-transformation serves as a baseline, against which we compare two data-driven models. All models and their parameters are described in the following subsections.

#### Heuristic baseline

The baseline model (**BL**) produces an argumentation structure that is *isomorphic* to the RST tree. RST relations are mapped to argumentative functions, based on the most frequently aligning class as reported in [Stede et al., 2016] – see Figure 6.4. For the two relations marked with an asterisk, no direct edge alignments could be found, and thus we assigned

**support**: background, cause, evidence, justify, list, motivation, reason, restatement, result
**rebut**: antithesis, contrast, unless
**undercut**: concession
**join**: circumstance, condition, conjunction, disjunction, e-elaboration, elaboration, evaluation-s, evaluation-n, interpretation*, joint, means, preparation, purpose, sameunit, solutionhood*

Figure 6.4: Mapping of RST relations to ARG relations used in the heuristic baseline.

them to the class of the non-argumentative *join*-relation. The argumentative *example* and *link*-relations were not frequent enough to be captured in this mapping.

We expect this baseline to be not an easy one to improve over. It will predict the central claim correctly already for 85% of the texts, due to the correspondence described in the aforementioned study. Also, 60% of the unlabelled edges should be mappable. Finally, the argumentative role is expected to be covered quite well, too: If the relation mapping is correct, the chain of supporting and attacking relations determining the role is very likely to be correct, even if attachment is wrongly predicted.

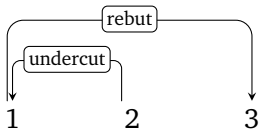
#### Naive aligner

Our naive aligner model (**A**) learns the probability of subgraphs in the RST structure mapping to subgraphs of the argumentative structure.

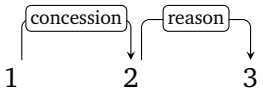
For training, this model applies a subgraph alignment algorithm yielding connected components with $n$ nodes occurring in the undirected, unlabelled version of both the RST and the argumentative structures. It extracts the directed, labelled subgraphs for these common components for both structures and learns the probability of mapping one to the other across the whole training corpus.

For prediction, all possible subgraphs of size $n$ in the input RST tree are extracted. If one maps to an argumentation subgraph according to the mapping learned on the training corpus, the corresponding argumentation subgraph is added to an intermediary multi-graph. After all candidate subgraphs have been collected, all equal edges are combined and their individual probabilities accumulated. Finally, a tree structure is decoded from the intermediary graph using the minimum spanning tree (MST) algorithm [Chu and Liu, 1965, Edmonds, 1967].

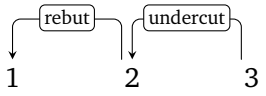
The model can be instantiated with different subgraph sizes $n$. Choosing $n = 2$ only learns a direct mapping between RST and ARG edges. Choosing larger $n$ can reveal larger structural patterns, including edges that cannot be directly aligned, such as those depicted in Figure 6.5. Most importantly, the model can be trained with more than one subgraph size $n$: For example, model **A-234** simultaneously extracts subgraphs of the size $n = [2, 3, 4]$, so that the edge probabilities of differently large subgraphs add up.



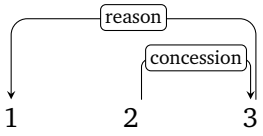

(a) Example 1: ARG




(b) Example 1: RST




(c) Example 2: ARG




(d) Example 2: RST

Figure 6.5: Two examples for common components between RST and ARG for attack- and counter-attack constructions, involving edges that cannot be directly mapped.

The collected edges of all candidate subgraphs do not necessarily connect all possible nodes. In this case, no spanning tree could be derived. We thus initialise the intermediary multi-graph as a total graph with low-scored default edges of the type *unknown*. These should only be selected by the MST algorithm when there is no other evidence for connecting two unconnected subgraphs. The number of predicted unknown edges thus serves as an indicator of the coverage of the learnt model. In the evaluation, unknown edges are interpreted as the majority relation type, i.e. as support.

Finally, we added an optional root-constraint (**+r**) to the model: It forbids outgoing edges from the node corresponding to the RST central nucleus, and therefore effectively enforces the ARG structure to have the same root as the RST tree.

### Evidence graph model

We use the evidence graph model (**EG**) to predict argumentation structures from RST. Our model differs from the previous ones in that our base classifiers are trained exclusively on a new feature set reflecting aspects of the input RST tree, and do not use any linguistic features. The segment features are shown in Figure 6.6. We distinguish three feature groups: base features including edges (EG-2), base features plus 3-node subgraph features (EG-23), and the latter plus 4-node subgraph-features (EG-234). Base classifiers for the cc, ro, and fu-level are trained on segment features. The at-level base classifier is trained on segment features for the source and the target node, as well as on relational features, shown in Figure 6.7.

As in the original model, the base classifiers perform an inner cross-validation over the training data in order to first optimise the hyperparameters of the log-linear SGD classi-

| |
|---|
| **base features incl. 2-node subgraph features:** |
| - absolute and relative position of the segment in the text |
| - binary feature whether it is the first or the last segment |
| - binary feature whether it has incoming/outgoing edges |
| - number of incoming/outgoing edges |
| - binary feature for each type of incoming/outgoing edge |
| **3-node subgraph features:** |
| - all relation chains of length 2 involving this segment |
| **4-node subgraph features:** |
| - all relation chains of length 3 involving this segment |

Figure 6.6: Segment feature sets

| |
|---|
| - direction of the potential link (forward or backward) |
| - distance between the segments |
| - whether there is a RST relation between the segments |
| - type of the RST relation between the segments or None |

Figure 6.7: Segment-pair features

fier [Pedregosa et al., 2011]. We do not optimise the weighting of the base classifiers for score combination here, because we had shown in the original experiments that an equal weighting yields competitive results (see Section 6.5.3).

### 6.8.3 Results

The evaluation results are shown in Table 6.17. All alignment models including at least subgraphs of size n=3 (A-23*) improve over the baseline (BL) in predicting the relation type (fu) and the attachment (at). Considering larger subgraphs helps even more, and it decreases the rate of unknown edges.[15] On the role level, the baseline is unbeaten. For central claim identification, the alignment model performs poorly. Adding the root constraint yields exactly the baseline prediction for the central claim, but also improves the results on all other levels, with the cost of an increased rate of unknown edges. The clear improvement over the baseline for the relation type (fu) indicates that the probability estimates of the alignment models capture the relations better than the heuristic mapping to the most frequently aligning class in the baseline. Furthermore, extraction of larger subgraphs gradually increases the result on both the fu and the at level, showing us that there are subgraph regularities to be learnt which are not captured when assuming isomorphic trees.

For the evidence graph model, we will first investigate the performance of the base classifiers (EG-bc-*), before we discuss the results of the decoder. The difference between the

[15] Note that when testing the A-234 model on training data, only very few unknown edges are predicted (less than 1%), which indicates that more data might help to fully cover all of them.

| model | cc | ro | fu | at | unknown |
|---|---|---|---|---|---|
| BL | .861 | **.896** | .338 | .649 | |
| A-2 | .578 | .599 | .314 | .650 | 10.6% |
| A-23 | .787 | .744 | .398 | .707 | 7.5% |
| A-234 | .797 | .755 | .416 | .719 | 7.0% |
| A-2345 | .794 | .762 | .424 | .721 | 6.8% |
| A-2+r | .861 | .681 | .385 | .682 | 13.9% |
| A-23+r | .861 | .783 | .420 | .716 | 11.3% |
| A-234+r | .861 | .794 | .434 | .723 | 10.8% |
| A-2345+r | .861 | .800 | .443 | .725 | 10.7% |
| EG-bc-2 | .899 | .768 | .526 | .747 | |
| EG-bc-23 | .907 | .845 | .525 | .749 | |
| EG-bc-234 | .906 | .847 | .526 | .750 | |
| EG-2 | .918 | .843 | .522 | .744 | |
| EG-23 | **.919** | .869 | .526 | **.755** | |
| EG-234 | .918 | .868 | **.530** | .754 | |

Table 6.17: Evaluation scores of all models on the gold RST trees reported as macro-avg. F1

three feature sets is most important here. Comparing the classifier that only uses the basic feature set (EG-bc-2) against the one with extra features for 3-node subgraphs (EG-bc-23), we find the greatest improvement on the argumentative role level with an extra +7.7 points macro F1 score. Central claim identification also profits with a minor gain of +0.8 points. Interestingly, the local models for function and attachment are not effected by the richer feature sets. Extending the features even to 4-node subgraphs (EG-bc-234), does not further improve the results on any level.

The evidence graph decoding models (EG-*) combine the predictions of the base classifiers to a global optimal structure. The model using the base classifiers with the smallest feature set (EG-2) already outperforms the best alignment model on all levels significantly and beats the baseline on all levels but argumentative role. We attribute this improvement to three aspects of the model: First, the learning procedure of the base classifiers is superior to that of the alignment model. Second, the base classifiers not only learn regularities between RST and ARG but also positional properties of the target structures. Finally, the joint prediction of the different levels in the evidence graph model helps to compensate weaknesses of the local models by enforcing constraints in the combination of the individual predictions: Comparing the base classifier's predictions (EG-bc-2) with the decoded predictions (EG-2), we observe a boost of +7.5 points macro F1 on the role level and a small boost of +1.9 points for central claim through joint prediction.

Adding features for larger subgraphs further improves the results: EG-23 beats EG-2 on all levels, but the improvement is significant only for role and attachment. EG-234, though, differs from EG-23 only marginally and on no level significantly. Note that the gain from joint prediction is less strong with better base classifiers, but still valuable with +2.4 points on the role level and +1.2 points for central claim.

In conclusion, the baseline model remained unbeaten on the level of argumentative role. This was already expected, as the sequence of contrastive relations in the RST tree is very likely to map to a correct sequence of proponent and opponent role assignments. On all other levels, the best results for mapping gold RST trees to fine-grained argumentation structures are achieved by the EG-23(4) model.

It is worth investigating which argumentative functions can be well recovered from the rhetoric structures using the best-scoring EG-23(4) model: Segments serving as the central claim of the argument can be predicted best with F1=0.86. Support and rebut is also predicted quite reliably with F1=0.73. Undercuts are harder to predict, but still score with an F1=0.57. The join function is challenging: The best model reaches an F1=0.41 here. For the very infrequent linked structures and example support, scores are very low, with F1=0.22 and F1=0.17 respectively. This again suggests that a much larger annotated resource would be required in order to reliably predict these structures. In conclusion, the important distinctions can be modelled quite successfully given gold RST trees, even the distinction between rebutting and undercutting attacks.

Finally, we want to compare these results against those of the previous section, where we predicted the very same EDU-segmented, fine-grained structures using the linguistic pipeline with natural-language based features. Using gold RST trees as input obviously makes the task of predicting argumentation structure easier: The RST-based results are better on every level, the largest improvement found on the role level with +13 points F1, and +8 as well as +7 points F1 for function and central claim. The attachment classification only improved by +3 points. On the level of argumentative function, we observe a strong improvement from F1=0.51 to F1=0.73 for rebuts, but undercuts and support and central claim also score around 10 points better. The very infrequent relations support by example and linked structures are also better covered using RST trees, but still predicted unreliably. Interesting is the loss when predicting join relations: Here, the NLP pipeline scored much better with F1=0.58 than the RST pipeline with F1=0.39. In general though, the NLP models of the last section achieve, depending on the level, between 85% to 96% of the performance of the gold RST tree based models here. This is noteworthy for two reasons: On the one hand, each gold RST tree already represents a deep conceptualisation of the intentional structure of the texts, a large share of which is relevant to argumentation structure. This indicates that our NLP-based models are more than just a first throw towards recognising the structure of argumentative texts. On the other hand, we should bear in mind that the RST-based models "only" achieve a labelled attachment score of 0.515.

Hence, there still remains a considerable part of argumentation that is not covered by RST structures and that is worth to be investigated in future research.

### 6.8.4 Conclusion

In this section, we compared three mechanisms to map gold RST trees to argumentation structures: A heuristic baseline that transforms RST trees to isomorphic trees with corresponding argumentative relations; a simple aligner, extracting matching subgraph pairs from the corpus and applying them to unseen structures; and the evidence graph model, which we have used in previous sections, but this time trained only on features extracted from the input RST trees. The evidence graph model achieved promising results, outperforming the other models, and we hope to have shown the versatility of this approach.

Comparing the results of the evidence graph model trained on RST trees compared to those where it was trained on natural language input, we conclude that our NLP based models are, despite all limitations, already doing a good job in decoding the structure of argumentation. However, not every aspect of argumentation is represented in rhetorical structures. It would thus be very desirable to investigate this more practically and elicit whether both models learn similar things about argumentation structures, or whether they complement one another and could be unified to provide even better predictions. This could be approached in several ways, for example by informing a potential RST parser with argumentation based features, or the other way around by augmenting a NLP-pipeline with successful features of a RST parser, or finally by integrating predictions of state-of-the-art RST parsers into the argumentation structure decoding. All of this, we have to leave for future work.

## 6.9 Conclusions

In this chapter we have presented our efforts to automatically predict the argumentation structures of the microtext corpus. We will summarise the results of the different studies, and present them in an updated comparison with the related work, see Table 6.18.

- Our aim in Section 6.3 was to approach the challenging endeavour of argumentation mining by starting small. In order to get an understanding of the difficulty of the various tasks involved, we developed local models for predicting various aspects of argumentation structure of the microtexts. We compared different machine learning algorithms and investigated the impact of features. For basic distinctions such as between argumentative roles or general functions, the results were promising, while some more fine-grained functions could not be recognised reliably. Identifying the relations between segments, especially non-adjacent links, could not be successfully

predicted in this first approach: Approaching this task as a segment-wise classification task turned out to be too restrictive. Also, the feature space only covered a context window to segments adjacent to the source. Nonetheless, we learned about interactions between certain levels, such as for example role and function, that might guide the prediction, and we observed the need to make global decisions about argumentation structure.

- In Section 6.4 we then focused exclusively on switches of argumentative role, which occur when the author mentions an objections or counter-considerations to his argument. We already achieved good results in predicting them on the microtext corpus and thus compared this with what is achievable using the same techniques on a corpus of Pro & Contra news commentaries. Although segments with opponent role were less frequent in the ProCon corpus, they still exist in two thirds of the newspaper texts. We observed that automatic recognition is more challenging on this corpus than for the microtexts, but also that both corpora differ in the distribution of some very indicative contrastive discourse markers.

- Section 6.5 developed the proposed approach of this thesis to recognise argumentation structures: the evidence graph model. Our aim was twofold: We wanted to derive globally optimal structures, and make use of the interactions between different tasks to guide prediction. Both is achieved in the evidence graph model: The predictions of a local attachment model over pairs of segments are decoded with the MST algorithm to yield the globally optimal tree configuration over the whole sequence of input segments. Secondly, the predictions of three other base classifiers for central claim identification, role and function classification are integrated into the evidence graph in such a way that all four aspects of argumentation structure can be predicted jointly, guiding each other to the best overall structure. In a detailed comparison with various baselines and a competing MST-parser pipeline, we demonstrated the superiority of this approach. The evidence graph model is, to the best of our knowledge, the first data-driven model of argumentation structure that optimises structures globally, and the first model to jointly tackle segment type classification, relation identification, and relation type classification.

- Since the initial publication of the evidence graph models, other decoders for argumentation structures have been proposed, which use ILP for deriving argumentation structures. In Section 6.6, we compare various ILP decoders against the MST-based evidence graph model. We replicate two ILP decoders and propose a novel ILP decoder with several sets of argumentation structure theoretic constraints, each of which we could demonstrate the impact of. All decoders including the evidence graph model use the same, improved local models, and are evaluated on the same corpus. The re-

sults show that the evidence graph model and the novel ILP decoder are on par, while the replicated decoders' performance falls back since they do not exploit the dialectics of argumentative roles. We thus argue that MST-based methods such as our proposed approach suffice to derive globally optimal argumentation structures. Whether ILP-based decoders can improve on this by modelling corpus- or genre-specific properties through extra constraints remains to be shown in each specific case. The top-scoring models also present the state of the art on the microtext corpus, improving over both published and replicated results of related decoders.

- In Section 6.7 we then applied the evidence graph model to more complex versions of the microtext corpus. While all previous results had been produced on structures using the ADU segmentation and the reduced relation set with only the binary distinction between support and attack, this section presented the results for all variations of increased complexity: using an underlying EDU segmentation, or the full fine-grained relation set, or even both. In order to assess the impact of both levels of complexity, we provided a detailed comparison of heuristic baseline models as a lower bound and of the evidence graph models' results. We found that EDU segmentation alone does not make the task much harder, allowing us to use the output of automatic discourse segmenters. In combination with the full relation set, however, argumentation mining becomes even more challenging. For some of the distinctions, such as between rebutting and undercutting attacks, we found promising results, while other fine-grained relation types are too hard or too underrepresented in the data to be successfully modelled.

- Finally, we let the evidence graph model predict argumentation structures not from written text, but from corresponding RST structures. These had been annotated as one of the extra layers of the microtext corpus. This experiment was reported in Section 6.8. The evidence graph models proved to yield the best results, compared with a tree-transformation baseline and a sub-graph matcher. Besides the theoretic implications of showing that RST and argumentation structures are not isomorphic "enough" to simply map them to each other, we have presented a practical model that can go further and learn to map where structures do not align. This could be used in future work in order to profit from the improvements achieved in the last two decades of RST parsing research, and to predict argumentation structures from the results of off-the-shelf RST parsers.

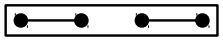
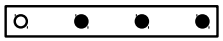
| approach | text genre | input | tasks/architecture | full structure | end-to-end |
|---|---|---|---|---|---|
| | | | BD AI AC RI RC | | |
| Mochales Palau and Moens [2009] | court decisions | sentence |  | ✓ | |
| Kang and Saint-Dizier [2014] | instructional | token |  | ✓ | |
| Lawrence et al. [2014] | phil. text | token |  | (✓) | |
| Stab and Gurevych [2014a] | student essays | ADU |  | | |
| Lawrence and Reed [2015] | AIFdb | token |  | (✓) | |
| Stab and Gurevych [2016] | essays/micro | token |  | ✓ | |
| Persing and Ng [2016a] | student essays | clause |  | ✓ | ✓ |
| Ch. 6.3 [Peldszus, 2014] | microtexts | ADU |  | | |
| Ch. 6.5 [Peldszus and Stede, 2015a] | microtexts | ADU |  | ✓ | |
| Ch. 6.6 [Afantenos et al., under review] | microtexts | ADU |  | ✓ | |
| Ch. 6.7 | microtexts | clause |  | ✓ | |
| Ch. 6.8 [Peldszus and Stede, 2016b] | microtexts | RST tree |  | ✓ | |

Table 6.18: Comparison of related work and the work presented here. For notational details, see Table 6.1.

# 7 Conclusion

In this last chapter, we summarise the work presented in this dissertation. We will discuss the results and provide an outlook for future work.

## 7.1 Contributions

In the following, we present the contributions of this thesis. We will first focus on the main contributions, and then present a list of a few novel methodological tools that have been developed in the course of this work.

**Main contributions**

- We provide an extensive literature review of theories describing the structure of discourse in general and the structure of argumentation in particular, and assess their suitability to represent and model argumentation structure in authentic, monological text.
- Based on the results of this review, we develop a scheme for annotating the structure of argumentation, and show by annotation experiments that it can be used to reliably annotate argumentation structures in text.
- We present the first parallel (German and English) corpus of short argumentative texts. Each text has been annotated with argumentation structure using the proposed scheme. Both corpus and annotations have been made publicly available.
- We provide an extensive literature review of approaches to predict argumentation structures or parts of it.
- We propose local models for automatically recognising different aspects of argumentation structure: the central claim of the text, argumentative role and argumentative function of each text segment, and argumentative relations holding between text segments.
- We propose an approach to derive globally optimal argumentation structures using the predictions of our local models: the 'evidence graph' model. It is the first data-driven model of argumentation structure that optimises structures globally, and the

first model to jointly tackle segment type classification, relation identification, and relation type classification.

- We systematically compare our results with other models that have been proposed after the initial publication of the evidence graph model. We show that our approach (and an equivalent model) are superior, because they are able to exploit more interactions between different aspects of argumentation structure. Our best models improve over published and replicated results and present the state of the art on our corpus.

**Minor contributions**

- We propose a method to cluster annotators, in order to investigate the similarities and differences among them and to identify systematic disagreements.
- We employ a group-wise stratification of classification instances, in order to ensure that the class distribution is similar among training and test-set, even though instances can only be sampled group-wise.
- We propose a dependency transformation of argumentation graphs. While the argumentation structures of our scheme are graph-theoretically rich, this (under typical circumstances) revertible transformation allows us to represent them as dependency trees. As a result, we are able to apply already established methods for structure prediction.
- We systematically simulate better base classifiers, in order to measure the influence of one improved classifier on the overall result. This way, we were able to demonstrate the impact of jointly predicting multiple aspects of argumentation structure.

## 7.2 Discussion and future work

In our conclusion in Chapter 6, we mainly emphasised what was achieved through our experiments. Now, reflecting on the whole project, we want to address possible limitations of our approach and offer paths of future work that could help overcome these.

**Scaling up for longer texts**

Our approach tests for every combinatorially possible pair of segments whether they should be attached or not. As a results, the evidence graph is a fully connected graph. The MST algorithm does a complete search in order to determine the best structure. For very long texts, this may lead to unacceptable prediction times. This has been experienced also in RST

parsing, where long texts are quite frequent, and lead to approaches that run in linear-time and built structures in a greedy fashion.

How pressing is this problem for the field of argumentation mining? Firstly, it certainly depends on the domain, whether one is to expect long argumentative texts or not. Even if long texts are typical in the target domain, one easing factor is that not every single segment of the text might be argumentatively relevant. A preprocessing step ruling out argumentatively irrelevant segments might thus significantly reduce the number of ADUs participating in the structure.

Apart from this, different solutions are possible. If the text has a document structure (e.g. composed of paragraphs or sections), one possible solution is to first predict argumentation structures for each paragraph or section, and then predict a global document-level structure using the central claims of each paragraph. The underlying assumption here is that there are no argumentative relations across paragraph boundaries except with the central claim of the paragraph, a heuristic that has been used for example in multi-paragraph student essays [Stab, 2017]. Another solution is to restrict the extraction of candidate pairs to a certain context window, allowing for example only relations to segments not further away than $n$ segments [Persing and Ng, 2016a]. This restriction has to be carefully checked against the distributions of relations in the target corpus and domain, though, in order to prevent losing coverage on long-distance dependencies. Both solutions could be implemented in the evidence graph model in a straightforward manner.

Finally, it will be worth investigating the impact of greedy structure building algorithms that run in linear time. How much accuracy is lost when relying on their incomplete search in the space of possible structures, compared to the total search of e.g. the MST algorithm? The challenge for such algorithms will be to keep track of the obligations the writer has towards the reader in presenting his argument and thus to allow structural attachments to ADUs which might still be attacked, or which still require further support.

#### Applying to other corpora

The proposed approach has so far only been applied to the microtext corpus. We would like to test it on other corpora in future work. One candidate would be the Pro & Contra commentary section from the PCC [Stede and Neumann, 2014], which we already used in some of our experiments. Unfortunately, only a few texts have been annotated with full argumentation structures, and a significant amount of work is yet to be done to reach a considerably large resource of news commentaries annotated with argumentation structures.

An immediate next step would therefore be to apply the evidence graph model to the corpus of student essays [Stab and Gurevych, 2016]. We would like to note that the level of argumentative role is not a peculiarity of the microtext corpus, as argued in that paper. The argumentative role of each segment can be derived from *any* argumentation structure

that has both supporting and attacking relations, thus likewise from the annotations of the student essays. Nonetheless, the authors are right in stressing that the proportion of argumentative attacks is much smaller in their corpus. It will be very interesting to investigate whether the dialectical view on the texts through the level of argumentative role will still pay off when applied to the student essays.

#### Identifying ADUs

One task that was only partially tackled in our experiments is ADU identification, i.e. deciding whether a certain text span is argumentatively relevant or not. The reason for this is that, by design our corpus aims to reduce the amount of argumentatively irrelevant material as far as possible. We covered one aspect of ADU identification, however, when using the more fine-grained EDU-segmentation of the corpus: The model had to decide whether an EDU forms an ADU in its own right, or has to be combined with an adjacent EDU by a join relation to form one single ADU.

A next step would thus be to consider corpora featuring argumentatively irrelevant material, such as the aforementioned corpus of Pro & Contra commentaries or the student essays corpus. For the latter, the task of ADU identification can be considered to be solved to a good degree using the methods of Stab and Gurevych [2016]. For the Pro & Contra commentaries, however, we know from our annotation experiments in Chapter 4.4 that the distinction between non-argumentative background segments and those that are argumentatively relevant is most often difficult. We thus expect more work to be done here in order to predict argumentative relevancy reliably in newswire text.

Especially in the case where determining argumentative relevancy is hard, it might be worth considering whether its prediction could be improved by the guidance from other tasks, such as ADU classification or relation identification. Consider for instance the interesting twist of the approach of Lawrence et al. [2014], where relevancy was decided as the last step in the pipeline. We would propose to make ADU identification part of the joint modelling. Persing and Ng [2016a] achieved this to an extent by combining the tasks of ADU identification and classification. In the evidence graph model, this would amount to an additional base classifier whose prediction is added to the combined edge scores, plus an attachment convention to allow MST to derive trees with irrelevant nodes. We have to leave a systematic investigation of this question to future work.

#### Weighting base classifiers

The fact that our base classifiers can receive different weightings in the evidence graph model was considered by one conference paper reviewer as problematic or at least as inelegant. As we understand it, this critique amounts to the complaint that there is yet another hyper-parameter to be tuned, making the evaluation more complex. Our intuition behind

testing different weighting was that the tasks tackled are not equally hard. If one base classifier is better, its predictions might be more trustworthy than those of other base classifier. However, it turned out in all our experiments that the differences between an optimised weighting and an equal weighting are negligible. In contrast to other approaches that used different base-classifiers without testing different weightings, we possess the empirical evidence that using an equal weighting is a reasonable choice.

#### Adding constraints for relations types

In the evidence graph model, by default all relations types are allowed to occur in any direction, at any distance. This is perfectly adequate for the different supporting and attacking relations that were used. In the setting with fine-grained EDU-segmentation, the join relation was also used, which is intended to hold only between adjacent text segments and only in backward direction. Similarly, a relation representing restatements (which we do not have in our corpus yet) would be intended to hold only between non-adjacent segments in likewise only backward direction.

In future work, we propose to restrict the prediction of these relation types to fulfil the intended constraints. For the ILP-based decoder presented in Chapter 6.6, this could be achieved by adding special constraints for these relation types. In the evidence graph model it would be equally straightforward to achieve this by constraining the population of the graph with edges of this type accordingly.

#### Using lexical knowledge in the lack of linguistic signals

One problem across all approaches to argumentation mining is the question of how to increase coverage on relations which are not explicitly signalled, e.g. by a discourse marker or other highly indicative linguistic cues.

Different features have been proposed in the literature which all aim at modelling lexical knowledge in one or the other way; for instance heuristically by extracting word-pairs or brown clusters; indirectly by relying on similarity scores based on word-overlap, TF-IDF, LDA topics, averaged word vector representations, or the scores of textual entailment systems; or more explicitly by exploiting lexical resources such as WordNet or FrameNet. Some of these features yielded valuable improvements – we refer the interested reader to the literature review (6.1).

In our experiments, we have not yet studied in depth the impact of such features. But contrary to the findings in related work, when we did test them, they usually could not contribute a significant improvement. We experimented with different similarity scores in our initial study on local models (6.3) and with averaged word vector representations in our improved local models (6.6), both without success.

One possible explanation is probably connected to the characteristics of our corpus. By design, the texts are very compact and the space restriction might enforce writers to be more explicit about their intention within each segment. Our corpus is therefore relatively rich in explicit linguistic cues. As a consequence, the aforementioned features may have less impact in our corpus than in less restricted corpora. Another explanation regarding in particular the minor impact of word-embeddings and similarity scores based on them, was that we have used pre-trained word-embeddings and did not learn the vector representations specifically for our task.

In future work, we would like to study the use of such features in more detail, and conduct a systematic feature analysis on different corpora. To this end, it would also be wise to review in more detail the literature on shallow discourse parsing that is concerned with the recognition of implicit discourse relations.

#### Learning structure, globally, jointly

The evidence graph model relies on a decoding approach, where different local models are trained, whose predictions are then integrated and decoded by a global model. One important feature of our approach is that different base classifiers jointly *predict* a structure through the interactions inherent in the combination of individual predictions. Nevertheless, these base classifiers are trained separately. One line of future research could investigate the impact of jointly *learning* the different classifiers. Another would be to employ structured learning, as advised by Moens [2013].

Note that the mstparser model we used as a competitor in Chapter 6.5 implemented a structured learning approach, which yielded competitive scores for the unlabelled structures. For labelled structures, however, the evidence graph model was superior, even though we tried to make all the evidence of the various base classifiers available to the mstparser model as well. A possible reason is that the mstparser model was still lacking a loss-function that included all aspects relevant to argumentation structures. As a consequence, the model optimised the predicted structure only for attachment, but not at the same time for argumentative role, function, and directly for picking a suitable central claim. To the best of our knowledge, no such loss function that adequately covers all aspects relevant for argumentation structure parsing has yet been proposed. We thus consider a definition of such a loss function as a first step towards using more powerful structured learning approaches.

#### Interfacing argumentation and discourse structure

An abundance of possible future work could address the interface between argumentation and discourse structure. Our corpus, which has been annotated not only with argumentation structures but also according to RST and SDRT, allows for the first time an empirical analysis –be it qualitatively or quantitatively– of the interrelations between these theories.

Furthermore, we have shown practically in Chapter 6.8 that RST and argumentation structures are not isomorphic 'enough' to simply map them to each other, and have thus applied the evidence graph model to learn to map where structures do not align. This was done successfully on the gold rhetoric structures, but we had to leave experiments with the results of off-the-shelf RST parsers for future work. We look forward to investigating the degree to which their output could be used to effectively model argumentation structures.

Finally, it would be worthwhile studying whether the predictions or argumentation structure from text and from discourse structures would complement one another and could be unified to provide even better predictions. This could be achieved by testing the benefit of argumentation structures in discourse parsing or of discourse structures in argumentation structure prediction, either by sharing features that have proven successful in the corresponding tasks, or by using the predicted structures themselves. All of this we have to set aside for future work.

# A Guidelines for the annotation of argumentation structure

The following appendix section contains the original annotation guidelines used in the annotation experiments presented in Chapter 4.2. An extended version of these guidelines was later published as [Peldszus et al., 2016].

# Richtlinien zur Annotation von Argumentationsstruktur

**Andreas Peldszus & Saskia Warzecha**

Die vorliegenden Annotationsanweisungen sind auf Texte anzuwenden, in denen ein Autor für oder gegen eine These argumentiert und dienen dem Zweck einer vollständigen schematischen Diagramm-Abbildung dieser Argumentation.

Der Annotationsprozess gliedert sich in drei Schritte. Gegeben ist ein schon segmentierter Text, für den zuerst die Gesamtthese bestimmt wird. Dann ist für jedes einzelne Textsegment festzustellen, welche argumentative Stimme es repräsentiert. Schließlich wird die argumentative Funktion das Segments und der genaue Anknüpfungspunkt an die bestehende Struktur bestimmt. Jeder der genannten Schritte wird im Folgenden einzeln beschrieben.

Resultat des Annotationsprozesses ist eine Argumentationsstruktur: ein Graph, dessen Knoten den Segmenten des Textes entsprechen und deren Verküpfungen den argumentativen Zusammenhang zwischen den Segmenten repräsentieren.

## Schritt 1: Identifikation der Gesamtthese

**Die Gesamtthese des Textes wird ermittelt.**

Zunächst muss von den Segmenten des Textes dasjenige ausgewählt werden, welches am ehesten die Gesamtthese des Textes wiedergibt. Es soll also die „Grundaussage“ des Textes gefunden werden, von der der Autor den Leser letztlich überzeugen möchte. Diese Gesamtthese steht für sich selbst, während alle anderen Textsegmente dazu dienen, durch Stützung und Anfechtung die Überzeugungskraft der Gesamtthese zu verstärken. Sie kann prinzipiell an jeder Stelle eines argumentativen Textes auftreten, sei es gleich zu Beginn, in der Mitte oder erst ganz zum Schluss.

Bei der Gesamtthese handelt es sich oft um Handlungsanweisungen oder -empfehlungen („Wir/Man/Peter sollte X tun.“). Genauso können aber auch andere Texthandlungen, wie Bewertungen („X ist schlecht.“) oder Vermutungen („Wahrscheinlich ist X der Fall.“) etc als Gesamtthese fungieren.

## Schritt 2: Zuweisung der argumentativen Stimme

**Für jedes Segment wird bestimmt, wessen Stimme es repräsentiert.**

Die Gesamtthese des Textes kann vom Autor kontrovers diskutiert werden. Er präsentiert eventuell nicht nur Gründe, die für die Annahme der These sprechen, sondern zieht auch mögliche Gegenargument in Betracht. Dies kann man analog zu einem Streitgespräch

verstehen, in dem ein Proponent eine These vertritt und diese gegen die Angriffe eines Opponenten verteidigt.

Der nächste Schritt besteht deswegen darin für jedes Segment festzustellen, ob dessen Aussage im analogen Streitgespräch vom Verteidiger der Gesamtthese, also vom Proponenten vorgebracht werden würde, oder vom Opponenten, der die These und ihre Begründung kritisch hinterfragt.

In Beispiel (1) lässt der Autor eine mögliche Gegenstimme („Aber die Kleidung ist so günstig!“) zu Wort kommen, also würde das zweite Segment der Stimme des Opponenten zugeordnet werden, während das erste Segment wie jede Gesamtthese die Stimme des Proponenten repräsentiert.

(1) H&M sollte man nicht unterstützen [1], auch wenn die Preise für Jeans und T-Shirts verlockend sind [2].

Um diese Unterscheidung im Diagramm deutlich zu machen, repräsentieren wir jene Segmente, die zur Stimme des Proponenten gehören, in einem Kreis und -falls es solche gibt- jene des Herausforderers in einem Kasten.

Bisher existieren in der angestrebten Argumentationssturktur nur Knoten (unterschiedlicher Art), aber die Knoten sind noch unverbunden. Die letzten beiden Annotationsschritte bestimmen deshalb die Art der Verbindung (also die Sorte des Pfeils) und die Stelle im Graphen, an die sich ein Knoten anschließt.

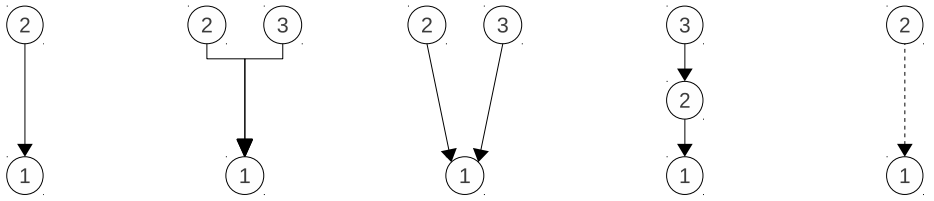
## Schritt 3: Zuweisung der argumentativen Funktion und des Bezugspunkts

**Für jedes Segmente wird bestimmt, welche Funktion sie haben und wie sie sich in die bestehende Struktur integrieren.**

Es gibt im Argument zwei elementare Funktionen für ein Segment: „stützen“ und „anfechten“. Die Gesamtthese ist hiervon ausgeschlossen, schließlich dienen alle Textsegmente der zu ihr führenden Argumentation. Jedes Segment sollte nur eine Funktion haben, d.h. im Graphen geht von jedem Knoten maximal ein Pfeil ab.

### Stützen

Es gibt mehrere Möglichkeiten, im Argument etwas zu stützen. Allen Möglichkeiten ist es aber gemein, dass das, was gestützt wird, durch das, was stützt, glaubhafter gemacht werden soll. Die Stützung liefert eine Begründung für die Annahme einer Aussage, sie soll deren argumentative Kraft vergrößern. Indem der Autor ein Segment *A* durch ein anderes *B* stützt, beantwortet er indirekt die Frage des Lesers „Warum sollte ich *A* glauben/annehmen?“. Einen ersten Eindruck, ob es sich bei dem Segment *B* um eine Stützung von *A* handelt, gibt




(a) einfache Stützung (b) gemeinsame Stützung (c) mehrfache Stützung (d) serielle Stützung (e) Stützung durch Beispiel

Abbildung A.1: Stützungsrelationen.

daher der warum-Test: Wenn „*A*. - Warum? - Weil *B*." kein geglückter Mini-Diskurs ist, dann ist es auch unwahrscheinlich, dass *B* eine Stützung von *A* ist.

**Einfache Stützung:** Ein einfaches Beispiel für eine Stützung wäre (2):

(2) Wir sollten das Gebäude abreißen [1], denn es ist völlig asbestverseucht [2].

Die Stützungsrelation wird durch einen normalen Pfeil repräsentiert. Der Argumentationsstrukturgraph für dieses kurze Beispiel ist in Abbildung 1a dargestellt. Es sei darauf hingewiesen, dass die Reihenfolge von dem stützenden und dem gestützen Segment auch andersherum sein kann. So in Beispiel (2'), wo erst das stützende Segment genannt wird und dann das dadurch gestützte. Im Graph wäre der Pfeil zwischen Segment 1 und 2 dann entsprechend andersherum.

(2') Das Gebäude ist völlig asbestverseucht [1]. Wir sollten es daher abreißen [2].

**Gemeinsame Stützung:** Neben dieser einfachen Stützung gibt es auch den komplexeren Fall einer gemeinsamen Stützung. Hier können zwei Segmente einzeln für sich genommen nicht zur Stützung einer Aussage beitragen, wohl aber gemeinsam wenn sie beide der Fall sind. Beispiel:

(3) Wir sollten das Gebäude abreißen [1], denn es ist völlig asbestverseucht [2], und verseuchte Gebäude müssen abgerissen werden [3].

Das Argument funktioniert nur, wenn beide Prämissen gleichzeitig wahr sind: Wenn verseuchte Gebäude nicht abgerissen werden müssten, spräche eine vorhandene Verseuchung nicht für den Abriss. Andersherum: Wenn verseuchte Gebäude abgerissen werden müssen, eines aber nicht verseucht ist, muss es auch nicht abgerissen werden. Im Graphen wird eine solche Struktur durch ein Stützungspfeil mit mehreren Startpunkten aber einem Zielpunkt angezeigt (siehe Abbildung 1b).

Es ist zu bemerken, dass die Regel, die hier im dritten Segment explizit genannt ist, im vorherigen Beispiel (2) implizit bleibt und daher mitangenommen werden muss. In der Annotation werden aber nur explizit in Segmenten ausgedrückte Prämissen berücksichtigt.

**Mehrfache Stützung:** Von der gemeinsamen Stützung unterschieden wird der Fall, in dem mehrere Prämissen unabhängig voneinander stützen. Dies ist Beispiel (4) der Fall:

(4) Wir sollten das Gebäude abreißen [1], denn es ist völlig asbestverseucht [2], und die Anwohner in der Nachbarschaft haben es sowieso nie haben wollen [3].

Selbst wenn das Gebäude nicht asbestverseucht ist, wäre die Unbeliebtheit bei den Anwohnern noch ein (mehr oder wenig starker) Grund für den Abriss. In der Argumentationsstruktur wird diese Konstellation folglich durch zwei einzelne Stützungspfeile repräsentiert (siehe Abbildung 1c).

**Serielle Stützung:** Natürlich sind auch andere Anknüpfungspunkte für ein stützendes Segment möglich. Das Abrissbeispiel könnte auch durch ein drittes Segment fortgesetzt werden, welches nicht die Gesamtthese stützt, sondern die Zwischenthese. In Beispiel (5) stützt der Proponent die Glaubhaftigkeit des zweiten Segments, indem er auf das Ergebnis der Expertenkommission verweist.

(5) Wir sollten das Gebäude abreißen [1], denn es ist völlig asbestverseucht [2]. Die Expertenkommission bescheinigte eine beträchtliche Kontamination [3].

Auf diesem Weg ergibt sich rekursiv eine serielle Stützungsstruktur, wie sie in Abbildung 1d zu sehen ist.

**Stützung durch Beispiele:** Ein besonderer Fall von Stützung liegt vor, wenn eine These dadurch gestärkt wird, dass auf ein Beispiel verwiesen wird:

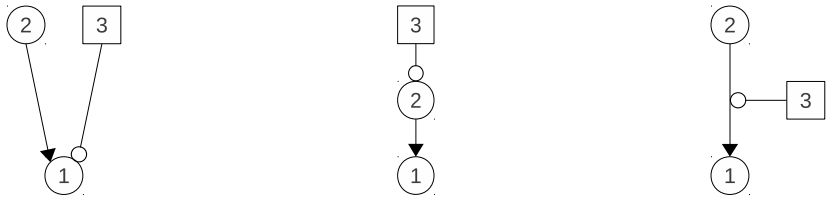
(6) Eine Bürgerinitiative kann die lokalen Autoritäten zwingen ein Gebäude abzureißen [1]. So hat es in München eine Gruppe geschafft den Bürgermeister zum Abriss eines unansehnlichen, leerstehenden Bürogebäude zu bewegen [2].

Diese spezielle Form von der Argumentation wird im Graphen durch einen gestrichelten Stützungspfeil dargestellt (siehe Abbildung 1e).

Sowohl der Proponent als auch ein potenzieller Herausforderer kann sich einer Stützung seiner eigenen Behauptungen bedienen, es sind an dieser Stelle nicht alle Kombinationen abgebildet. Dieselben Stützungs-Relationen, die für den Proponenten beschrieben wurden, sind also auch zwischen den Segmenten des Opponenten möglich, z.B. wenn dieser für seinen Einwand noch zusätzliche Gründe ins Spiel bringt. Man sollte aber beachten, dass sowohl der Proponent wie auch der Opponent ausschließlich ihre jeweils eigenen Segmente stützen werden.

### Anfechten

Entsprechend der Funktionsweise einer Stützung soll mit einer Anfechtung das Gegenteil erreicht werden: Das, was angefochten wird, soll durch die Anfechtung widerlegt oder in




(a) Rebutter gegen die Konklusion

(b) Rebutter gegen die Prämisse

(c) Undercutter gegen die Relation

Abbildung A.2: Angriffe des Opponenten auf die Argumentation des Proponenten.

seiner argumentativen Kraft geschwächt werden. Alle Anfechtungen werden im Diagramm als Pfeil mit Kugelspitze dargestellt. Es wird zwischen zwei Sorten von Anfechtung unterschieden: „Rebutter“ greifen Aussagen an, „Undercutter“ greifen Beziehungen zwischen Aussagen an.

**Rebutter:** Ein „Rebutter“ ficht eine Aussage an. Er behauptet also, dass sie aus bestimmten Gründen nicht gilt. So im folgenden Beispiel:

(7) Wir sollten das Gebäude abreißen [1], denn es ist völlig asbestverseucht [2]. Andererseits hat man eine sehr gute Aussicht vom Dach [3].

Die in Segment 3 angeführte gute Aussicht vom Dach spricht *gegen* einen Abriss des Gebäudes. Für einen entsprechenden Argumentgraph siehe Abbildung 2a. Ein Rebutter kann unter Umständen auch als Argument *für die Negation* der angefochtenen Aussage verstanden werden: Die gute Aussicht stützt die Aufforderung, das Gebäude nicht abzureißen.

Wie auch bei den Stützungen, gibt verschiedene mögliche Anknüpfungspunkte. Während im vorherigen Beispiel (7) die Konklusion angefochten wurde, wird im folgenden Beispiel (8) die Prämisse des Arguments attackiert, siehe Abbildung 2b.

(8) Wir sollten das Gebäude abreißen [1], denn es soll völlig asbestverseucht sein [2]. Aber eigentlich hat noch niemand eine genaue Einschätzung des Grads der Kontamination [3].

Die Tatsache, dass es noch keine genaue Zahlen über den Grad der Kontamination gibt, spricht gegen die Vermutung, dass es völlig verseucht ist.

**Undercutter:** Ein „Undercutter“ ficht im Gegensatz zum Rebutter nicht die Gültigkeit einer Aussage an, sondern eine Beziehung zwischen Aussagen, z.B. eine Stützungsrelation.

(9) Wir sollten das Gebäude abreißen [1], denn es ist völlig asbestverseucht [2]. Allerdings könnte man es auch sanieren [3].

In diesem Beispiel wird eine Ausnahme als Gegenargument anführt. Es wird weder angefochten, dass das Gebäude asbestverseucht ist, noch wird darüber eine Aussage getroffen,

dass man es abreißen sollte, sondern es wird die Schlussfolgerung von Verseuchung zum Abriss angegriffen. Nur weil das Gebäude verseucht ist, muss man es ja nicht gleich abreissen. Man könnte es ja auch sanieren. Ein Undercutter wird im Argumentdiagramm folglich Angriffspfeil gegen einen anderen Pfeil dargestellt, siehe Abbildung 2c.

Ob ein Einwand des Opponenten nun die Gültigkeit einer vom Proponent vorgebrachten Konklusion anficht (Rebutter), oder durch das Anführen einer Ausnahme die Gültigkeit der Schlussfolgerung von Prämisse zu Konklusion anficht (Undercutter), ist nicht immer offensichtlich und dies zu entscheiden erfordert eine genaue Prüfung. Mitunter ist es dabei hilfreich zu testen, wie geglückt die Anfechtung ist, wenn die Prämisse nicht da wäre. Ein Undercutter ergibt nur Sinn, wenn es ein Schlussfolgerung gibt, die er untergräbt. Lässt man in Beispiel (7) die Prämisse weg (Segment 2), so ist die Anfechtung immer noch wirkunsgvoll. Ließe man die Prämisse in Beispiel (9) weg, ergibt die Anfechtung keinen Sinn mehr, weil sie abhängig von der Prämisse ist. Wenn man die Prämisse weglassen kann, liegt also wahrscheinlich eher ein Rebutter gegen die Konklusion vor als ein Undercutter.

### Anfechtungen erwidern

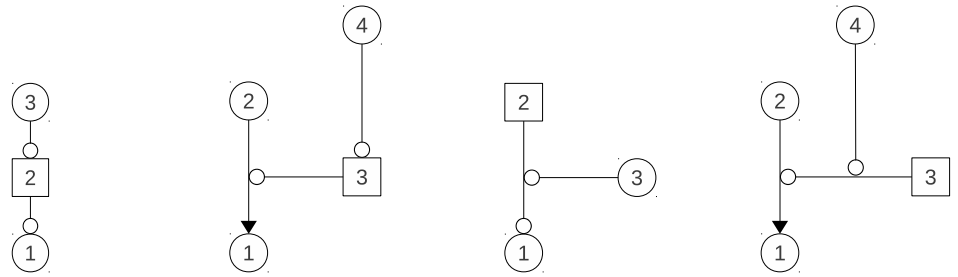
Bis zu diesem Punkt haben wir nur Anfechtungen des Opponenten gesehen, der sich gegen die Argumente des Proponenten wendet. Natürlich setzt sich der Proponent auch zur Wehr und verteidigt seine Argumente, indem er im Gegenzug die Angriffe des Opponenten anficht. Dabei ergeben sich folgende Kombinationen:

**Einen Rebutter rebutten.** Der Opponent hatte einen Grund gegen eine Aussage vorgebracht. Der Proponent entkräftet diesen Einwand, indem er widerum einen Grund gegen den Einwand vorbringt. So wird in Beispiel (10) der Mangel an gesichteten Touristengruppen vorgebracht um zu zeigen, dass das Gebäude keine Touristenattraktion ist. Für den entsprechenden Graph siehe Abbildung 3a.

(10) Wir sollten das Gebäude abreißen [1], auch wenn es eine Touristenattraktion sein soll [2]. Ich hab da jedenfalls noch nie Touristengruppen gesehen [3].

**Einen Undercutter rebutten.** Der Opponent hatte durch seinen Hinweis auf eine Ausnahmebedingung eine Schlussfolgerung angefochten. Im Beispiel (11): Wenn die Ausnahmebedingung gelten würde und Asbest harmlos wäre, dann müsste man ein verseuchtes Gebäude nicht abreißen. Der Proponent kann aber zeigen, dass Ausnahmebedingung nicht erfüllt ist, indem er auf die fragwürdige Informationsquelle hinweist. Für den entsprechenden Graph siehe Abbildung 3b.

(11) Wir sollten das Gebäude abreißen [1], denn es soll völlig asbestverseucht sein [2]. Zwar sagt eine neue wissenschaftliche Studie, dass Asbest harmlos ist [3]. Aber das ist ja wohl eine Zeitungs-Ente [4].




(a) Rebutter gegen Rebutter (b) Rebutter gegen Undercutter (c) Undercutter gegen Rebutter (d) Undercutter gegen Undercutter

Abbildung A.3: Erwiderung des Proponenten auf die Anfechtungen des Opponenten.

**Einen Rebutter undercutten.** Der Opponent hatte einen Grund gegen eine Aussage vorgebracht, bzw. einen Grund für die Negation der Aussage. Im Gegenzug zeigt der Proponent, dass der vermeintliche Grund diese Schlussfolgerung gar nicht erlaubt. Im Beispiel (12) kann der Proponent hinnehmen, dass das Gebäude tatsächlich eine Touristenattraktion ist. Aber er zeigt, dass dies kein Grund gegen den Abriss ist, weil die Ausnahmebedingung erfüllt ist, dass der Neubau eine noch größere Attraktion sein wird. Für den Graphen siehe Abbildung 3c.

(12) Wir sollten das Gebäude abreißen [1], auch wenn es eine Touristenattraktion sein soll [2]. Die werden bestimmt eine neue und größere Attraktion dahin bauen [3].

**Einen Undercutter undercutten.** Der Opponent hatte durch seinen Hinweis auf eine Ausnahmebedingung eine Schlussfolgerung angefochten. Der Proponent erwidert, indem er für die Ausnahme selbst wieder eine Ausnahme findet. Im Beispiel (13) zeigt der Proponent, dass die Möglichkeit eines Abrisses des Gebäudes wegen des zu hohen Preises irrelevant ist. Er prüft also gar nicht erst, ob die Ausnahmebedinung nun gilt oder nicht, ob eine Sanierung technisch möglich ist oder nicht, sondern er verwirft diesen Möglichkeit von vornherein als unrealisierbar. Für den entsprechenden Graphen siehe Abbildung 3d.

(13) Wir sollten das Gebäude abreißen [1], denn es soll völlig asbestverseucht sein [2]. Zwar könnte man es sanieren [3], aber das wäre viel zu teuer [4].

## Schematische Übersicht der Schritte im Annotationsprozess

**Vorbereitung:**

1. Den Text einmal komplett lesen.
2. Die Gesamtthese ermitteln.

**Für jedes einzelne Segment bestimmen:**

1. Wessen Stimme:
    - Proponent
    - Opponent
2. Welche Funktion und welcher Bezugspunkt:
    - Stützung
        - einfache Stützung
        - gemeinsam Stützung
        - Beispielsstützung
    - Anfechtung
        - Rebuttal (gegen Aussagen)
        - Undercutter (gegen Relationen zwischen Aussagen)

**Überprüfung:**

- dass der Knoten, zu dem alles führt, die ermittelte Gesamtthese ist,
- dass von jedem Knoten max. ein Pfeil abgeht,
- dass Proponent und Opponent jeweils nur ihre eigenen Knoten stützen und nur die des jeweils anderen anfechten.

# Bibliography


Satomi Adachi-Bähr. *Kontrastive Analyse von Gliederungsprinzipien in argumentativen schriftlichen Texten im Deutschen und Japanischen*, volume 20, Ausgabe 1 of *Arbeitspapiere und Materialien zur deutschen Sprache*. Institut für Deutsche Sprache, Mannheim, 2006.

Stergos Afantenos and Nicholas Asher. Testing SDRT's Right Frontier. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1–9, Beijing, China, August 2010. Coling 2010 Organizing Committee.

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cecile Fabre, Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paul Pery-Woodley, Laurent Prevot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal, 2015.

Stergos Afantenos, Andreas Peldszus, Jérémy Perret, and Manfred Stede. Comparing Decoding Mechanisms for Parsing Argumentative Structures. *Argument & Computation*, under review.

AIFdb. AraucariaDB, 2016. URL `http://www.arg.dundee.ac.uk/aif-corpora/zip/araucaria`. Last accessed online: 2016-11-23; md5sum of the zipfile 479b5b2be57320e2aebedf4397620cc1.

Sergio J. Alvarado. *Understanding editorial text: a computer model of argument comprehension*. Kluwer, Boston, 1990.

Sergio J. Alvarado, Michael G. Dyer, and Margot Flowers. Editorial Comprehension in OpEd through Argument Units. In *Proceedings of the Fifth National Conference on Artificial Intelligence, August 11-15, 1986 Philadelphia*, pages 250–256, Menlo Park, California, 1986. The AAAI Press.

Jaime Arguello and Kyle Shaffer. Predicting speech acts in MOOC forum posts. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, 2015.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December 2008.

Nicholas Asher. *Reference to abstract objects in discourse*. Kluwer, Dordrecht, 1993.

Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, Cambridge, 2003.

John L. Austin. *How to do things with words*. Harvard University Press, Cambridge/MA, 1975.

Moshe Azar. Argumentative text as rhetorical structure: An application of rhetorical structure theory. *Argumentation*, 13:97–114, 1999.

Sebastian Bachmann and Michael Brandt. Argumentative Zoning bei Kommentaren. Term paper in the 2005 summer semester course Prof. M. Stede: "Textstruktur", 2005.

Jason Baldridge, Nicholas Asher, and Julie Hunter. Annotation for and robust parsing of discourse structure on unrestricted texts. *Zeitschrift für Sprachwissenschaft*, 26:213–239, 2007.

Monroe C. Beardsley. *Practical Logic*. Prentice-Hall, New York, 1950.

Maria Becker, Alexis Palmer, and Anette Frank. Argumentative texts and clause types. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 21–30, Berlin, Germany, August 2016. Association for Computational Linguistics.

Farah Benamara, Nicholas Asher, Yvette Yannick Mathieu, Vladimir Popescu, and Baptiste Chardon. Evaluation in Discourse: a Corpus-Based Study. *Dialogue and Discourse*, 7(1): 1–49, 2016. doi: 10.5087/dad.2016.101.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1263.

Heike Bieler, Stefanie Dipper, and Manfred Stede. Identifying formal and functional zones in film reviews. In *Proceedings of the Eighth SIGDIAL Workshop*, Antwerp, 2007.

Bernd Bohnet. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 89–97, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Wauter Bosma. Query-based summarization using rhetorical structure theory. In Ton van der Wouden, Michaela Poß, Hilke Reckman, and Crit Cremers, editors, *Computational Linguistics in the Netherlands 2004, Selected Papers from the Fifteenth CLIN Meeting, December 17, Leiden Centre for Linguistics*. LOT Utrecht, 2004.

Ulrik Brandes, Markus Eiglsperger, Ivan Herman, Michael Himsolt, and M.Scott Marshall. Graphml progress report structural layer proposal. In Petra Mutzel, Michael Jünger, and Sebastian Leipert, editors, *Graph Drawing*, volume 2265 of *Lecture Notes in Computer Science*, pages 501–512. Springer Berlin Heidelberg, 2002. ISBN 978-3-540-43309-5.

Margareta Brandt and Inger Rosengren. Zur Illokutionsstruktur von Texten. *Zeitschrift für Literaturwissenschaft und Linguistik*, 86:9–51, 1992.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yakorska. A model for processing illocutionary structures and argumentation in debates. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Jill Burstein and Daniel Marcu. A machine learning approach for identification thesis and conclusion statements in student essays. *Computers and the Humanities*, 37(4):455–467, 2003.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*, pages 85–112. Kluwer, Dordrecht, 2003.

Gavin C. Cawley and Nicola L.C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, August 2010. ISSN 1532-4435.

Carlos Chesñevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. Towards an argument interchange format. *The Knowledge Engineering Review*, 21(04):293–316, 2006.

Yoeng-Jin Chu and Tseng-Hong Liu. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400, 1965.

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, pages 840–850, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Robin Cohen. *A computational model for the analysis of arguments*. PhD thesis, University of Toronto, Toronto, Canada, 1983.

Robin Cohen. Analyzing the structure of argumentative discourse. *Computational Linguistics*, 13(1-2):11–24, 1987a.

Robin Cohen. Interpreting clues in conjunction with processing restrictions in arguments and discourse. In Kenneth D. Forbus and Howard E. Shrobe, editors, *AAAI*, pages 528–533. Morgan Kaufmann, 1987b.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. Learning to classify email into "speech acts". In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 309–316, Barcelona, Spain, July 2004. Association for Computational Linguistics.

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

Laurence Danlos. Comparing RST and SDRT Discourse Structures through Dependency Graphs. In *Proceedings of the Workshop on Constraints in Discourse (CID)*, Dortmund/Germany, 2005.

Isin Demirsahin, Adnan Ozturel, Cem Bozsahin, and Deniz Zeyrek. Applicative structures and immediate discourse in the turkish discourse bank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 122–130, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, December 2006. ISSN 1532-4435.

Gerhard Dillmann. *Sprechintentionen in deutschen und japanischen Zeitungskommentaren. Illokutionstypologie und kontrastive Analysen*. dissertation, Albert-Ludwigs-Universität Freiburg im Breisgau, 2008.

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, September 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)*, pages 2249–2255, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Jack Edmonds. Optimum Branchings. *Journal of Research of the National Bureau of Standards*, 71B:233–240, 1967.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008. ISSN 1532-4435.

Adam Faulkner. Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. In *Florida Artificial Intelligence Research Society Conference*, pages 174–179, 2014.

Valéria D. Feltrim, Simone Teufel, Maria Graças V. das Nunes, and Sandra M. Aluísio. Argumentative zoning applied to critiquing novices' scientific abstracts. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 233–246. Springer Netherlands, 2006. doi: 10.1007/1-4020-4102-0_18.

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, pages 987–996, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 60–68, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

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)*, pages 511–521, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Peter Filzmoser, Bettina Liebmann, and Kurt Varmuza. Repeated double cross validation. *Journal of Chemometrics*, 23(4):160–171, 2009. ISSN 1099-128X. doi: 10.1002/cem.1225.

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*, pages 49–54, Sofia, Bulgaria, 2013.

Kate Forbes-Riley, Fan Zhang, and Diane Litman. Extracting pdtb discourse relations from student essays. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 117–127, Los Angeles, September 2016. Association for Computational Linguistics.

James B. Freeman. *Dialectics and the Macrostructure of Argument*. Foris, Berlin, 1991.

James B. Freeman. *Argument Structure: Representation and Theory*. Argumentation Library (18). Springer, 2011.

Gerald Gamrath, Tobias Fischer, Tristan Gally, Ambros M. Gleixner, Gregor Hendel, Thorsten Koch, Stephen J. Maher, Matthias Miltenberger, Benjamin Müller, Marc E. Pfetsch, Christian Puchert, Daniel Rehfeldt, Sebastian Schenker, Robert Schwarz, Felipe Serrano, Yuji Shinano, Stefan Vigerske, Dieter Weninger, Michael Winkler, Jonas T. Witt, and Jakob Witzig. The SCIP optimization suite 3.2. Technical Report 15-60, ZIB, Takustr.7, 14195 Berlin, 2016.

Maria Paz Garcia Villalba and Patrick Saint-Dizier. Some facets of argument mining for opinion analysis. In *COMMA 2012: Proceedings of the 4th International Conference on Computational Models of Argument*, Vienna, 2012. IOS Publishing.

Anna Gastel, Sabrina Schulze, Yannick Versley, and Erhard Hinrichs. Annotation of implicit discourse relations in the TüBa-D/Z treebank. In Hanna Hedeland, Thomas Schmidt, and Kai Wörner, editors, *Multilingual Resources and Multilingual Applications - Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*, Arbeiten zur Mehrsprachigkeit (Folge B, Nr. 96, 2011), pages 99–104. Universität Hamburg, Hamburg, 2011.

Jeroen Geertzen and Harry Bunt. Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, SigDIAL '06, pages 126–133, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, T. Raymond Ng, and Bita Nejat. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1168.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany, August 2016. Association for Computational Linguistics.

Thomas F. Gordon. An overview of the carneades argumentation support system. In Chris Reed, editor, *Dialectics, dialogue and argumentation. An examination of Douglas Walton's theories of reasoning and argument*, pages 145–156. King's College London, London, 2010.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news, blogs, and the social web. *International Journal on Artificial Intelligence Tools*, 24(05):1540024:1–22, 2015. doi: 10.1142/S0218213015400242.

Trudy Govier. *A Practical Study of Argument*. Wadsworth, Belmont, CA, 1st edition, 1985.

Trudy Govier. More on counter-considerations. In *Proceedings of International Conference of the Ontario Society for the Study of Argumentation (OSSA)*, pages 1–10, Windsor/Ontario, 2011.

Nancy Green. Annotating evidence-based argumentation in biomedical text. In *Proceedings of the IEEE Workshop on Biomedical and Health Informatics*, 2015.

Nancy L. Green. Representation of argumentation in text with rhetorical structure theory. *Argumentation*, 24:181–196, 2010. ISSN 0920-427X. doi: 10.1007/s10503-009-9169-4.

Günther Grewendorf. Argumentation in der Sprachwissenschaft. *Zeitschrift für Literaturwissenschaft und Linguistik*, 10(38/39):129–150, 1980.

Barbara Grosz and Candace Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

Yufan Guo, Roi Reichart, and Anna Korhonen. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 928–937, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

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

Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, 2017.

Ben Hachey and Claire Grover. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345, 2006.

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, November 2009. ISSN 1931-0145.

Saidul Kazi Hasan and Vincent Ng. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356. Asian Federation of Natural Language Processing, 2013.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd edition, 10th printing edition, 2013.

Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, Las Cruces/NM, 1994.

Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

Hugo Hernault, Hemut Prendinger, David duVerle, and Mitsuru Ishizuka. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3): 1–33, 2010.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

Jerry Hobbs. Coherence and coreference. *Cognitive Science*, 3:67–90, 1979.

Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, 2015.

Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1250–1259, Singapore, August 2009. Association for Computational Linguistics.

Yangfeng Ji and Jacob Eisenstein. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435, september 2015.

Hans Kamp. A theory of truth and semantic representation. In Martin J. B. Stokhof, Jeroen A. G. Groenendijk, and Theo M. V. Janssen, editors, *Formal Methods in the Study of Language*, pages 277–322. Matematisch Centrum, Amsterdam, 1981.

Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Kluwer, Dordrecht, 1993.

Juyeon Kang and Patrick Saint-Dizier. A discourse grammar for processing arguments in context. In *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 43–50. IOS Press, 2014.

Manfred Kienpointner. *Argumentationsanalyse*, volume 56 of *Innsbrucker Beiträge zur Kulturwissenschaft*. Verlag des Instituts für Sprachwissenschaft, Innsbruck, 1983.

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*, pages 1–11, Denver, CO, June 2015. Association for Computational Linguistics.

Wolfgang Klein. Argumentation und Argument. *Zeitschrift für Literaturwissenschaft und Linguistik*, 10(38/39):9–56, 1980.

Josef Kopperschmidt. *Methodik der Argumentationsanalyse*. Frommann-Holzboog, Stuttgart, 1989.

Klaus Krippendorff. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA, 1980.

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*, pages 76–81. Digital Government Society of North America, 2007. ISBN 1-59593-599-1.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorial data. *Biometrics*, 33(1):159–174, March 1977.

John Lawrence and Chris Reed. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, Denver, CO, June 2015. Association for Computational Linguistics.

John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. Aifdb: Infrastructure for the argument web. In Bart Verheij, Stefan Szeider, and Stefan Woltran, editors, *Computational Models of Argument - Proceedings of COMMA 2012, Vienna, Austria, September 10-12, 2012*, volume 245 of *Frontiers in Artificial Intelligence and Applications*, pages 515–516. IOS Press, 2012.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Alan Lee, Rashmi Prasad, Aravind Joshi, and Nikhil Dinesh. Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax? In *Proceedings*

*of the 5th International Workshop on Treebanks and Linguistic Theories*, page 12, Prague, Czech Republic, Dezember 2006.

Alan Lee, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Departures from tree structures in discourse: Shared arguments in the Penn Discourse Treebank. In *Proceedings of the Constraints in Discourse III Workshop*, Potsdam, Germany, July 2008.

Hartmut E. H. Lenk. Sprachhandeln im Zeitungskommentar: Die Illokutionsanalyse (ISA) als Textbeschreibungsmodell. In Elisabeth Wåghäll Nivre, Brigitte Kaute, Bo Andersson, Barbro Landén, and Dessislava Stoeva-Holm, editors, *Begegnungen. Das VIII. Nordisch-Baltische Germanistentreffen in Sigtuna vom 11. bis zum 13. 6. 2009*, Stockholmer Germanistische Forschungen 74, pages 165–181, Sweden, Stockholm, 2011. Acta Universitatis Stockholmiensis.

Jiwei Li, Rumeng Li, and Eduard Hovy. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069, Doha, Qatar, October 2014a. Association for Computational Linguistics.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 25–35, Baltimore, Maryland, 2014b.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Corpora for the conceptualisation and zoning of scientific papers. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin R. Batchelor, and Dietrich Rebholz-Schuhmann. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000, 2012.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. What to do with an airport? mining arguments in the german online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153, Berlin, Germany, August 2016. Association for Computational Linguistics.

William Mann and Sandra Thompson. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281, 1988.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Daniel Marcu. A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 365–372, College Park/MD, 1999.

Daniel Marcu. *The theory and practice of discourse parsing and summarization*. MIT Press, Cambridge/MA, 2000.

James R. Martin. *English text: system and structure*. John Benjamins, Philadelphia/Amsterdam, 1992.

James R. Martin and Peter R. R. White. *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan, Houndsmills/New York, 2005.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Ryan McDonald, Koby Crammer, and Fernando Pereira. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 91–98, Ann Arbor, Michigan, June 2005a. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada, October 2005b. Association for Computational Linguistics.

Stephen Merity, Tara Murphy, and James R. Curran. Accurate argumentative zoning with maximum entropy models. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 19–26, Suntec City, Singapore, August 2009. Association for Computational Linguistics.

Eleni Miltsakaki, Aravind Joshi, Rashmi Prasad, and Bonnie Webber. Annotating discourse connectives and their arguments. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 9–16, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

Yoko Mizuta and Nigel Collier. An annotation scheme for a rhetorical analysis of biology articles. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1737–1740, Lisbon, Portugal, may 2004. European Language Resources Association (ELRA). ISBN 2-9517408-1-6.

Raquel Mochales and Marie-Francine Moens. Study on the structure of argumentation in case law. In *Proceedings of the 2008 Conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*, pages 11–20, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press. ISBN 978-1-58603-952-3.

Raquel Mochales and Marie-Francine Moens. Argumentation mining. *Artificial Intelligence and Law*, 19:1–22, 2011. ISSN 0924-8463.

Raquel Mochales Palau. *Automatic Detection and Classification of Argumentation in a Legal Case (Automatische detectie en classificatie van de argumentatie in een juridische zaak)*. PhD thesis, Faculty of Engineering Science, July 2011. Moens, Marie-Francine and De Schreye, Daniel (supervisors).

Raquel Mochales Palau and Aagje Ieven. Creating an argumentation corpus: do theories apply to real arguments? In *Proceedings of the ICAIL 2009*, pages 21–30. Barcelona, Spain, 2009.

Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the ICAIL 2009*, pages 98–109. Barcelona, Spain, 2009.

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 4th and 5th Workshops of the Forum for Information Retrieval Evaluation*, FIRE '12 & '13, pages 2:1–2:6, New York, NY, USA, 2013. ACM. doi: 10.1145/2701336.2701635.

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*, pages 225–230, 2007.

Johanna Moore and Martha Pollack. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544, 1992.

Wolfgang Motsch. Zur Illokutionsstruktur von Feststellungtexten. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 40(1):45–67, 1987.

Wolfgang Motsch. Zur Sequenzierung von Illokutionen. In Wolfgang Motsch, editor, *Ebenen der Textstruktur: Sprachliche und kommunikative Prinzipien*, Reihe Germanistische Linguistik 164, pages 189–208. Niemeyer, Tübingen, 1996.

Wolfgang Motsch and Renate Pasch. Bedeutung und illokutive Funktion sprachlicher Äußerungen. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 37(4):471–489, 1984.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India, December 2012a. The COLING 2012 Organizing Committee.

Philippe Muller, Marianne Vergez-Couret, Laurent Prevot, Nicholas Asher, Farah Benamara, Myriam Bras, Anne Le Draoulec, and Laure Vieu. Manuel d'annotation en relations de discours du projet ANNODIS. *Carnets de Grammaire*, 21, 2012b.

Huy Nguyen and Diane J. Litman. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, Denver, CO, June 2015. Association for Computational Linguistics.

Huy Nguyen and Diane J. Litman. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137, Berlin, Germany, August 2016. Association for Computational Linguistics.

Huy Nguyen, Wenting Xiong, and Diane Litman. Iterative design and classroom evaluation of automated formative feedback for improving peer feedback localization. *International Journal of Artificial Intelligence in Education*, pages 1–41, 2017. doi: 10.1007/s40593-016-0136-6.

Günther Öhlschläger. *Linguistische Überlegungen zu einer Theorie der Argumentation*. Niemeyer, Tübingen, 1979.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Silvia Pareti. A database of attribution relations. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Joonsuk Park and Claire Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Joonsuk Park, Arzoo Katiyar, and Bishan Yang. Conditional random fields for identifying appropriate types of support for propositions in online user comments. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 39–44, Denver, CO, June 2015. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Andreas Peldszus. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 88–97, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

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*, pages 196–204, Sofia, Bulgaria, August 2013a. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. From argument diagrams to automatic argument mining: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31, 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 (EMNLP)*, pages 938–948, Lisbon, Portugal, September 2015a. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. Towards detecting counter-considerations in text. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 104–109, Denver, CO, June 2015b. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015 / Vol. 2*, pages 801–816, London, 2016a. College Publications.

Andreas Peldszus and Manfred Stede. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 103–112, Berlin, Germany, August 2016b. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. Inhaltszonen. In Stede [2016a], pages 133–144.

Andreas Peldszus, André Herzog, Florian Hofmann, and Manfred Stede. Zur Annotation von kausalen Verknüpfungen in Texten. In Angelika Storrer, Alexander Geyken, Alexander Siebert, and Kay-Michael Würzner, editors, *Proceedings of the 9th Conference on Natural Language Processing (KONVENS 2008). Ergänzungsband Textressourcen und lexikalisches Wissen*, pages 71–83, Berlin, 2008. Zentrum Sprache BBAW.

Andreas Peldszus, Saskia Warzecha, and Manfred Stede. Argumentationsstruktur. In Stede [2016a], pages 185–208.

Jérémy Perret, Stergos Afantenos, Nicholas Asher, and Mathieu Morey. Integer linear programming for discourse parsing. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 99–109, San Diego, California, 2016.

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*, pages 1384–1394, San Diego, California, 2016a.

Isaac Persing and Vincent Ng. Modeling stance in student essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2174–2184, Berlin, Germany, August 2016b. Association for Computational Linguistics.

Livia Polanyi. A formal model of the structure of discourse. *Journal of Pragmatics*, 12: 601–638, 1988.

Livia Polanyi and Remko Scha. A syntactic approach to discourse semantics. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics, Proceedings of COLING '84*, pages 413–419. ACL, 1984.

Livia Polanyi and Martin van den Berg. Discourse structure and sentiment. In Myra Spiliopoulou, Haixun Wang, Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu, editors, *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, Vancouver, BC, Canada, December 11, 2011*, pages 97–102, Vancouver, BC, Canada, 2011. IEEE Computer Society. doi: 10.1109/ICDMW.2011.67.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. Sentential structure and discourse parsing. In *Proceedings of the Workshop on Discourse Annotation*, pages 80–87, Barcelona, Spain, 2004a. Association for Computational Linguistics.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. A rule based approach to discourse parsing. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 108–117, Cambridge, Massachusetts, USA, April 30 - May 1 2004b. Association for Computational Linguistics.

John L. Pollock. *Cognitive Carpentry: A Blueprint for How to Build a Person*. MIT Press, Cambridge, MA, 1995.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. Annotating attribution in the penn discourse treebank. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 31–38, Sydney, Australia, July 2006. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008a. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. The penn discourse treebank 2.0 annotation manual. Technical Re-

port IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia PA, 2008b.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950, december 2014.

Matthew Purver. Topic segmentation. In G. Tur and R. de Mori, editors, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 291–317. Wiley, 2011. ISBN 978-0-470-68824-3.

Ashequl Qadir and Ellen Riloff. Classifying sentences as speech acts in message board posts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 748–758, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

James Ravenscroft, Anika Oellrich, Shyamasree Saha, and Maria Liakata. Multi-label annotation in scientific articles - the multi-label cancer risk assessment corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Jonathon Read and John Carroll. Annotating expressions of appraisal in english. *Language Resources and Evaluation*, 421–447(3), 2012a.

Jonathon Read and John Carroll. Weakly-supervised appraisal analysis. *Linguistic Issues in Language Technology*, 8(2), 2012b.

Chris Reed and Katarzyna Budzynska. How dialogues create arguments. In *Proceedings of the 7th Conference on Argumentation of the International Society for the Study of Argumentation ISSA*. Rozenberg Quarterly, 2010.

Chris Reed, Douglas Walton, and Fabrizio Macagno. Argument diagramming in logic, law and artificial intelligence. *The Knowledge Engineering Review*, 22(1):87–109, March 2007.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. Language resources for studying argument. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA).

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. In Nicoletta Calzolari (Conference Chair),

Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, pages 1168–1171, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

Niall Rooney, Hui Wang, and Fiona Browne. Applying kernel methods to argumentation mining. In *Proceedings of the 25th FLAIRS Conference*, 2012.

Patrick Saint-Dizier. Processing natural language arguments with the TextCoop platform. *Journal of Argumentation and Computation*, 3(1):49–82, 2012.

Patrick Saint-Dizier and Juyeon Kang. Argument compound mining in technical texts: linguistic structures, implementation and annotation schemas. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015 / Vol. 2*, pages 895–906, London, 2016. College Publications.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66, Denver, CO, June 2015. Association for Computational Linguistics.

Holger Schmitt. *Zur Illokutionsanalyse monologischer Texte*. Peter Lang, Frankfurt, 2000.

Thomas Schröder. *Die Handlungsstruktur von Texten*. Narr, Tübingen, 2003.

William A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325, 1955.

John R. Searle. *Speech acts: an essay in the philosophy of language*. Cambridge University Press, Cambridge, 1969.

John R. Searle. A classification of illocutionary acts. *Language in Society*, 5(1):1–23, 1976.

Uladzimir Sidarenka, Andreas Peldszus, and Manfred Stede. Discourse Segmentation of German Texts. *JLCL*, 30(1):71–98, 2015.

Stefanie Simon, Erhard Hinrichs, Sabrina Schulze, and Yannick Versley. Handbuch zur annotation expliziter und impliziter diskursrelationen im korpus der tübinger baumbank des deutschen (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Deutschland, März 2011.

Noah A. Smith. *Linguistic Structure Prediction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool, 2011.

A. Francisca Snoeck Henkemans. State-of-the-art: The structure of argumentation. *Argumentation*, 14:447–473, 2000. ISSN 0920-427X. 10.1023/A:1007800305762.

A. Francisca Snoeck Henkemans. Complex argumentation in a critical discussion. *Argumentation*, 17:405–419, 2003. ISSN 0920-427X.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 170–179. Association for Computational Linguistics, 2009.

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*, pages 69–78, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Jonathan Sonntag and Manfred Stede. GraPAT: a tool for graph annotations. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Christian Stab. *Argumentative Writing Support by means of Natural Language Processing*. PhD thesis, Darmstadt University of Technology, Germany, 2017.

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

Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland, August 2014b. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. Parsing Argumentation Structures in Persuasive Essays. *ArXiv e-prints*, 2016. URL `https://arxiv.org/abs/1604.07370v2`. Version 2.

Peter Staudacher. Zur Semantik indefiniter Nominalphrasen. In Brigitte Asbach-Schnitker and Johannes Roggenhofer, editors, *Neuere Forschungen zur Wortbildung und Historiographie der Linguistik*, number 284 in Tübinger Beiträge zur Linguistik, pages 239–258. Gunter Narr Verlag, Tübingen, Germany, 1987.

Manfred Stede. DiMLex: A Lexical Approach to Discourse Markers. In Vittorio Di Tomaso Alessandro Lenci, editor, *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria, Italy, 2002.

Manfred Stede. The Potsdam Commentary Corpus. In Bonnie Webber and Donna K. Byron, editors, *ACL 2004 Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Manfred Stede. *Korpusgestützte Textanalyse. Grundzüge der Ebenen-orientierten Textlinguistik*. Narr, Tübingen, 2007.

Manfred Stede. *Discourse Processing*. Morgan and Claypool, 2011.

Manfred Stede. *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0*. Universitätsverlag Potsdam, 2016a.

Manfred Stede. Rhetorische Struktur. In *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0* Stede [2016a], pages 145–184.

Manfred Stede and Arne Neumann. Potsdam commentary corpus 2.0: Annotation for discourse research. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Manfred Stede and Andreas Peldszus. The role of illocutionary status in the usage conditions of causal connectives and in coherence relations. *Journal of Pragmatics*, 44(2):214–229, 2012. ISSN 0378-2166. doi: http://dx.doi.org/10.1016/j.pragma.2012.01.004.

Manfred Stede and Antje Sauermann. Linearization of arguments in commentary text. In *Proceedings of the Workshop on Multidisciplinary Approaches to Discourse*. Oslo, 2008.

Manfred Stede, Stergos Afantenos, Andreas Peldszus, Nicholas Asher, and Jérémy Perret. Parallel discourse annotations on a corpus of short texts. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portoroz, Slovenia, 2016.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974. ISSN 00359246.

Rajen Subba and Barbara Di Eugenio. An effective discourse parser that uses rich linguistic information. In *NAACL '09: Proceedings of Human Language Technologies – The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 566–574. Association for Computational Linguistics, 2009.

John M. Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge, 1990.

Maite Taboada and William Mann. Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4):567–588, 2006.

Luis Tari, Saadat Anwar, Shanshan Liang, James Cai, and Chitta Baral. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics*, 26(18):547–553, 2010.

Simone Teufel. *Argumentative Zoning: Information Extraction from Scientific Text*. dissertation, University of Edinburgh, 1999.

Simone Teufel. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Studies in Computational Linguistics. CSLI Publications, 2010.

Simone Teufel and Min-Yen Kan. Robust argumentative zoning for sensemaking in scholarly documents. In Raffaella Bernadi, Sally Chambers, Björn Gottfried, Frédérique Segond, and Ilya Zaihrayeu, editors, *Advanced Language Technologies for Digital Libraries*, volume 6699 of *Lecture Notes in Computer Science*, pages 154–170. Springer Berlin Heidelberg, 2011.

Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, December 2002. ISSN 0891-2017.

Simone Teufel, Jean Carletta, and Marc Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 110–117, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore, August 2009. Association for Computational Linguistics.

Stephen N. Thomas. *Practical Reasoning in Natural Language*. Prentice-Hall, New York, 1974.

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

Frans H. van Eemeren and Rob Grootendorst. *Speech acts in argumentative discussions: A theoretical model for the analysis of discussions directed towards solving conflicts of opinion*. Dordrecht: Floris Publications, 1984.

Frans H. van Eemeren and Rob Grootendorst. *Argumentation, Communication, and Fallacies: A Pragma-dialectical Perspective*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1992.

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

Antoine Venant, Nicholas Asher, Philippe Muller, Pascal Denis, and Stergos Afantenos. Expressivity and comparison of models of discourse structure. In *Proceedings of the SIGDIAL 2013 Conference*, pages 2–11, Metz, France, August 2013. Association for Computational Linguistics.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. Modeling review argumentation for robust sentiment analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 553–564. Dublin City University and Association for Computational Linguistics, 2014.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

Douglas Walton. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1996.

Douglas Walton. Objections, rebuttals and refutations. In *Proceedings of International Conference of the Ontario Society for the Study of Argumentation (OSSA)*, pages 1–10, Windsor/Ontario, 2009.

Douglas Walton. How to refute an argument using artificial intelligence. *Studies in Logic, Grammar and Rhetoric*, 23(36):123–154, 2011.

Douglas Walton. Building a system for finding objections to an argument. *Argumentation*, pages 1–23, 2012. ISSN 0920-427X. doi: 10.1007/s10503-012-9261-z.

Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

Saskia Warzecha. Klassifizierung und Skopusbestimmung deutscher Negationsoperatoren. Bachelor thesis, Potsdam University, 2013.

Bonnie Webber and Aravind Joshi. Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse. In *COLING-ACL Workshop on Discourse Relations and Discourse Markers*, pages 86–92, 1998.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. Discourse relations: A structural and presuppositional account using lexicalised tag. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 41–48, College Park, Maryland, USA, June 1999. Association for Computational Linguistics. doi: 10.3115/1034678.1034695.

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587, 2003.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6): 80–83, 1945. ISSN 00994987.

Dieter Wunderlich. Pro und Kontra. *Zeitschrift für Literaturwissenschaft und Linguistik*, 10 (38/39):109–127, 1980.

Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor Bench-Capon. Semi-automated argumentative analysis of online product reviews. In *COMMA 2012: Proceedings of the 4th International Conference on Computational Models of Argument*, Vienna, 2012. IOS Publishing.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China, July 2015. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. Conll 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Ann Arbor, Michigan, August 2016. Association for Computational Linguistics.

Mark A. Young and Robin Cohen. Determining intended evidence relations in natural language arguments. *Computational Intelligence*, 7(2):110–118, 1991. doi: 10.1111/j.1467-8640.1991.tb00386.x.

Fan Zhang, Diane Litman, and Katherine Forbes-Riley. Inferring discourse relations from pdtb-style discourse labels for argumentative revision classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2615–2624, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

Le Zhang. *Maximum Entropy Modeling Toolkit for Python and C++*, December 2004.