Automatic Analysis of Arguments about Controversial Educational Topics in Web Documents

Automatische Argumentanalyse in Webdokumenten zu kontroversen Bildungsthemen
Master-Thesis von Roland Kluge
April 2014

TECHNISCHE
UNIVERSITÄT
DARMSTADT

UBIQUITOUS
KNOWLEDGE
PROCESSING

Automatic Analysis of Arguments about Controversial Educational Topics in Web Documents
Automatische Argumentanalyse in Webdokumenten zu kontroversen Bildungsthemen

Vorgelegte Master-Thesis von Roland Kluge

1. Gutachten: Prof. Dr. Iryna Gurevych
2. Gutachten: Dr. Judith Eckle-Kohler

Tag der Einreichung:

Abstract

Decision making in social communities, such as families, companies, or parties, builds on debates and discussions, where arguments on particular topics are exchanged. With this work, we contribute to the efforts in automatically processing arguments for decision making, which is embedded in the field of Argumentation Mining.

Since only few corpora for Argumentation Mining exist, we first built a corpus of argumentative German Web documents, containing 79 documents from 7 educational topics, which were annotated by 3 annotators according to the claim-premise argumentation model. The corpus comprises 70,000 tokens, annotated with 5,000 argument units, i.e., claims and premises.

We found that the annotators performed similarly with regard to surface statistics such as the distribution of argument unit types or lengths. Each annotator's annotations cover on average ca. 74 of the tokens, which indicates the argumentative nature of the dataset. The inter-annotator agreement evaluates to ca. 44 % for Fleiss' κ and to ca. 40 % for Krippendorff's unitized alpha. We found that agreement correlates slightly negatively with annotation time demand per document.

Finally, we present a number of experiments on the role of 360 discourse markers for discriminating claims from premises. Our results show that several intensifying discourse particles are distinctive for claims and premises. Furthermore, we confirmed expectations from the literature that the discourse relation *concession* introduce counter-arguments. The discourse relations *comparison/contrast* and *result* frequently indicate claims, while the discourse relations *alternative*, *reason*, and *sequence* tend to indicate premises.

Another experiment investigated the role of discourse markers as features for Machine Learning. Using a Naïve Bayes classifier, we found that discourse markers as sole features for discriminating claims and premises yielded an improvement of 13 percentage points over the majority class baseline.

Zusammenfassung

Diskussionen und Debatten sind fundamental für soziale Entscheidungsprozesse beispielsweise in der Familie, im Unternehmen oder in der Politik. Diese Arbeit trägt dazu bei, Methoden zu entwickeln, um Entscheidungsprozesse durch automatische Verarbeitung von Argumenten zu unterstützen, und ist im Forschungsfeld *Argumentation Mining* angesiedelt.

In diesem Bereich existieren nur relativ wenige annotierte Korpora, sodass unser erster Schritt eine Annotationsstudie war, in der 3 Annotatoren 79 deutschsprachige Webdokumente zu 7 Bildungsthemen gemäß des *premise-claim* Argumentationsmodells annotiert haben. Das Korpus umfasst ca. 70.000 Wörter und es wurden ca. 5.000 Argumenteinheiten (*claims* und *premises*) annotiert.

Aus einer Reihe von Statistiken ist zu erkennen, dass das Annotationsverhalten gleichmäßig ist, was oberflächliche Eigenschaften wie beispielsweise die Länge und Verteilung der Argumenteinheiten angeht. Im Schnitt überdecken die Annotationen eines Annotators ca. 74 % der Wörter im Korpus, was auf dessen argumentative Natur hindeutet. Das Inter-Annotator Agreement liegt

für Fleiss’ κ bei ca. 44 % und für Krippendorffs α_u bei ca. 40 %. Wir haben herausgefunden, dass das Agreement (leicht) negativ mit der Annotationsdauer pro Dokument korreliert.

Abschließend stellen wir eine Reihe von Experimenten vor, die untersuchen, wie sich bestimmte Diskursmarker in den beiden Arten von Argumenteinheiten *claim* und *premise* verhalten. Einige steigernde oder abschwächende Diskurspartikel (*intensifiers/downtoners*) sind jeweils charakteristisch für *claims* und *premises*. Zudem konnten wir bestätigen, dass Gegenargumente durch die Diskursrelation *concession* eingeleitet werden können. Die Diskursrelationen *comparison/contrast* und *result* zeigten sich als markant für *claims*, wohingegen *alternative*, *reason* und *sequence* charakteristisch für *premises* sind.

In einem weiteren Experiment untersuchten wir die Eignung von Diskursmarkern als Features für das Maschinelle Lernen. Es zeigte sich, dass Diskursmarker als einzige Features bereits ausreichen, um bei der binären Klassifikation zwischen *claims* und *premises* die Majority Class-Baseline um 13 Prozentpunkte zu schlagen, wenn ein Naïve Bayes-Klassifizierer eingesetzt wird.

Contents

List of Abbreviations

α	significance level
A_i	annotator/annotation set i
A-Po	abbreviation for post-claim attack
A-Pr	abbreviation for pre-claim attack
α_u	Krippendorff's unitized alpha agreement
A_e	expected agreement (in general)
agr	helper function for calculating the *overlap score*
AIF	Argument Interchange Format
AM	Argumentation Mining
MON	maximum overlap normalizer
A_o	observed agreement (in general)
$A_{o,s}$	observed sentence-based agreement metric
$A_{o,t}$	observed token-based agreement metric
`ARG`	generalization level that aggregates all AUs of an argument into one span annotation
`AU`	generalization level that ignores all labels
AU	argument unit
C-Re	abbreviation for restatement
C	abbreviation for claim
D_e	expected disagreement (in general)
DiMLex	Discourse Marker Lexicon, a lexicon of ca. 170 German discourse markers
DKPro	Darmstadt Knowledge Processing Repository
DM	discourse marker
D_o	observed disagreement (in general)
DR	discourse relation
EXCITEMENT	Exploring Customer Interactions through TExtual EntailMENT (project)
GW	the external annotator in the annotation study
IAA	inter-annotator agreement

IQR	inter-quartile range (the difference between 25 %- and 75 %-quartile)
j	Jaccard-based agreement metric
JEK	Judith Eckle-Kohler (as annotator)
JSON	JavaScript Object Notation (data format)
κ_s	sentence-based kappa agreement metric
κ_t	token-based kappa agreement metric
MC	majority-class classifier
MP	Multi-layer Perceptron (classifier)
n	number of annotators
NB	Naïve Bayes (classifier)
NLP	natural language processing
NO-R	generalization level that drops all relational attributes (pre-/post-claim, restatement)
ORIG	generalization level that represents the original argument units
p	significance value
PDTB-DM	a custom lexicon of 51 DMs from the Penn Discourse Treebank
PDTB	Penn Discourse Treebank
pp.	percentage points
PR	generalization level that ignores the polarity of a premise (support/attack), but not its direction (pre-/post-claim)
Q_1, Q_3	25 %-/75 %-quartile
RF	Random Forest (classifier)
ρ_p	Pearson's rank correlation coefficient
ρ_s	Spearman's rank correlation coefficient
RK	Roland Kluge (as annotator)
RST	Rhetorical Structure Theory
RTE	Recognizing Textual Entailment
S-Po	abbreviation for post-claim support
S-Pr	abbreviation for pre-claim support
SVM	Support Vector Machine (classifier)
τ	Kendall's rank correlation coefficient

UIMA	Unstructured Information Management Architecture (framework)
ω_l	Wilson-Wiebe kappa agreement metric
ω_o	Wilson-Wiebe overlap agreement metric
X, Y	generic annotation labels

1 Introduction

Search engines help us to cope with the overwhelming amount of (textual) data on the Web. While traditional search engines were focused on document retrieval based on a user-independent ranking, recent approaches strive to consider the user's background and intent (search history, location, type of desired information) and additional semantic information.

Guha et al. [2003] distinguish between navigational and research (on-line) search. While users apply navigational search to navigate to a particular document (e.g., a weather report), research search (sometimes also called exploratory search) comes into play when users research about and explore a new topic (e.g., entry options for Ph.D. graduates in chemistry).

The latter example shows that educational issues (in a wide sense) are an application domain for research search, especially because decisions in this field dramatically affect one's further way of life. Other educational decisions are, for instance, which school (type) one's children should visit; whether to go to university or to do a training after school; and whether to take a Master's after one's Bachelor degree.

Owing to the big impact of these decisions, their answers should best be elicited based on arguments. Current search engines still lack the ability to explicitly (i.e., semantically) search for arguments. This work addresses first steps towards argumentative, exploratory search of arguments for making personal decisions with a focus on the educational domain. Vovk [2013] implemented a proof-of-concept system for retrieving arguments from Web documents in the educational domain. Figure 1.1 shows a mock-up of its user interface.

While Vovk used a coarse-grained representation of arguments, this work explores and applies the fine-grained claim-premise argumentation model in an annotation study on the same dataset. We hope to gain more insight into the appropriate level of detail for representing arguments in our application domain. For this purpose, we collected extensive statistics of and performed several experiments on the annotated dataset.

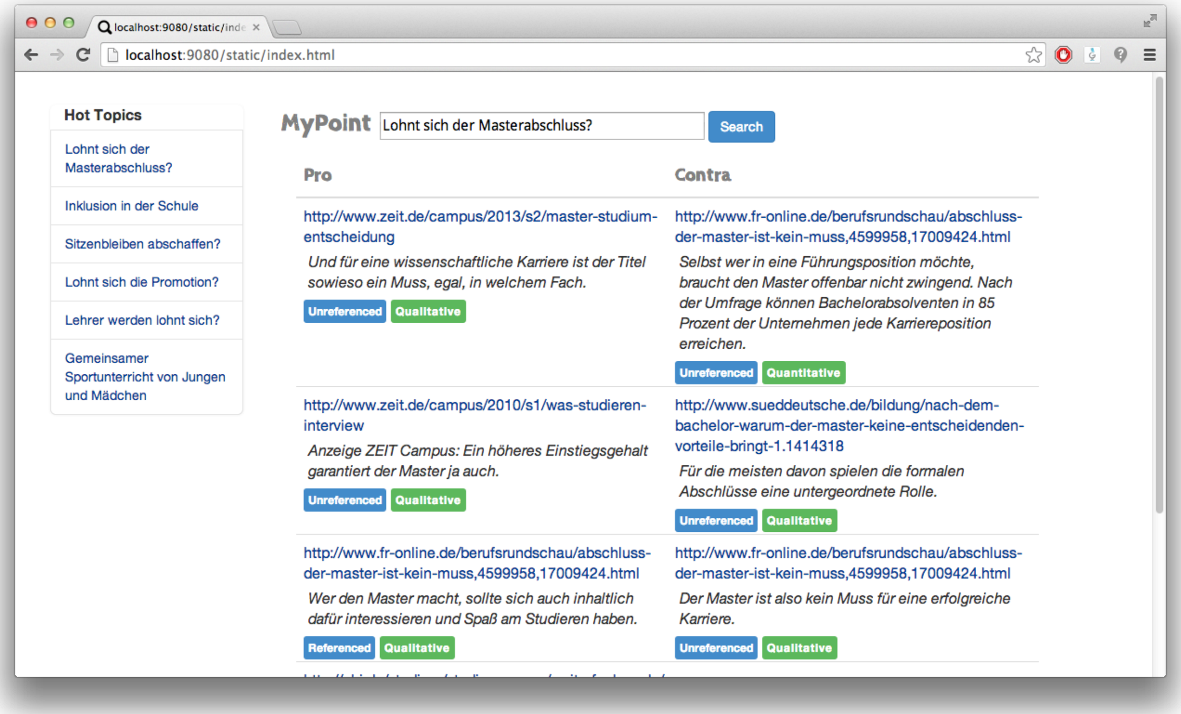
Related work shows that the importance of argumentative search is now starting to become recognized. Examples are legal cases (*What were arguments in precedent cases?* – [Sombekke et al., 2007; Moens et al., 2007] and policy making (*What are possible objections against a new policy?* – [Florou et al., 2013]).

Thesis Structure

The rest of this thesis is structured as follows:

Chapter 2 presents a survey of related work in the field of Argumentation Mining. We describe several application domains and introduce the argumentation models relevant for this work. Furthermore, we list existing resources and reason why we decided to build our own one. Finally, we describe related work in the field of discourse processing that relates to our experiments.

Chapter 3 describes the development of our annotated corpus. After summarizing the preprocessing of the corpus and the annotation study structure, we justify our annotation scheme and introduce the custom-made annotation tool. We conclude the chapter with a report on the post-processing and error corrections.



Hot Topics

- Lohnt sich der Masterabschluss?
- Inklusion in der Schule
- Sitzenbleiben abschaffen?
- Lohnt sich die Promotion?
- Lehrer werden lohnt sich?
- Gemeinsamer Sportunterricht von Jungen und Mädchen

MyPoint Lohnt sich der Masterabschluss? Search

Pro	Contra
http://www.zeit.de/campus/2013/s2/master-studium-entscheidung *Und für eine wissenschaftliche Karriere ist der Titel sowieso ein Muss, egal, in welchem Fach.* Unreferenced Qualitative	http://www.fr-online.de/berufsrundschau/abschluss-der-master-ist-kein-muss,4599958,17009424.html *Selbst wer in eine Führungsposition möchte, braucht den Master offenbar nicht zwingend. Nach der Umfrage können Bachelorabsolventen in 85 Prozent der Unternehmen jede Karriereposition erreichen.* Unreferenced Quantitative
http://www.zeit.de/campus/2010/s1/was-studieren-interview *Anzeige ZEIT Campus: Ein höheres Einstiegsgehalt garantiert der Master ja auch.* Unreferenced Qualitative	http://www.sueddeutsche.de/bildung/nach-dem-bachelor-warum-der-master-keine-entscheidenden-vorteile-bringt-1.1414318 *Für die meisten davon spielen die formalen Abschlüsse eine untergeordnete Rolle.* Unreferenced Qualitative
http://www.fr-online.de/berufsrundschau/abschluss-der-master-ist-kein-muss,4599958,17009424.html *Wer den Master macht, sollte sich auch inhaltlich dafür interessieren und Spaß am Studieren haben.* Referenced Qualitative	http://www.fr-online.de/berufsrundschau/abschluss-der-master-ist-kein-muss,4599958,17009424.html *Der Master ist also kein Muss für eine erfolgreiche Karriere.* Unreferenced Qualitative

Figure 1.1.: Mock-up of Vovk's prototype of an argumentative search engine. The user enters a query – in this case, whether obtaining a Master's degree pays off – and the system presents relevant arguments from the web in divided into pro and contra arguments and labeled with additional attributes (blue and green boxes).

Chapter 4 contains statistics of the annotated corpus. We characterize the corpus in terms of its topics and document categories and present further statistics on surface counts (sentence, tokens, etc.). To find out more about the annotation behavior, we investigated various properties of the annotations such as length, counts, and argumentation patterns.

Chapter 5 is dedicated to an analysis of inter-annotator agreement (IAA). We first delineate the applied IAA metrics and examine their characteristics with respect to other dimensions such as generalization level or confidence. We propose the maximum overlap normalization algorithm as a means for preparing a gold standard.

Chapter 6 describes three experiments on the role of discourse markers for distinguishing the argument units claim and premise. The discourse markers stem from three resources, including the Penn Discourse Treebank and the DiMLex lexicon. We explain the setup and results of each experiment and relate the results to previous findings in the literature.

Chapter 7 concludes this work with a summary of our contributions and an outlook on potential next steps.

Chapter A contains implementation notes, for instance concerning the corpus file format.

Chapter B through Chapter E contain tabulated information about the corpus and detailed data from the corpus analysis.

2 Literature Survey

This chapter presents a collection of related literature. We delineate the roots of Argumentation Mining (AM), describe existing resources for AM, and survey related work on discourse processing.

2.1 Argumentation Mining

Argumentation Mining (AM) deals with automatically identifying and analyzing arguments in coherent texts [Palau and Moens, 2009]. Walton [2009] states that the four main tasks of AM are the identification, analysis, evaluation, and invention (generation) of arguments. According to this scheme, our work is mainly concerned with the identification and analysis of arguments. The term *Argumentation Mining* (AM)[1] was probably invented in the field of legal argumentation in 2009 [Palau and Moens, 2009][2]. However, considerably earlier related work exists, especially with respect to scientific publications.

Argument Extraction is a subtask of AM and deals with the automatic identification of arguments from text. Once arguments have been extracted from a text, they need to be presented appropriately (*argument mapping*). Kirschner et al. [2003] survey a number of argument mapping tools, and Bentahar et al. [2010] propose to develop a joint representation for argumentation models in order to promote exchange between researchers. Trees and (in general) graphs seem to be appropriate structures to capture argumentative structures.

The interest in AM is increasing steadily as several recent workshops show, e.g., ARGAIP 2010[3] and 2013[4], or the *First Workshop on Argumentation Mining* at the upcoming ACL meeting in June 2014[5].

Other disciplines have influenced AM, e.g., Opinion Mining, Sentiment Analysis, argumentation theory, Artificial Intelligence, or discourse theory. Discourse theory is particularly important for AM because arguments form interlinked structures (argumentation lines) that could be described in terms of discourse relations.

2.1.1 Argumentation Mining for Newspapers and the Web

Today, semi-professional or lay writers produce argumentative content (including news) on-line, blurring the line between classical (printed) newspapers and the Web. Therefore, we address both domains in a common section.

Legal and scientific texts have a rather well-defined and formal structure, whereas language in newspapers and user-generated content (such as blogs) appears less structured. Especially enthymemes – arguments where premises are omitted – are typical for these genres as illustrated by Walton [2008].

[1] It seems that the lowercase version *argumentation mining* is more common, but for consistency with terms such as *Opinion Mining* or *Sentiment Analysis*, we use the uppercase form in this work.

[2] `http://scholar.google.de/scholar?q="argumentation+mining"&as_ylo=&as_yhi=2008`

[3] `http://wsarg2010.ing.unibs.it/`

[4] `https://sites.google.com/site/argaip2013/`

[5] `http://www.uncg.edu/cmp/ArgMining2014/`

User-generated content is challenging for classic NLP algorithms that often expect well-formed syntax (e.g., see [Gimpel et al., 2011]). Schneider et al. [2012] examined the particularities of such content with respect to AM and [Llewellyn, 2012] proposes a project for her Ph.D. that aims at extracting arguments from social media content.

Another line of research aims at structuring arguments in ongoing on-line debates to depict the current state of the debate. For the same task, Cabrio and Villata [2012] explored textual entailment techniques, while Heras et al. [2010] investigated which argumentation schemes from Walton et al. [2008] could be found in on-line discussions. Wyner and van Engers [2010] suggest to harvest user-generated argumentation with an on-line interface to support e-government. Florou et al. [2013] intend to build a Web crawler for arguments to improve public policy making in the first place.

Stede and Sauermann [2008] investigate argumentative structures in editorial comments using a graph-based argumentation model proposed by Freeman [1991].

Previous research on identifying arguments on political topics in newspaper texts is closely related to our work. Bal and Saint-Dizier [2009] aimed at identifying arguments on Nepalese political topics in newspaper articles collected from several sources with the objective to verify these arguments. A recent contribution presents ongoing efforts in creating a corpus of suitable texts [Bal and Dizier, 2010].

Our work follows recent work by Vovk [2013], who worked on German Web documents containing controversial educational topics. In Vovk's annotation scheme, an argument consists of one or more sentences and each argument can be for or against the topic of the document. Furthermore, each argument is classified according to whether it contains quantitative evidence and whether the origin (reference) of the argument is known.

The following sections briefly describe the current state of AM in other application domains. We suppose that some insights could also be transferred to the newspaper/Web domain, e.g., regarding argument visualization (teaching) or argumentation models (argumentative zoning in scientific publications).

2.1.2 Argumentation Mining for Scientific Publications

Works on argument identification in scientific documents started with the CARS model[6] that describes how authors motivate and justify their results in the introduction of publications [Swales, 1990].

Later, Teufel et al. [1999] extended the CARS model to cover whole scientific publications, which became known as *argumentative zoning* [Teufel, 1999]. Still today, there is an active line of research around argumentative zoning (e.g., [Teufel et al., 2009; Guo et al., 2013]).

In the same community, other groups aim to identify claims within scientific publications because claims are often used to relate one's own work to the scientific context [Park and Blake, 2012; Ahmed et al., 2013]. Tbahriti et al. [2006] explored how four argumentative moves (*purpose*, *methods*, *results*, and *conclusion*) may serve as features to find similar papers and found *purpose* and *conclusion* to be useful features.

[6] *Creating A Research Space*

2.1.3 Argumentation Mining for Legislation

Mining argumentation in legal texts is attractive because jurists spend considerable time in searching for arguments from precedent cases that may be applied in the current situation. As the jurisdiction differs from country to country, several national and international projects in this field have been started, such as ACILA[7] (2006 – 2010) or ARGUMENTUM [8] (2012 – present).

Researchers in the ARGUMENTUM project have begun to compile a corpus from decisions of the German Federal Constitutional Court [Houy et al., 2013]. Maarek [2010] delineates a project to support IT companies in drafting software contracts automatically by extracting argumentative knowledge from previous French IT contract cases. Sombekke et al. [2007] envisions an argument management system, where users manually arguments from legal dossiers, which may then be retrieved for future cases.

2.1.4 Argumentation Mining for Intelligent Writing Support

Argumentation Mining may improve intelligent writing support systems by providing students with feedback about their argumentative style. Moreale and Vargas-Vera [2003] suggest to integrate AM into the correction process of student essays. Their argumentation model builds on results from the scientific domain. Abbas and Sawamura [2008] propose an easy-to-query relational argument database to teach argumentation to students.

2.1.5 Argumentation Mining for Project Management

Another interesting application of AM is project management: Liu et al. [2009] expect that AM could help track design decisions and rationales in large-scale software projects and Browne et al. [2011] tried to detect design changes and inconsistencies in the documentation of aerospace projects.

2.1.6 Argumentation Mining and Textual Entailment

Recognizing Textual Entailment (RTE) emerged as an abstraction of several established Information Retrieval tasks such as Question Answering or Machine Translation. Textual entailment is defined as a directed relation between two texts, T (text) and H (hypothesis). We say that T entails H if a typical reader reading T will probably infer that H is true [Dagan et al., 2013, p. 3]. This definition shows that RTE is a highly relevant field for AM. For instance, we may expect that the set of premises entails the claim of an argument.

Several publications connect AM and RTE (e.g., [Hogenboom et al., 2010]). Cabrio and Villata [2012] used a TE system to extract argument networks from online debates. In a follow-up publication they conclude that the support relation (in argumentation) is related to the entailment relation (in TE) and that the attack relation corresponds to the contradiction relation [Cabrio and Villata, 2013].

[7] *Automatic Detection and Classification of Arguments in a Legal Case*, `http://www.cs.kuleuven.be/groups/liir/projects.php?project=125`

[8] *ARGUMENTUM - Analyse und Synthese von Argumentationsstrukturen durch rechnergestützte Methoden am Beispiel der Rechtswissenschaft* `http://argumentum.eear.eu/`

EXCITEMENT[9] is a collaborative effort to unite existing RTE systems into a common, open architecture based on UIMA[10], which will probably boost efforts in applying RTE for AM purposes.

2.2 Argumentation Models

Argumentation models offer formal representations to capture argumentation in natural text. This section presents the argumentation models that are relevant for our research. We detail on the model of claim and premise, on Toulmin's general argumentation scheme, and on Walton's collection of 96 argumentation schemes.

2.2.1 Claim and Premise

Argumentation models are by far older than AM, in fact the claim-premise model (also called premise-conclusion model) dates back to Aristotle's notion of deduction, known as *syllogism* [Smith, 2014]. A syllogism describes how supposed or given thoughts (*premises*, *protasis*) lead to a conclusion (*sumperasma*), also called the *claim*.

We define an argument according to the claim-premise argumentation model as follows: An *argument* consists of a number of so-called *argument units* (AUs), which can be either premises or claims. Premises either support or attack a claim. A claim that restates a previous claim within the same argument is called a *restatement*.

Note that this definition forecloses nested argumentative structures, where a premise may consists of more fine-grained argumentation units. In Section 3.3, we detail on the representation of claims and premises.

2.2.2 Toulmin's Model

In 1958, Stephen E. Toulmin proposed a general purpose argumentation scheme in his book *The Uses of Arguments* [Toulmin, 1958].

Toulmin was dissatisfied with the the claim-premise model and sought for other ways to represent roles within arguments. He came up with an argumentation model that defines six components: *claim*, *grounds*, *qualifier*, *warrant*, *backing*, and *rebuttal*.

The components interact as follows: An argument builds around the central *claim* that a proponent underpins by means of accepted facts (*grounds*). A *qualifier* modifies the strength of the claim, either downtoning (e.g., by using presumably) or amplifying (e.g., by using always) it. The conditions of *rebuttal* define exceptional cases in which the claim will not hold. The *warrant* justifies why the grounds support the claim (generalization relation). Since the *warrant* itself may be arguable, Toulmin introduced the *backing* that supports the warrant.

Figure 2.1 shows Toulmin's original example that illustrates his argumentation model. A formulation as continuous text could be: Harry was born in Bermuda, which is testified by his certificate of birth and other witnesses. Every man born in Bermuda will generally be a British subject as stated in the following statutes and legal provisions: [. . .]. So, presumably, Harry is a British subject, unless Harry has become a naturalized American citizen in the meantime.

[9] *EXploring Customer Interactions through Textual entailMENT*, `http://excitement-project.eu/`, [Wasserblat et al., 2012]

[10] *Unstructured Information Management Architecture*, `https://uima.apache.org/`

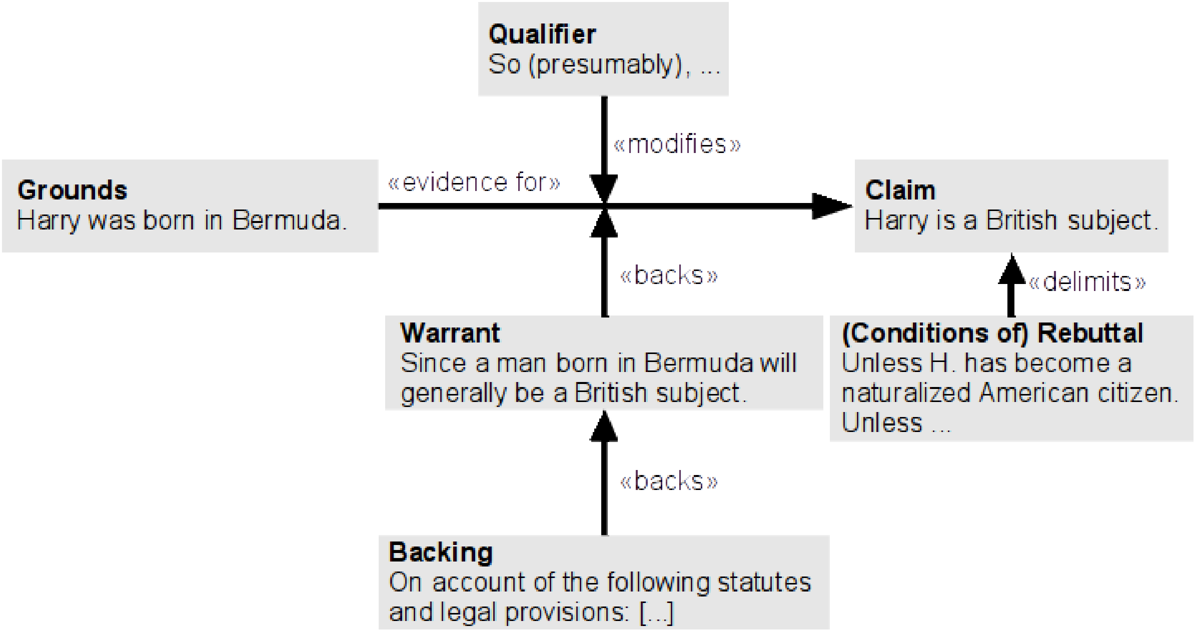

Qualifier
So (presumably), ...
«modifies»
Grounds
Harry was born in Bermuda.
«evidence for»
Claim
Harry is a British subject.
«backs»
«delimits»
Warrant
Since a man born in Bermuda will generally be a British subject.
(Conditions of) Rebuttal
Unless H. has become a naturalized American citizen.
Unless ...
«backs»
Backing
On account of the following statutes and legal provisions: [...]

Figure 2.1.: Example of Toulmin's argumentation model: *Harry is a British subject*. The example is taken from Toulmin [1958].

Ramage et al. [2008] bring about three major reasons, why Toulmin's scheme needs adaption in practical applications: First, the model was not intended to be a rhetorical one and so each argument is modeled in isolation. A second problem seems to be the argumentation „direction" from the grounds to the claim, which makes it difficult to integrate the frequent *result* relation. A third issue is the missing support for recursive argumentation: Toulmin did not allow for arguable grounds. In contrast, the backing serves as recursive support for the warrant, but for some applications this may still not suffice.

In a survey publication, Newman and Marshall [1992] describe further experience with Toulmin's model and several adaptations.

2.2.3 Walton's Argumentation Schemes

Walton et al. [2008] present a collection of 96 argumentative patterns, also called *argumentation schemes*. An argumentation scheme describes a pattern of interlinked propositions (several premises and a claim/conclusion) that form an argument. Critical questions are essential to argumentation schemes: Given an existing argument, they test whether the scheme has been applied correctly by questioning the truth of the premises.

Argumentation schemes experienced vivid attention in the AM community (e.g., [Walton, 2011, 2012; Feng and Hirst, 2011; Feng, 2010]). One reason may be that the publicly available argumentation corpus AraucariaDB (see Section 2.3) is labeled with argumentation schemes according to [Walton et al., 2008].

Argumentation schemes can be considered the contrary of Toulmin's general purpose argumentation model. An expectable downside is that annotating argumentation schemes needs extensive training (to learn each argumentation scheme) and that each scheme only covers a relatively small proportion of arguments [Walton, 2012].

2.3 Resources for Argumentation Mining

The lack of large corpora annotated with arguments is still a limiting factor on the progress in AM research [Mochales and Moens, 2011]. In this section, we list potential sources of annotated data for AM.

2.3.1 Argumentation Platforms – Arguments from the Wild

In the recent years, a number of debating websites have come up. In contrast to traditional forums, debating websites allow users to structure their arguments, e.g., into pro and contra or into argument clusters.

Debate.org[11] was founded in 2007 and claims to be the first of its kind[12]. Community members can choose among a traditional forum, opinion polls (yes/no with arguments), and a one-on-one debate mode where a proponent and a challenger exchange arguments over several rounds and the community finally crowns the winner.

All Our Ideas[13] started as a research project at Princeton University in 2010. Its concept is a mixture of interview and survey: Users are presented with two suggestions for solving a particular problem and decide which one is better. Suggestions are user-generated and are restricted to 140 characters.

While more debate platforms can be found on the Web[14], we failed to find a resource of user-generated content that provides adequate annotated data.

Groza and Indrie [2013] describe a promising project that intends to build a collaborative argumentation platform comparable to Wikipedia for so-called *mass argumentation*.

2.3.2 AraucariaDB (2005) and ArgDB-pl (2010)

AraucariaDB is one of the most popular corpora for AM [Reed and Rowe, 2004; Reed, 2006]. It was built with and is accessible through the argument mapping tool Araucaria [15], which currently supports three different argumentation schemes: Walton [1996], Pollock [1995], and Katzav and Reed [2004]. The corpus contains texts from several domains, including newspaper editorials, advertising, parliamentary records.

ArgDB-pl[16] is a related project in Polish. The corpus has been created with Araucaria-PL, a fork of Araucaria [Budzynska, 2011].

2.3.3 AIFdb (2013)

AIFdb is an extension of AraucariaDB, created in 2013, that additionally includes argumentation in blog posts [Lawrence et al., 2012]. Its database is compliant with the Argument Interchange

[11] http://www.debate.org/

[12] However, several similar projects existed before, such as Room 5, a social argument mapping web application [Loui et al., 1997]

[13] http://www.allourideas.org

[14] DebateGraph, RationalWiki, EssayForum are two interesting examples: http://debategraph.org, http://rationalwiki.org/wiki/Category:Debates, http://www.essayforum.com/

[15] http://araucaria.computing.dundee.ac.uk/doku.php

[16] http://argumentacja.pdg.pl/argdbpl/

Format (AIF) and its declared aim is to bridge existing argumentation tools such as Araucaria (see above), Carneades [Gordon and Walton, 2006], or Rationale [Van Gelder, 2007]: The tool-specific formats can be read, visualized, and written. Besides a user interface, the AIFdb project also defines web services.

2.3.4 Vovk's corpus (2013)

In 2013, Vovk created an annotated corpus consisting of ca. 90 German Web documents on educational topics [Vovk, 2013]. After a 2-step annotation study with 3 annotators, he identified ca. 500 arguments.

2.4 The Role of Discourse Processing

This section surveys related work from the field of discourse processing. We describe particularities of argumentative discourse and the role of discourse markers in argumentation.

2.4.1 Argumentative Discourse

Shell phrases are a discourse element that is particular for argumentative discourse. They serve as organizational parts that do not contribute to the argumentative content and, therefore, play a role similar to discourse markers in general discourse [Madnani et al., 2012]. Discourse markers such as *in conclusion*, *to sum up*, but also whole sentences such as the following can serve as shell phrases:

> **Example 2.1: On-sentence shell phrase**
> *Let's shed some light on another argument against staying down.*
> *Now, we come back to the original claim.*
> *There is even more evidence for this.*

The role of argumentative discourse in newspapers has received particular attention in the discourse processing community: Smirnova [2009] points out that citations serve as integral part of argumentation in newspapers and Iedema et al. [2003] emphasize that nucleus-satellite structures (in the sense of Rhetorical Structure Theory [Mann and Thompson, 1988]) are a good model for particular media sub-genres (e.g., „hard news“); see [Iedema et al., 1994] for detailed information.

2.4.2 Discourse Markers

In this section, we lay focus on so-called discourse markers (DMs). DMs are tokens, n-grams, and discontinuous word tuples that indicate discourse relations (DRs) between text segments as depicted in the following example:

Example 2.2: Examples of DMs:
In the following sentences, the token DM *however* indicates the *contrast* relation, the n-gram DM *as a result* signals the *result* relation, and the discontinued DM *either+or* expresses the *alternative* relation:

(1) *However*, her bike basket was too small for all the clothes she would have liked to buy.
(2) *As a result*, she had to decide:
(3) *Either* she could buy two pairs of jeans *or* she could afford to purchase the winter boots.

In the following, we investigate the role of DMs in corpus analysis and as features for classification tasks.

DMs in corpus analysis

The Penn Discourse Treebank (PDTB) [Prasad et al., 2008] is probably the most prominent discourse-annotated corpus. It builds on the 1 million word Wall Street Journal corpus and is annotated with DRs – such as *concession*, *contrast*, or *result* – and their relation arguments. A subset of the PDTB annotations consists of DRs that are lexically expressed or signaled by DMs, so-called *explicit* discourse connectives. For instance, the *concession* relation can be expressed by the DMs *however* and *but*. Some explicit DMs – such as *while* that appears in 12 DRs – seem to be highly polysemous.

A number of discourse tree banks in other languages have emerged in recent years, e.g., for Czech [Poláková et al., 2013] or Arabic [Al-Saif and Markert, 2010].

Torabi Asr and Demberg [2013] analyzed the DMs and their corresponding DRs annotated in the PDTB and addressed the question which information is conveyed by discourse connectives in the context of human sentence processing, i.e., how they contribute in the process of inferring a particular DR.

Taboada [2006] performed a corpus-based analysis of DRs annotated in the Rhetorical Structure Theory (RST) Discourse Treebank [Carlson et al., 2003]. The most frequent relation in the RST Discourse Treebank is *concession*, and this relation also received particular attention in the corpus linguistics literature: Taboada and Gómez-González [2012] present a corpus-based comparative study of DMs that express *concession* across English and Spanish in different genres. A classification of DMs signaling *concession* across English and German is presented by Grote et al. [1997]. They also point out the importance of *concession* in argumentative discourse: DMs expressing *concession* are often used to introduce counter-arguments in an argumentation line.

Cabrio et al. [2013] present an annotation study that aims to connect discourse processing with argumentation schemes (in the sense of Walton et al. [2008]). They selected four argumentation schemes – *Example, Cause to Effect and Effect to Cause, Practical Reasoning, Inconsistency* – and identified DRs in the PDTB that are typical for these schemes.

DMs as features for classification

Regarding the use of DMs as features for classification tasks, there is previous work in sentiment classification. Taboada et al. [2011] successfully employed discourse particles as features for the calculation of polarity scores in automated sentiment analysis. They focused on particles that act as intensifiers, i.e., which modify the semantic intensity of the lexical item they refer to (e.g., amplifiers such as *very* increase the semantic intensity, while downtoners such as *slightly* decrease it). Mukherjee and Bhattacharyya [2012] demonstrate that using discourse connectives as features in a system for sentiment classification of Twitter posts significantly improves classification accuracy over state-of-the art systems that do not consider DMs as features.

3 Annotation Study

This chapter describes the annotation study. We start with an overview of the two phases of the study – the pre-study and the main study – and describe the pre-processing steps that we performed. Afterwards, we explain the argumentation scheme and its implementation. We conclude the chapter with a description of the per-document annotation process, a presentation of the annotation tool, and a few words on the post-processing of the corpus.

3.1 Course of the Annotation Study

The annotation study divides into two parts: a pre-study and a main study. Table 3.1 summarizes the key facts about both phases. While the pre-study served to build a common understanding of the annotation task and to develop the annotation tool and guidelines, we produced the annotated corpus during the main study.

In the pre-study, we worked on a held-out development set consisting of documents from the topic *inklusion*. Two annotators participated in the pre-study: Judith Eckle-Kohler (JEK) and Roland Kluge (RK). For the main study, we additionally hired an inexperienced third annotator (abbreviated as GW)[1]. Adding a third annotator seemed sensible as this enables majority voting for creating a gold standard. Furthermore, an external annotator, who is not affected with the research question, may increase our confidence in the reproducibility of the annotations.

3.2 Document Acquisition and Pre-processing

Re-using the documents in Vovk's corpus appealed to us because they are in German and deal with controversial educational topics. Furthermore, we could assume that the documents were of high quality because they have been manually selected from the top 100 search engine hits[2]

[1] Two of the annotators are also author of and contributor to this work, which makes speaking about „the annotators" a little strange, but it allows us to distinguish between our two distinct roles – researcher and annotator – in the annotation study.

[2] Vovk used a controversial formulation of the topic as query, e.g., for *sitzenbleiben*, he issued the query „Sitzenbleiben abschaffen?"

	Pre-study	Main Study
Annotation time	ca. 3 weeks	ca. 6 weeks
Annotators	JEK, RK	JEK, RK, GW
#Topics	1	6
Topics	*inklusion*	*master, lehrer, promovieren* *g8, sport, sitzenbleiben*
#Documents	8	80

Table 3.1.: Information about pre-study and main study. The topics are given in the order of annotation.

and 1,000 crawled documents per topic. For more details on the document collection process, see [Vovk, 2013, Sec. 6.2].

However, the documents in Vovk's corpus lacked structural information such as paragraphs and headings. Our preliminary analysis of the data showed that it is helpful to see the original layout of a document during annotation, even for short documents.

For that reason, we redownloaded and pre-processed all documents again, which is described in the following sections.

3.2.1 Manual Pre-processing

Vovk's research target was to build a search engine for arguments. Even though the document selection was eventually done manually, the data acquisition ran automatically, which caused 8 documents to be incomplete because only the first of several pages was crawled[3].

We also omitted two documents from the corpus: Document *lehrer5* is a duplicate of *lehrer3*, and document *sitzenbleiben15* turned out to be a list of current practices (concerning when a student has to repeat a class) in various German states.

We recognized that paragraphs and headings were not always enclosed in the appropriate HTML tags (e.g., `<p>`, `<h1>`, `<h2>`). Instead, paragraphs occasionally appeared as two line breaks (`<br/><br/>`) and headings consisted of a paragraph in bold font (e.g., `<p><b>Heading</b></p>`). Due to the subsequent automatic pre-processing step, we had to inspect and correct every document carefully to normalize the representation of paragraphs and headings to standard HTML tags: Paragraphs were encoded with `<p>` tags, the title of each article was normalized to the `<h2>` tag, and subheadings were normalized to `<h3>`.

We reduced each document to the bare content of the article, excluding publishing date, author, images, boilerplate code, scripts, and advertisement. We manually converted list items (`<li/>`) to paragraphs in order to preserve the formatting as well as possible. Finally, we also normalize quotation marks and hyphens because the automatic HTML parser had problems with several special characters.

3.2.2 Automatic Pre-processing

While Vovk implemented the pre-processing in the Python framework *NLTK*[4], the complexity of preserving the paragraphs and headings throughout the automatic pre-processing made us decide for using *DKPro Core*[5], an extensive collection of NLP components, based on *Apache UIMA*[6].

The pre-processing consists of the following steps and is implemented in the class `PreprocessingApplication`:

1. **Input:** Reads the manually pre-processed HTML files
2. **Preserving paragraphs and headings:** Replaces `<p>` and `<h1>`…`<h6>` tags with special placeholders

[3] The extended documents are marked in Table B.1

[4] *Natural Language Toolkit*, `http://nltk.org/`

[5] *Darmstadt Knowledge Processing Repository*, [Gurevych et al., 2007], `https://code.google.com/p/dkpro-core-asl/`

[6] *Unstructured Information Management Architecture*, [Ferrucci and Lally, 2004], `https://uima.apache.org/`

3. **Removing other HTML code:** Removes any non-preserved HTML code using *jsoup* [7]
4. **Tokenization**: Splits text into tokens and sentences using the `LanguageToolSegmenter` component of *DKPro Core*.
5. **Token indexing:** Assigns a unique index to each token that is not a placeholder in order to make it addressable in the annotation tool
6. **HTML Rendering:** Creates HTML code ready to be visualized in the annotation tool
7. **Output:** Writes pre-processed corpus as JSON, HTML, and plain text

3.3 Annotation Scheme

This section describes the development of the annotation scheme. We tried to find a scheme that expresses the argumentative structures in the documents to a satisfying level of granularity and that is, at the same time, easy to apply.

During the pre-study we chose among three argumentation models: the claim-premise model, Toulmin's general argumentation model, and Walton's argumentation schemes.

3.3.1 Observed Document Structure

We found that most documents in our corpus comprise three major parts: an introductory part, a main part, and a concluding part, which play the following roles: The introduction summarizes the document content in one or two paragraphs and evokes the reader's interest. The main part consists of a sequence of arguments. The optional conclusion summarizes provides an outlook.

The introduction and conclusion also contain arguments, which are concise and on a high level, due to the abstract nature of these parts. The key point was to pick an appropriate (internal) argument representation.

3.3.2 Trying out Argumentation Schemes

At first, we considered to apply a subset of the argumentation schemes collected by Walton et al. [2008].

During a preliminary literature research, we recognized that the coverage may be a major problem: Walton [2012] found that only around 37 % of the arguments in their study could be matched to the argumentation schemes under consideration. The time demand to select an appropriate subset of the 96 schemes and to train the annotators is another disadvantage.

For these reasons, we refrained from using argumentation schemes in our study.

3.3.3 Trying out Toulmin's Scheme

Another option was Toulmin's scheme, which is not as specific as the argumentation schemes [Toulmin, 1958]. Furthermore, the annotators would only need to learn one argumentative pattern instead of 96. Unfortunately, we encountered issues with Toulmin's scheme as well.

[7] `http://jsoup.org/`

First, annotating the 8 documents in the development set took so long that the expected annotation time for the remaining 79 documents was unacceptable. Second, in most cases, we were only able to fill few slots of the scheme; most often we found grounds and claim. This may be a either due to insufficient training or due to the fact that the scheme is inappropriate for the documents.

So, we also abstained from Toulmin's scheme and resorted to the claim-premise argumentation model.

3.3.4 Trying out Claim-premise model

After our experience with argumentation schemes and Toulmin's model, we evaluated the claim-premise model on the development set and found it to be considerably more promising: Not only was the estimated time demand considerably smaller, also the agreement among the annotators increased dramatically.

Therefore, we adopted the this model for our annotation study. The next section clarifies how we represent the relations from the premises and restatement(s) to the claim.

3.3.5 Representing Relations

Two types of relations exist: (1) Premises may either support or attack a claim (*support* and *attack* relations), and (2) one claim may restate another one (*restatement* relation). We considered two representation forms for relations: graph-based and segmentation-based.

Graph-based representation

In the graph-based representation, AUs are annotated as spans, and the links between AU spans encode all all further information. This enables each AU to take multiple roles, e.g., as a premise of a high-level argument and as a claim of a low-level argument. The „main claims" of a text are those AUs with an outdegree of zero.

This annotation scheme is relatively versatile and promotes a two-step annotation process, where AUs are identified in a first step and the relations between the AUs are marked in a second step. The downside of this representation is the increased effort of implementing the annotation tool. Furthermore, the inter-annotator agreement analysis has to deal with the graph representation.

Segmentation-based representation

The segmentation-based representation also assumes a segmentation of the text into AUs, but it assigns a fixed role (*claim*, *support*, or *attack*) to each AU.

Therefore, this representation cannot model nested argumentative structures and only allows for linear argumentation, which means that there are no AUs of another argument between a premise/restatement and its associated claim. While this representation of relations appears to be more restrictive, it is easier to implement and to annotate than the graph-based representation.

Selecting a relation representation

Our choice for a relation representation is based on our experience in the pre-study. Occasionally, we found nested argumentative structures in the documents, which would necessitate

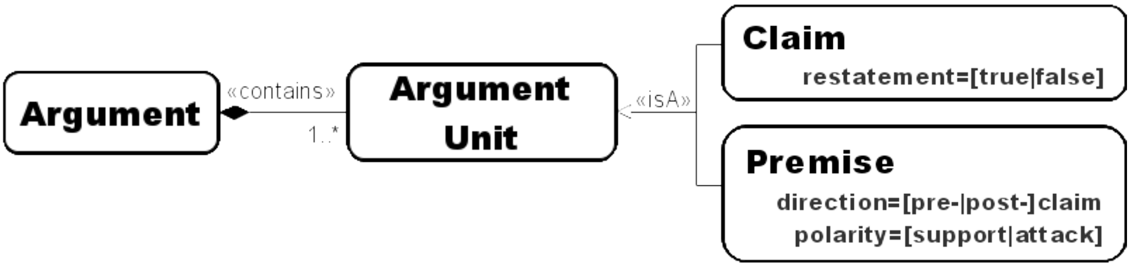


Figure 3.1.: Overview of the annotation scheme. A premise that appears before its associated claim is called *pre-claim premise*, a premise that appears after the claim is called *post-claim premise*, encoded in the *direction* attribute. The *polarity* of a premise signifies whether the premise supports or attacks the claim. Whether a claim is a restatement or not is modeled with the Boolean attribute *restatement*.

the graph-based representation. However, the agreement on the level of fine-grained arguments was so low and the discussions to sort out the disagreement were so time-consuming that we finally decided in favor of the segmentation-based representation. The decision to implement the annotation tool by ourselves (see Section 3.6) also backed this decision since a graph-based representation would have been more difficult to implement.

3.3.6 Final Argumentation Model

Our final argumentation model, summarized in Figure 3.1, is an adoption of the well-known claim-premise model. We decided for a segmentation-based representation of relations, assigning a fixed label to each AU, which is the only difference between our argumentation model and the general definition in Section 2.2.1.

3.3.7 Argument Unit Representation

Contrary to the hierarchical argumentation model, our implementation represents AUs as a single UIMA type (called `ArgumentUnit`) with a *label* feature that takes one of six values, summarized in Table 3.2. The labels are defined as constants in `ArgumentUnitLabel`.

3.3.8 Examples

We conclude this section with several illustrative examples (Examples 3.1, 3.2, and 3.3), which stem from the annotated corpus. At least 2 annotators agreed on each argument.

> **Example 3.1: First example of an argument (*g80*)**
> **Claim[DE]**: „Die Umstellung [zu G8] war schwierig“, sagt Diana.
> **Support[DE]**: In den Sommerferien nach dem Sitzenbleiben holte sie das nach, was ihr die G8er voraus hatten: Lateinvokabeln, Stochastik, Grammatik. „Den Vorteil, durch das Wiederholen den Stoff noch mal zu machen, hatte ich nicht.“
> –

Argumentation model	Label	Abbrev.
Claim[restatement=false]	claim	C
Claim[restatement=true]	claim-re	C-Re
Premise[polarity=support,direction=pre-claim]	support-pre	S-Pr
Premise[polarity=support,direction=post-claim]	support-post	S-Po
Premise[polarity=attack,direction=pre-claim]	attack-pre	A-Pr
Premise[polarity=attack,direction=post-claim]	attack-post	A-Po

Table 3.2.: Comparison of argumentation model and implementation. The UIMA type `ArgumentUnit` represents the different types of AUs as *label* feature.

Claim[EN]: „The change [to G8] was difficult,“ says Diana.
Support[EN]: [Since] After staying down, she had to catch up with the G8 students during her summer holiday, studying Latin vocabulary, stochastics, and grammar. „I did not have the advantage of reviewing previous material.“

Example 3.2: Second example of an argument (*lehrer0*)
Claim[DE]: Lehrer wird man, weil das ein sicherer Beruf ist.
Support[DE]: So denken noch immer viele junge Leute, die sich für eine Pädagogenlaufbahn entscheiden. Gut acht von zehn Erstsemestern, die 2009 mit einem Lehramtsstudium anfingen, war dieser Aspekt ihres künftigen Berufs wichtig oder sogar sehr wichtig. Keine andere Studentengruppe, die die Hochschul-Informations-System GmbH HIS befragte, legt so viel Wert auf Sicherheit

–

Claim[EN]: People become teachers because it is a safe job.
Support[EN]: This is what more and more young people who decide to become a teacher think. Well over eight of 10 freshman students who started to study to become teachers in 2009 considered this an important or very important aspect. No other group of students interviewed by the HIS set that much value on safeness.

Example 3.3: Third example of an argument (*promovieren0*)
Claim[DE]: Für die Unis sind Doktoranden günstige Arbeitskräfte.
Support[DE]: Eine Bekannte hatte mit ihrem Doktorvater zu kämpfen, der versuchte, sie noch am Institut zu halten, als ihre Arbeit längst fertig war. Er hatte immer neue Ausreden, weshalb er noch keine Note geben konnte. Als sie dann auch ohne Note einen guten Job bekam, außerhalb der Uni, spielte sich eine Art Rosenkrieg zwischen den beiden ab. Bis heute verlangt er von ihr noch Nacharbeiten an der Dissertation. Sie schuftet jetzt spätabends und am Wochenende für ihren Ex-Prof, der natürlich immer nur an ihrem Fortkommen interessiert war.

–

Claim[EN]: At university, graduate students are cheap employees.

> **Support[EN]**: An acquaintance struggled with her Ph.D. supervisor, who tried to keep her in his group at any rate, even though she had already completed her thesis. He pled more and more excuses for not yet grading her work. When she finally found a good job outside university – even without a final grade – a martial strife arose. Still today, he asks her to rework her dissertation. Now, she is drudging for her ex-supervisor, who always only wanted the best for her, late in the evening or on the weekend.

3.4 Generalization Levels

For the corpus analysis, it may be of interest to abstract from certain properties of the AUs such as the distinction between support or attack or the relations between AUs. This abstraction is captured by the concept of so-called generalization levels. Table 3.3 provides a survey of the generalization levels that we apply in this work.

Implementation Note

In the implementation, the generalization levels are represented by the Java enumeration `GeneralizationLevel`, which also contains structural levels (tokens, sentences) for technical reasons.

The `LabelGeneralizingAnnotator` component generalizes the labels of the original AUs (levels `PR`, `PR`, and `AU`), while the `ArgumentAnnotator` implements the generalization level `ARG`. The `OriginalArgumentLabelsRestorer` restores the original AUs (level `ORIG`) and deletes any derived AUs.

3.5 Per-document Annotation Process

This section describes the process of annotating a single document. It is an adapted excerpt of the annotation guidelines [Kluge, 2013]. Annotators were instructed to process each document in four iterations, also called rounds, as illustrated in Figure 3.2. This procedure helps to focus on one particular task and level of abstraction at a time.

Round 1: Gathering an overview

In the first round, the annotator gathers an overview of the document. He identifies the subject-matter by reading the introduction and the conclusion, and by skimming through the main part.

Round 2: Identifying arguments

With the topic of the document in mind, the annotator starts reading one paragraph after the other and identifies the argumentation line. To this end, the annotator *only* marks claims and restatement(s) in order to remain on a level of abstraction high enough to keep track of the whole argumentation line.

Round 3: Annotating premises

Next, the annotator annotates the premises. The annotator should validate his choice of claims by means of selecting appropriate premises. He should reconsider his claim annotations if no

Level	Abbrev.	Description	Labels
Original AUs	ORIG	represents the original AUs as marked by the annotators	*claim,* *claim-re,* *support-pre,* *support-post,* *attack-pre,* *attack-post*
Premises	PR	abstracts from support and attack premises and labels both as premise.	*claim,* *claim-re,* *premise-pre,* *premise-post*
No Relations	NO-R	abstracts from any relational attributes (polarity and direction for premises, and restatement relation for claims)	*claim,* *premise*
Argument Units	AU	abstracts from any label, leaving only the pure annotated AU spans	*arg_unit*
Arguments	ARG	aggregates the claim and its corresponding premises and restatements in a *new* annotation	*argument*

Table 3.3.: Overview of generalization levels. **Labels:** The AU labels used in the implementation.

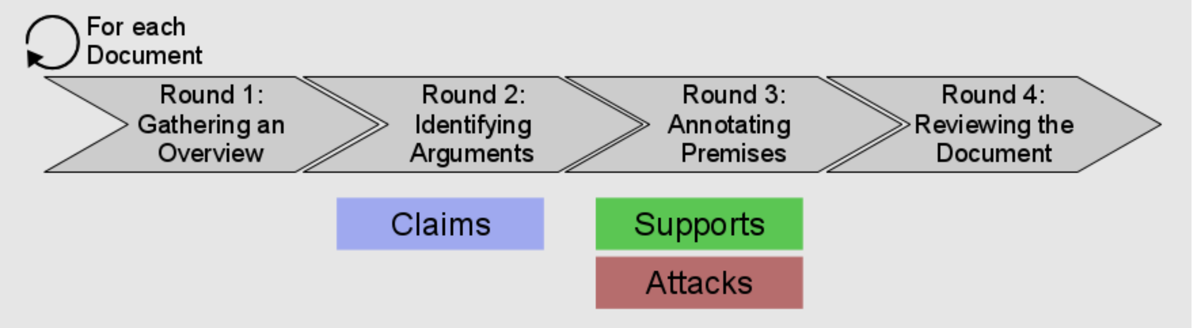

For each Document
Round 1: Gathering an Overview
Round 2: Identifying Arguments
Round 3: Annotating Premises
Round 4: Reviewing the Document
Claims
Supports
Attacks

Figure 3.2.: Overview of the per-document annotation process.

premises can be found for a claim. In the introduction and conclusion, claims without premises are acceptable, since the argumentation in these parts tends to be coarse-grained.

Round 4: Reviewing the document

During rounds 2 and 3, the annotator's view became more and more focused on the fine-grained argumentation structures. For this reason, we instructed the annotators to review the whole document in a forth iteration. To validate the annotations, the annotator answers (by himself) test questions such as the following:

- Are there largely unannotated passages? If yes, why?
- Are there claims without a premise? Are these statements arguable, actually? Could they be a restatement of another claim?
- Are the annotations structurally correct? For instance, a restatement is never directly preceded by a pre-claim premise since this premise is actually a post-claim premise for the corresponding claim.

3.6 Annotation Tool

This section describes the annotation tool. We summarize our requirements and explain why we think that enhancing Vovk's tool was the best choice. A survey of the implemented features concludes this section.

3.6.1 Requirements

The following list summarizes our requirements, which evolved during the pre-study.

- **Annotation spans:** The tool should allow to annotate arbitrary spans, which may be as short as a clause or as long as multiple sentences.
- **Annotation item properties:** Since the relations between AUs are encoded as property of the AUs, the tool should allow to assign attributes to annotations.
- **User-friendliness:** Using the tool should be easy to learn and annotators should be able to frequently and quickly save their annotations. The four-iteration annotation process should be supported, but not necessarily enforced.

- **Comparison perspective:** The tool should offer a perspective that compares annotations of two or three annotators side by side. This features is particularly important for a qualitative analysis of the annotated corpus.

- **Structured document visualization:** The tool should be able to display paragraphs and section headings appropriately. We found that annotating becomes more convenient when the annotators perceive the original layout of a document.

- **Multi-user support:** The tool should run on a central server and provide a web-based access for annotation and administration, to avoid installing the tool on each annotator's machine.

3.6.2 Existing annotation tools

This section provides an overview of existing annotation tools that we considered for our annotation study.

MAE annotation tool

The *Multi-purpose Annotation Environment* (MAE)[8] is an open-source, stand-alone Java application with a MySQLite backend. It is licensed under GNU GPL v3 and the most recent version is 0.9.6 from May 2012.

Annotation boundaries may be set at any character position and annotations are interlinked via labeled relations. Technical restrictions constrain the number of different annotation types to eleven.

Another open-source Java-based tool called *Multi-document Adjudication Interface* (MAI)[9] can be used to merge annotations by multiple annotators.

MMAX2

MMAX2[10] is an open-source Java-based, stand-alone annotation tool under Apache License 2.0. While the tool is used in several studies, we felt that maintenance has ceased, especially because the official project website[11] is no longer reachable and the latest version has been released in 2010.

brat rapid annotation tool

brat[12] is a web-based, open-source annotation tool under MIT license that is written in Python and designed to allow for maximum flexibility and quick setup. Documents are read from text files and tokenized according to whitespaces or regular expressions.

It allows to label spans and connect annotated spans with labeled arcs. Any number of annotators can participate and annotations by two annotators can be compared side-by-side. The in-window search looks up terms in Google and Wikipedia. Furthermore, a comparison perspective exists that shows two sets of annotations side by side.

[8] [Stubbs, 2011], `https://code.google.com/p/mae-annotation/`

[9] `https://code.google.com/p/mai-adjudication/`

[10] [Müller and Strube, 2006], `http://mmax2.net`

[11] `http://mmax.eml-research.de`

[12] `http://brat.nlplab.org/`

The major disadvantage of *brat* is its handling of long annotations: The whole annotation is displayed in one line.

WebAnno

WebAnno[13] is a web-based, open-source annotation tool under Apache License 2.0 that is written in Java and that builds upon *brat*'s visualization components. Multiple concurrent annotation layers exist (e.g., single-/multi-token level, arc annotations).

In contrast to *brat*, *WebAnno* offers a user management interface with different roles (user, curator, administrator) and has various readers for different input formats. Its curation perspective allows to compare and merge annotations by a number annotators. Furthermore, the annotated documents can be directly written to an XMI serialization compatible with the UIMA framework. This is convenient because our further analysis builds on UIMA.

Unfortunately, *WebAnno* inherited the problems with long lines from *brat*.

Vovk's tool

The tool that Vovk used for his study is written in Python and builds on Google App Engine[14]. It was tailored to his annotation process and is quite user-friendly. Import and export is carried out via a custom JSON format. Advanced functionality such as curation, process management are missing and we would need to adapt the tool to our own annotation model.

Our decision

MAE seems to be a mature annotation tool, but it lacks direct multi-user support; *MMAX2* seems to be discontinued, and so *WebAnno* was actually the tool of choice for our task because it is open-source and web-based, and exports annotations to XMI. Unfortunately, we had to refrain from it because of the annoying problems with long lines.

For this reason, we decided to adapt Vovk's tool for our purposes; the necessary implementation effort is the major downside of our decision.

3.6.3 Annotation Tool Features

The following paragraphs list the features that we implemented in the annotation tool.

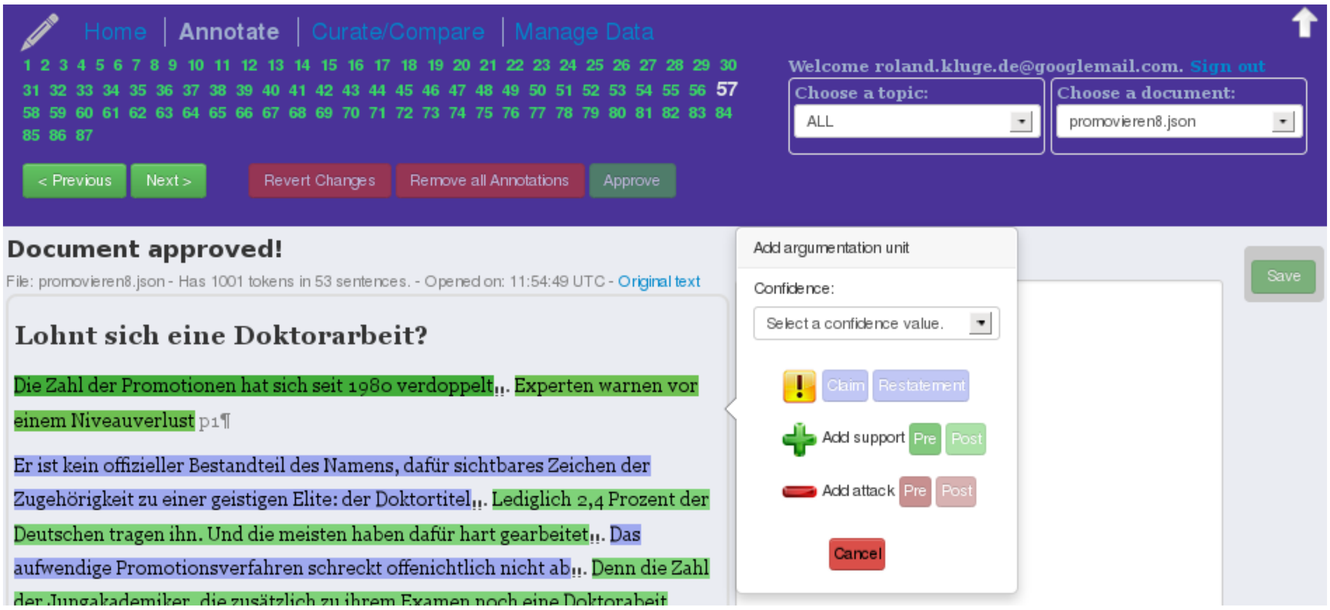
Annotation scheme

In the original tool, annotations consist of one or more sentences. In our implementation, we allow for arbitrary annotation boundaries. Figure 3.3 is a screenshot of the annotation perspective.

While Vovk enforced a two-step annotation process, we offer a single annotation perspective and encouraged the annotators to follow the proposed four-iteration process. Authors approve a document when they are done annotating it; afterwards, only a privileged user can unapprove it.

[13] [Yimam et al., 2013], `https://code.google.com/p/webanno/`

[14] `https://developers.google.com/appengine/`



Welcome roland.kluge.de@googlemail.com.

Choose a topic: ALL

Choose a document: promovieren8.json

Document approved!

File: promovieren8.json - Has 1001 tokens in 53 sentences. - Opened on: 11:54:49 UTC - Original text

Confidence: Select a confidence value.

Lohnt sich eine Doktorarbeit?

Die Zahl der Promotionen hat sich seit 1980 verdoppelt$_{!!}$. Experten warnen vor einem Niveauverlust p1¶

Er ist kein offizieller Bestandteil des Namens, dafür sichtbares Zeichen der Zugehörigkeit zu einer geistigen Elite: der Doktortitel$_{!!}$. Lediglich 2,4 Prozent der Deutschen tragen ihn. Und die meisten haben dafür hart gearbeitet$_{!!}$. Das aufwendige Promotionsverfahren schreckt offenichtlich nicht ab$_{!!}$. Denn die Zahl der Jungakademiker, die zusätzlich zu ihrem Examen noch eine Doktorarbeit

Figure 3.3.: Screenshot of annotation perspective. Annotations are created by selecting the first and last token. Afterwards, the annotator assigns a confidence score and the AU label. Subscript markers at the end of annotations indicate the confidence level (e.g., *!!* stands for high, and *!?* stands for low confidence).

Confidence level

Each annotation receives a so-called confidence score, which can be *high*, *medium*, or *low*. In most situations, annotators should be highly confident, that is, they are sure about the selected annotation span and label. If the argumentative structure is unclear or the annotator does not fully understand the author's point he should assign *medium* confidence. *Low* confidence score was reserved to mark border cases or candidates for further discussion.

Per-document notes

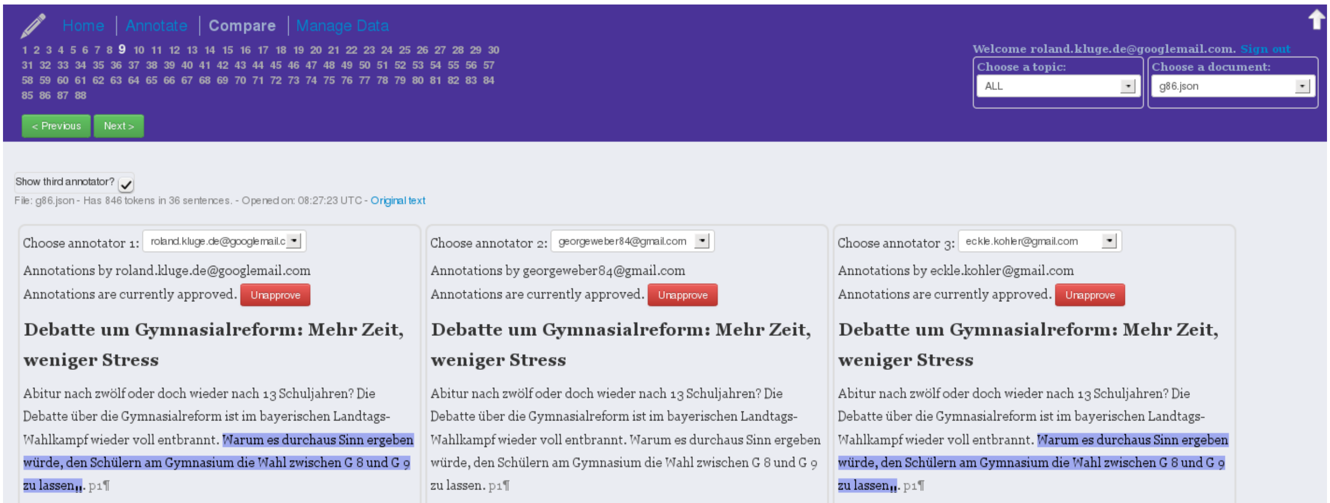
Each annotator can assign notes to a document to share his thoughts about difficult or remarkable cases with us. Particularly in case of medium or low confidence, we encouraged the annotators to note down their concerns.

Identity switching

We encountered several cases where annotations had to be corrected for technical reasons. In order to become independent of the annotators to respond to our correction requests, we allowed authenticated users to take the identity of any annotator in the annotation perspective.

Currently, switching identities is only possible through editing the URL. In the following example, document *g80* is opened for annotation from the viewpoint of *abc@gmail.com*:

> **Example 3.4: Temporary identity switching**
> `<site>/argunit/annotate?doc=g80.json&user=abc@gmail.com`



Welcome roland.kluge.de@googlemail.com.

Choose a topic: ALL

Choose a document: g86.json

Show third annotator? ☑

File: g86.json - Has 846 tokens in 36 sentences. - Opened on: 08:27:23 UTC - Original text

Choose annotator 1: roland.kluge.de@googlemail.c

Annotations by roland.kluge.de@googlemail.com

Annotations are currently approved.

Debatte um Gymnasialreform: Mehr Zeit, weniger Stress

Abitur nach zwölf oder doch wieder nach 13 Schuljahren? Die Debatte über die Gymnasialreform ist im bayerischen Landtags-Wahlkampf wieder voll entbrannt. Warum es durchaus Sinn ergeben würde, den Schülern am Gymnasium die Wahl zwischen G 8 und G 9 zu lassen. p1¶

Choose annotator 2: georgeweber84@gmail.com

Annotations by georgeweber84@gmail.com

Annotations are currently approved.

Debatte um Gymnasialreform: Mehr Zeit, weniger Stress

Abitur nach zwölf oder doch wieder nach 13 Schuljahren? Die Debatte über die Gymnasialreform ist im bayerischen Landtags-Wahlkampf wieder voll entbrannt. Warum es durchaus Sinn ergeben würde, den Schülern am Gymnasium die Wahl zwischen G 8 und G 9 zu lassen. p1¶

Choose annotator 3: eckle.kohler@gmail.com

Annotations by eckle.kohler@gmail.com

Annotations are currently approved.

Debatte um Gymnasialreform: Mehr Zeit, weniger Stress

Abitur nach zwölf oder doch wieder nach 13 Schuljahren? Die Debatte über die Gymnasialreform ist im bayerischen Landtags-Wahlkampf wieder voll entbrannt. Warum es durchaus Sinn ergeben würde, den Schülern am Gymnasium die Wahl zwischen G 8 und G 9 zu lassen. p1¶

Figure 3.4.: Screenshot of comparison perspective. The user can switch between two-annotator and three-annotator comparison.

Linking to paragraphs and tokens

Every paragraph and token can be linked to via anchors in the HTML site. While this appears to be a minor feature, it was helpful during the automatic analysis, as the analysis component could generate links to arbitrary positions in the documents. The following example demonstrates how to link to the first paragraph and the twelfth token in *g80*:

Example 3.5: HTML anchors to paragraphs and tokens

```
<site>/argunit/compare?doc=g80.json#p1
<site>/argunit/compare?doc=g80.json#t12
```

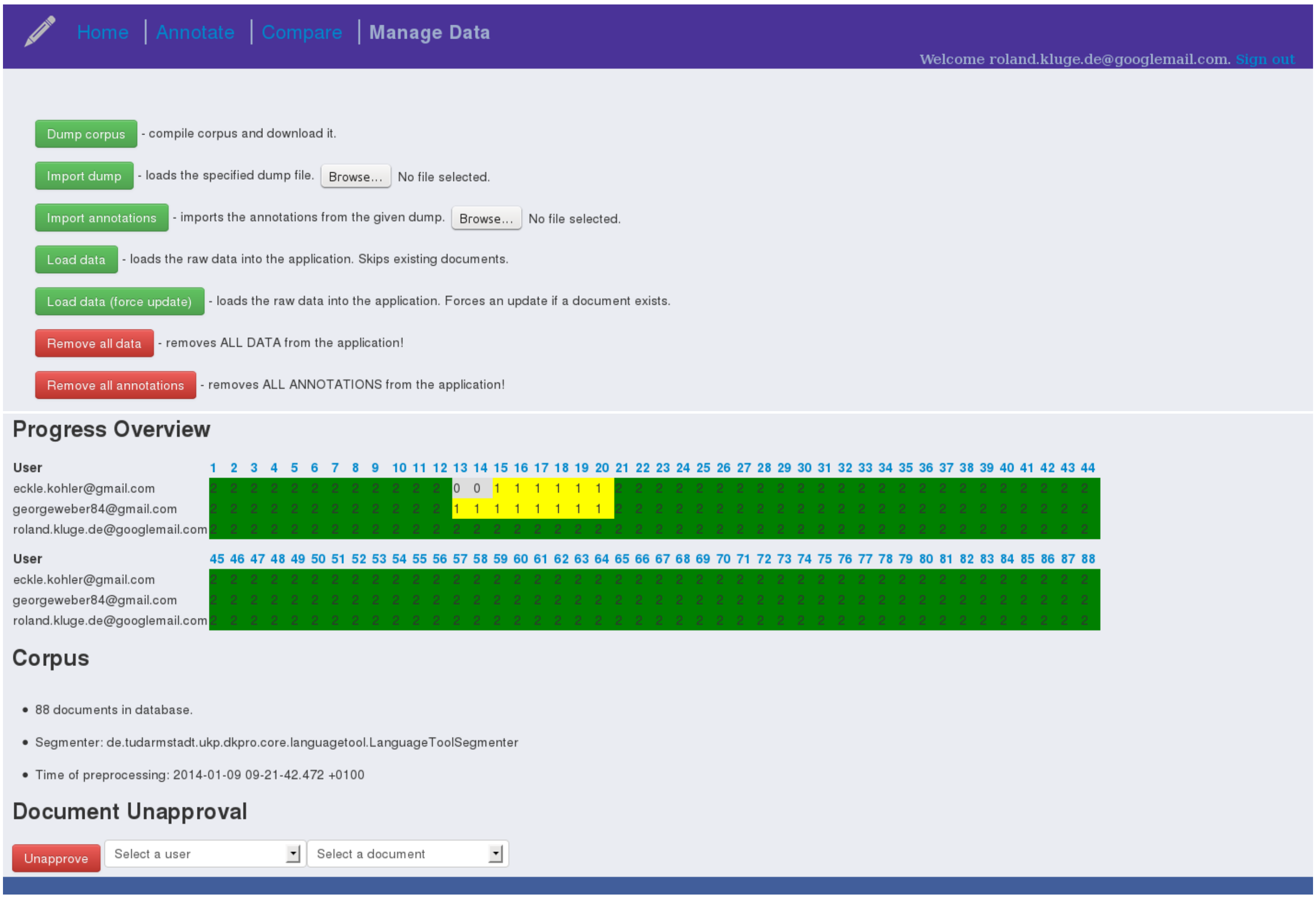
Comparison perspective

We also implemented a comparison perspective that shows annotations by up to three annotators side by side. Figure 3.4 shows a screenshot of the comparison perspective.

Administration perspective

The original tool could only import and export data via specific URLs. We implemented a convenient data administration perspective and improved the error handling (e.g., missing documents). All features are now accessible through a common user interface, depicted in Figure 3.5. The following list summarizes the new features.

- User-privileges determine the perspectives that a user has access to.
- Export of corpus (JSON), import of annotations or whole corpus from dump file
- Load and re-load pre-processed documents, e.g., in order to correct typos or conversion errors on-the-fly
- Overview of each annotator's progress
- Approval and unapproval documents



Dump corpus - compile corpus and download it.

Import dump - loads the specified dump file. Browse... No file selected.

Import annotations - imports the annotations from the given dump. Browse... No file selected.

Load data - loads the raw data into the application. Skips existing documents.

Load data (force update) - loads the raw data into the application. Forces an update if a document exists.

Remove all data - removes ALL DATA from the application!

Remove all annotations - removes ALL ANNOTATIONS from the application!

Progress Overview

User	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
eckle.kohler@gmail.com	2	2	2	2	2	2	2	2	2	2	2	2	0	0	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
georgeweber84@gmail.com	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
roland.kluge.de@googlemail.com	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

User	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88
eckle.kohler@gmail.com	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
georgeweber84@gmail.com	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
roland.kluge.de@googlemail.com	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Corpus

- 88 documents in database.
- Segmenter: de.tudarmstadt.ukp.dkpro.core.languagetool.LanguageToolSegmenter
- Time of preprocessing: 2014-01-09 09-21-42.472 +0100

Document Unapproval

Unapprove | Select a user | Select a document

Figure 3.5.: Screenshot of administration perspective. The first part provides functionality for data management (import, export, and remove data). The second part is an overview of the annotators' progress (unprocessed, in progress, and approved). At the bottom, the administrator can unapprove documents that need revision.

3.7 Post-processing

This section describes the post-processing steps that the corpus has undergone.

3.7.1 Annotating Paragraphs

Knowing the position of an argument unit within its paragraph is an interesting information, especially for future classification experiments. The `ParagraphAnnotator` component annotates the DKPro Core type `Paragraph`[15] based on the <p> tags in the HTML document.

3.7.2 Annotating Headings

Headings are often not terminated with a period and caused problems during sentence splitting. The `HeadingAnnotator` component annotates the DKPro Core type `Heading`[15] and corrects the sentence splitting so that headings appear in a separate sentence.

3.7.3 Manual Corrections

Structural annotation errors

The annotation tool was not able to automatically check for cases where annotators produced illegal AU patterns. For instance, a post-claim premise that directly follows a pre-claim premise is an annotation error because the necessary intermediate claim is missing.

We found ten structural errors and resolved them manually during the post-processing.

Overlapping annotations (per annotator)

The original annotation guidelines allowed annotations to overlap in order to model discontinuous annotation. Only one annotator used this feature in 11 cases.

In the first nine cases, the overlapping annotations were a claim and a support with identical span. In the tenth case, a whole sentence was annotated as claim and an embedded clause as support. In Example 3.6, the annotator interpreted the mentioning of the reputable broadcasting corporation ARD as evidence. The eleventh case was an unintended overlap, where an annotator overlooked a duplicate post-support that was concealed by another post-support.

> **Example 3.6: Example of overlapping annotation**
> **Claim[DE]**: „Die Bilanz nach zehn Jahren Turbo-Abi ist verheerend“, so fasste es die ARD in einem Beitrag über das G8 zusammen.
>
> **Support[DE]**: so fasste es die ARD in einem Beitrag über das G8 zusammen.
>
> –
>
> **Claim[EN]**: „Ten years of G8 leave us with a disaster,“ an ARD television report dealing with G8 concludes.
>
> **Support[EN]**: an ARD television report dealing with G8 concludes.

[15] see `https://code.google.com/p/dkpro-core-asl/wiki/TypeSystem`

The reason for the first 10 overlapping annotations was that an annotator tried to identify fine-grained annotation structures, which conflicts with our rule to annotate at least clauses. Consequently, he was not able to determine the exact AU boundaries and avoided annotating on a sub-clause level. We resolved all 10 cases by discussion to avoid overlapping annotations in the final corpus.

4 Corpus Statistics

This chapter collects a variety of statistics of the corpus. Some of them are rather informative (How many documents are there? From there do they stem? ...), others helped to decide which inter-annotator agreement metrics to apply (see Section 5) or how a gold standard could be created (see Section 7.2).

4.1 Topics

The corpus contains documents from seven topics. The documents of a topic are numbered starting from 0; for example, *sport0* is the first document in topic *sport*.

- ***g8***: Should students visit secondary school („Gymnasium“) for 8 years (G8) or 9 years (G9)?
- ***inklusion*** (also known as *mainstreaming*): Should students with special needs visit regular classes?
- ***lehrer***: Does it pay off to become a teacher? Should teachers be civil servants? Is being a teacher a hard job?
- ***master***: What are the advantages and disadvantages of achieving a Master's degree?
- ***promovieren***: What are the advantages and disadvantages of graduating?
- ***sitzenbleiben***: Should students with low grades repeat a class at school?
- ***sport*** (also known as co-education): Should schools hold mixed-sex sports lessons?

4.2 Document Categories

We observed that the texts in the corpus belong to different categories. We identified three obvious categories[1]:

1. **interview**: An interview is a printed representation of a dialog between somebody from the editorial staff and an interviewee, e.g., an expert or politician. It starts with an introductory part that describes the surrounding topic. Afterwards, a sequence of question-answer pairs follows with an optional concluding part.
2. **news**: Documents in this category are about current events such as recently published surveys, strikes, court decisions, or new policies.
3. **article/other**: This category groups all documents that are neither *interview* nor *news*. Most of the documents are articles, but there are also (a few) specific categories such as petitions, comments, essays, survey articles, or blog posts.

Table 4.1 shows the distribution of document categories.

[1] The selection of document categories is not backed by any formal theory.

Type	All	*inklusion*
interview	6	0
news	39	1
article/other	43	7
Sum	88	80

Table 4.1.: Document category distribution

4.3 Document Sources

Table 4.2 summarizes the 29 source domains of the 88 documents. The complete mapping from URL to document can be found in Appendix C.

Source	All	*inklusion*
spiegel.de	31	1
welt.de	10	2
sueddeutsche.de	8	1
focus.de	6	1
zeit.de	5	0
bildung-news.com	2	0
derwesten.de	2	0
faz.net	2	0
tagesspiegel.de	2	0
abi.de	1	0
badische-zeitung.de	1	0
berliner-zeitung.de	1	0
blog.initiativgruppe.de	1	0
bpb.de	1	1
change.org	1	0
christophburger.de	1	0
daserste.ndr.de	1	0
dw.de	1	0
fr-online.de	1	0
haus-der-sprache.de	1	0
heise.de	1	0
ingenieur.de	1	0
ismail-tipi.de	1	0
jobvector.de	1	0
ksta.de	1	0
myhandicap.de	1	1
neues-deutschland.de	1	0
nsfkn.de	1	1
rbb-online.de	1	0

Table 4.2.: Overview of source URLs and number of documents per source URL. *spiegel.de* contributes 35 % of the 88 documents.

Several newspaper agencies (e.g., *spiegel.de*, *welt.de*, *focus.de*) contribute the largest fraction of documents (ca. 83 %). All portals that focus on particular topics – *abi.de*, *ingenieur.de*, *bpb.de*, *jobvector.de*, *myhandicap.de* – contribute to only one topic each. Other sources are blogs (e.g., *christophburger.de*, *nsfkn.de*, *ismail-tipi.de*) and the petition website *change.org*.

It is notable that the documents *master0* (*sueddeutsche.de*) and *master5* (*fr-online.de*) build on the same background story; several passages are literally identical.

Topic	Documents	Paragraphs	Sentences	Tokens
g8	12	150	601	11,149
lehrer	19	248	876	16,074
master	7	103	407	7,783
promovieren	13	180	785	13,542
sitzenbleiben	21	297	976	18,188
sport	8	59	218	3,728
All (no *inklusion*)	80	1,037	3,863	70,464
inklusion	8	146	611	11,467
All	88	1,183	4,474	81,931

Table 4.3.: Per-topic statistics on token, sentence, paragraph, and document counts.

Annotator	Time [hh:mm]	Rate[mpd]	Rate[spm]	Rate[tpm]
RK	27:43	20.8	2.3	42
JEK	18:04	16.8	2.9	53
GW	22:20	13.6	3.6	64
Sum	67:43	–	–	–
Average	22:34	17.0	2.9	53

Table 4.4.: Annotation time per annotator. Annotation time was measured per document. With a number of 3,863 sentences and 70,464 tokens, we obtain annotation rates as minutes per document (**mpd**), sentences per minute (**spm**), and tokens per minute (**tpm**).

4.4 Token, Sentence, Paragraph, and Document Counts

Counting tokens, sentences, paragraphs, and documents per topic help to decide on how to split the documents into training and test set for a later classification task (Table 4.3).

4.5 Time Demand

We instructed the annotators to track their annotation time demand per document. The annotation rate (measured in tokens per minute or sentences per minute) may help to identify documents that were relatively hard to annotate.

All together, the annotation study took about 68 h and the average annotation time per annotator is about 22.5 h. Table 4.4 summarizes the total annotation time per annotator.

4.6 Argument Unit Count Statistics

This section analyzes the annotation label distribution. Table 4.5 gives an overview of the per-label distribution of the AUs; Table 4.6 shows the corresponding statistics for the generalization levels `NO-R` and `ARG`.

There is a notable imbalance between support and attack in the corpus (46.2 % vs. 6.6 %). Premises and claims appear almost similarly frequently (53 % vs. 47 %).

We see that all annotators made use of claims and restatements roughly to the same extent (45.3 % to 46.2 % , 1.3 % to 1.7 %). JEK particularly favored pre-claim attacks (3.6 % vs. 1.3 % and 1.4 %) and GW favored post-claim supports (39.0 % vs. 36.0 % and 35.0 %).

On average, the absolute difference between the per-annotator rows and the averages row is 0.7 pp, while the maximum difference is 2.5 pp which indicates a consistent annotation behavior with respect to the distribution of AU labels.

Annotator	C	C-Re	A-Pr	A-Po	S-Pr	S-Po	Sum
RK	781	28	22	76	175	609	1691
	(46.2 %)	(1.7 %)	(1.3 %)	(4.5 %)	(10.3 %)	(36.0 %)	(33.0 %)
JEK	759	23	60	67	164	577	1650
	(46.0 %)	(1.4 %)	(3.6 %)	(4.1 %)	(9.9 %)	(35.0 %)	(32.2 %)
GW	809	23	30	80	139	704	1785
	(45.3 %)	(1.3 %)	(1.7 %)	(4.5 %)	(7.8 %)	(39.4 %)	(34.8 %)
All	2349	74	112	223	478	1890	5126
	(45.8 %)	(1.4 %)	(2.2 %)	(4.4 %)	(9.3 %)	(36.9 %)	

Table 4.5.: Distribution of AU labels. This table breaks down the distribution of AUs per annotator (column **Sum**). Each of the first three rows displays the distribution of labels per annotator. In all but the last column the percentages are the fraction of the row total; in the last row and column, the percentages represent the proportion of all 5,126 AUs. **A/S-Pr/-Po** – pre-/post-claim attack/support; **C** – claim; **C-Re** – restatement.

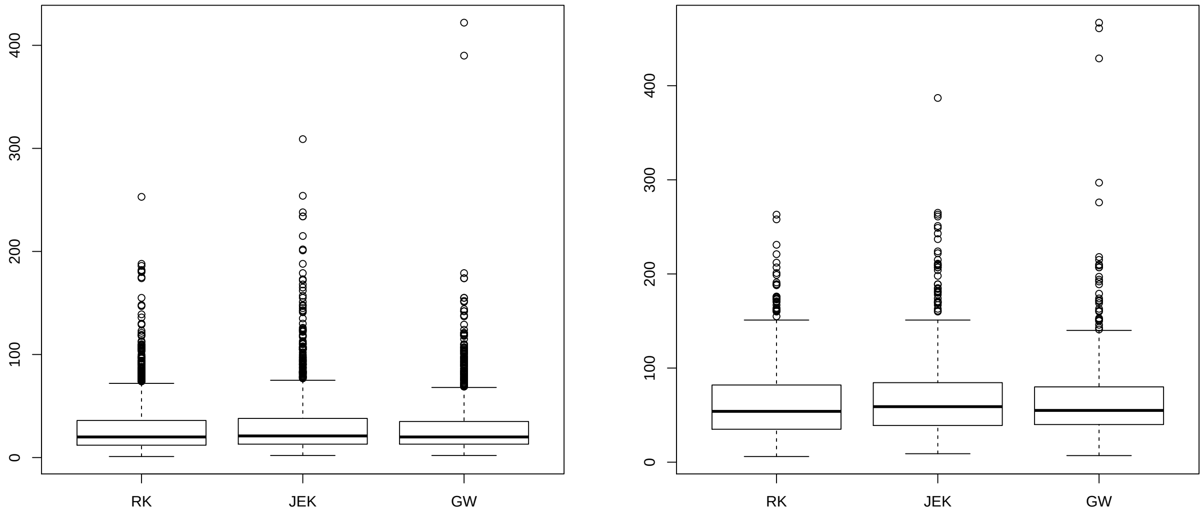
Anno.	*claim*	*premise*	Sum
RK	809 (47.8 %)	882 (52.2 %)	1691 (33.0 %)
JEK	782 (47.4 %)	868 (52.6 %)	1650 (32.2 %)
GW	832 (46.6 %)	953 (53.4 %)	1785 (34.8 %)
Sum	2,423 (47.3 %)	2,703 (52.7 %)	5126

(a) Distribution for `NO-R`

Anno.	*argument*
RK	781 (33.3 %)
JEK	759 (32.3 %)
GW	809 (34.4 %)
Sum	2349

(b) Distribution for `ARG`

Table 4.6.: Label distribution for `NO-R` and `ARG`. Percentages are defined as in Table 4.5.



(a) Lengths of original AUs

(b) Lengths of arguments

Figure 4.1.: Per-annotator length distributions (measured in tokens) as boxplot. The box represents the range from 25 % to 75 %-quartile with the median marked by a horizontal line. The whiskers denote the full range of values except for the outliers, which are marked with circles. Outliers are more than 50 % of the inter-quartile range (IQR) away from the 75 % quartile

4.7 Argument Unit Length Statistics

While the previous section presented statistics on the AU *counts*, this section focuses on the *length* of AUs. Figure 4.1 shows the distribution of AU lengths per annotator. Even though several „outliers“ exist, the general annotation behavior seems to be comparable across annotators as Table 4.7 shows in numbers.

An average AU is 29 tokens long, 25 % of AUs are shorter than 13 tokens, and 25 % of the AUs are longer than 37 tokens. The average argument is 65 tokens long, which corresponds to 2.2 AUs per argument.

4.8 Token Coverage Statistics

In this section, we analyze the distribution of AU labels on the 70,464 tokens in the corpus. Each token has a unique label because we resolved overlapping annotations during the post-processing.

Table 4.8 and Table 4.9 show that about three of four tokens (74.4 %) are covered by an AU, which indicates the argumentative nature of the corpus. The largest fraction of tokens is covered by post-claim supports (39.1 %) and claims (22.3 %). The smallest classes are restatement (0.6 %) and pre-claim attacks (1.2 %).

Annotator	Q_1	median	mean	Q_3
RK	12	20	28.49	36
JEK	13	21	30.67	38
GW	13	20	28.33	35

(a) Lengths of original AUs

Annotator	Q_1	median	mean	Q_3
RK	35	54	63.69	82
JEK	39	59	68.28	84.5
GW	40	55	64.84	80

(b) Lengths of arguments

Table 4.7.: Token count statistics for original AUs and for arguments. Q_1/Q_3 – 25 %-/75 %-quartile.

Anno.	C	C-Re	S-Pr	S-Po	A-Pr	A-Po	Non-AU
RK	16,718 (23.7 %)	465 (0.7 %)	6,581 (9.3 %)	25,289 (35.9 %)	515 (0.7 %)	2,488 (3.5 %)	18,408 (26.1 %)
JEK	16,514 (23.4 %)	431 (0.6 %)	5,870 (8.3 %)	27,507 (39%)	1,369 (1.9 %)	1,693 (2.4 %)	17,080 (24.2 %)
GW	13,885 (19.7 %)	391 (0.6 %)	4,783 (6.8 %)	29,890 (42.4 %)	629 (0.9 %)	2,288 (3.2 %)	18,598 (26.4 %)
Avg.	15,706 (22.3 %)	429 (0.6 %)	5,745 (8.2 %)	27,562 (39.1 %)	838 (1.2 %)	2,156 (3.1 %)	18,029 (25.6 %)

Table 4.8.: Token coverage statistics for original AUs as number of tokens covered by the different types of AU. The percentages represent the proportion of the 70,464 tokens. **Non-AU** – not annotated/argumentative

Anno.	AU	Non-AU
RK	52,056 (73.9 %)	18,408 (26.1 %)
JEK	53,384 (75.8 %)	17,080 (24.2 %)
GW	51,866 (73.6 %)	18,598 (26.4 %)
Average	52,435 (74.4 %)	18,028 (25.6 %)

Table 4.9.: Token coverage statistics for arguments. This table represents the segmentation of the text into argumentative and non-argumentative passages.

4.9 Confidence Statistics

This section provides an overview of the confidence score distribution per annotator (Table 4.10).

The vast majority of annotations is labeled with high confidence score (97.05 %), while only about 3 % received a medium confidence score. We see that RK uses medium-confidence annotations considerably more frequently than the other annotators. Low-confidence annotations are extremely rare (0.06 % of all annotations) and we discuss them in the following section.

Anno.	low	medium	high	Sum
RK	1 (0.06 %)	100 (5.91 %)	1,590 (94.03 %)	1,691 (32.99 %)
JEK	2 (0.12 %)	14 (0.85 %)	1,634 (99.03 %)	1,650 (32.19 %)
GW	0 (0.00 %)	34 (1.90 %)	1,751 (98.10 %)	1,785 (34.82 %)
All	3 (0.06 %)	148 (2.89 %)	4,975 (97.05 %)	5,126

Table 4.10.: Distribution of confidence scores by annotator. Percentages in the central field represent the fraction of AUs with a particular confidence per annotator. Percentages in the last column and last row correspond to the proportion with regard to all AUs.

4.9.1 Investigating Low-confidence Annotations

Five of the original annotations had a low confidence score, which was due to three reasons:

- Two cases were mistaken annotations that were supposed to get a medium confidence score.
- One annotator identified a premise but could not find the corresponding claim. In the original annotation guidelines we called such cases *implicit claims* (i.e., the claim is not stated explicitly) or *default claims* (i.e., the document's controversy is the claim). For the main study, we refrained from labeling implicit or default claims.
- The remaining two AUs are a pair of claim and premise, where the annotator could not figure out how the premise logically relates to the claim, even though the author's intention was clear.

4.10 Argumentation Patterns

This section analyzes the argumentation patterns in the corpus. By an *argumentation pattern* we understand the sequence of AUs of an argument.

Table 4.11 shows the 10 most frequent argumentation patterns; the list of all 51 argumentation patterns is located in Appendix D.

The shortest possible argumentation pattern is a single claim, which is the fourth most frequent pattern in our corpus (5.5 %). Almost three quarters of arguments (72.4 %) consist of one claim and one premise: 59.5 % for *C→S-Po* and 11.6 % for *S-Pr→C*. The corresponding patterns consisting of attack and claim are significantly less frequent (3.4 %): 2.1 % for *C→A-Po* and 1.3 % for *A-Pr→C* The ten most frequent argumentation patterns make up 94.3 % of all arguments.

The patterns *C→S-Po* and *C→S-Po→S-Po* are examples of seemingly equivalent patterns. Yet, we found that they are in fact different for two reasons: (1) Non-argumentative parts may appear between the supports, and (2) the two premises have different confidence scores.

The longest pattern with 9 AUs is *C→S-Po→A-Po→S-Po→S-Po→A-Po→S-Po→A-Po→A-Po*.

Rank	Pattern	Frequency	Percentage
1	C→S-Po	1398	59.5 %
2	S-Pr→C	272	11.6 %
3	S-Pr→C→S-Po	141	6.0 %
4	C	130	5.5 %
5	C→S-Po→A-Po	66	2.8 %
6	A-Pr→C→S-Po	51	2.2 %
7	C→A-Po	49	2.1 %
8	C→S-Po→C-Re	41	1.8 %
9	C→A-Po→S-Po	36	1.5 %
10	A-Pr→C	31	1.3 %
11–51	Other	134	5.7 %

Table 4.11.: List of the 10 most frequent argumentation patterns. **Pattern**: C – claim, C-Re – restatement, A/S-Pr/Po – pre-/post-claim attack/support. **Frequency** is the number of arguments having the given pattern. **Percentage** represents the portion of the pattern of all arguments.

4.10.1 Argumentation Patterns in Introduction and Conclusion

Table 4.12 shows the distribution of argumentation patterns in the first and last paragraph, which we use to analyze the differences in comparison to the overall statistics.

Introduction

Even though the proportion of single claims is larger compared to the overall statistics (21.8 %, compared to 5.5 %), the two combinations of support and claim are still the most frequent patterns with a share of 54.8 % of the 179 arguments intersecting with the first paragraph.

Conclusion

Contrary to the situation in introductions, single claims are the second most frequent pattern in conclusions (11.5 %). The proportion of 63.7 % of the 253 patterns, the fraction of claim-support patterns is even larger than in the introductions.

Rank	Pattern	Freq.	Pct.
1	C→S-Po	56	31.3 %
2	S-Pr→C	42	23.5 %
3	C	39	21.8 %
4	A-Pr→C	9	5.0 %
5	C→S-Po→A-Po	6	3.4 %
6	A-Pr→C→S-Po	5	2.8 %
7	S-Pr→C→S-Po	5	2.8 %
8	C→A-Po→S-Po	4	2.2 %
9	A-Pr→S-Pr→C	3	1.7 %
10	S-Pr→C→A-Po	3	1.7 %
11–17	Other	7	3.9 %

(a) Patterns in first paragraph

Rank	Pattern	Freq.	Pct.
1	C→S-Po	137	54.2 %
2	C	29	11.5 %
3	S-Pr→C	24	9.5 %
4	S-Pr→C→S-Po	14	5.5 %
5	C→S-Po→C-Re	10	4.0 %
6	C→A-Po	6	2.4 %
7	C→S-Po→A-Po	5	2.0 %
8	C→A-Po→S-Po	4	1.6 %
9	A-Pr→C	4	1.6 %
10	A-Pr→C→S-Po	3	1.2 %
11–21	Other	17	6.7 %

(b) Patterns in final paragraph

Table 4.12.: List of the 10 most frequent argumentation patterns in the final paragraph. The legend is equal to Table 4.11.

4.11 Pairwise Overlap of Argument Units

Measuring the average length of the original AUs and arguments is one way to compare annotation granularity among annotators (see Section 4.7). In this section, we use a custom metric, the pairwise overlap of AUs, to evaluate annotation granularity in another way..

The pairwise overlap for a pair of annotators (A_1, A_2) is the average number of A_2's annotations that each of A_1's annotations overlaps:

$$\mathrm{pwo}(A_1, A_2) = \frac{1}{|A_1|} \sum_{a_1\text{: annotation by } A_1} \left| \{a_2 \mid a_2 : \text{ annotation by } A_2 \wedge a_1 \text{ intersects with } a_2\} \right|$$

The larger the pairwise overlap is, the more coarse-grained are A_1's annotations compared to A_2.

Table 4.13 summarizes the pairwise overlap statistics for the original AUs; comprehensive statistics of pairwise overlap is located in Appendix E. In general, JEK tends to create longer annotations compared with the other annotators. Conversely, RK creates shorter annotations. However, the differences are not very large, so it may be doubted whether a significant difference exists.

A similar experiment with arguments (instead of AUs) yielded even less distinctive values as shown in Table 4.14.

$\downarrow A_1$ | $A_2 \rightarrow$	**RK**	**JEK**	**GW**
RK	–	1.015 (−0.002)	1.017 (−0.001)
JEK	1.017 (+0.002)	–	1.028 (+0.015)
GW	1.018 (+0.001)	1.013 (−0.015)	–

Table 4.13.: Average pairwise overlap on level ORIG for every annotator pair (A_1, A_2), with A_1 in rows and A_2 in columns. The values in parentheses are the difference to the entry symmetric to the major diagonal, which represents the pairwise overlap of (A_2, A_1).

$\downarrow A_1$ | $A_2 \rightarrow$	**RK**	**JEK**	**GW**
RK	–	1.006 (+0.002)	1.004 (−0.002)
JEK	1.004 (−0.002)	–	1.006(±0.000)
GW	1.006 (+0.002)	1.006 (±0.000)	–

Table 4.14.: Average pairwise overlap on level ARG. For more explanations, see Table 4.13.

5 Inter-annotator Agreement Analysis

We dedicate a whole chapter to the analysis of inter-annotator agreement (IAA) because it is such an important metric in corpus analysis. The term *inter-annotator agreement* describes the extent to which labels assigned to items agree.

Inter-annotator agreement metrics assess the reliability (*Can the data be reproduced?*) and validity (*Are the data meaningful with respect to the task?*) of the data [Artstein and Poesio, 2008, p. 556]. Artstein and Poesio [2008] explain in-depth when and how to apply several classic IAA metrics.

5.1 Agreement Metrics

Defining IAA in terms of annotation items is problematic in our scenario because we allowed for arbitrary annotation boundaries, which means that each annotator creates his „own“ annotation items.

There are two general strategies to solve this problem: Either we map the arbitrary-span annotations to annotation items and apply an item-based IAA metric, or we select an IAA metric that directly operates on span annotations. The following sections introduce the IAA metrics that are used in the corpus analysis.

5.1.1 Observed Agreement

Observed agreement – often denoted as A_o – is an item-based, baseline IAA metric: For two annotators, it calculates the percentage of items that the annotators agree on relative to the number of all items [Artstein and Poesio, 2008, p. 558].

A straight-forward extension for more than two annotators is to average the observed agreement over all (unordered) pairs of annotators [Fleiss, 1971].

5.1.2 Fleiss’ Kappa and Carletta’s K

Observed agreement does not take into consideration that agreement by chance. For example, two random annotators choosing uniformly among two categories will agree with a probability of 50 %.

Chance-corrected IAA metrics tackle this problem by taking into account the expected (dis)agreement, which can be estimated from the observed category distribution in the data.

For following analyses, we apply a generalized version of Scott’s π [Scott, 1955] that is known as Fleiss’ κ [Fleiss, 1971] or Carletta’s K [Carletta, 1996][1]. This metric supports an arbitrary number of annotators and categories. It assumes an identical label distribution for each annotator and is calculated as

$$\kappa = \frac{A_o - A_e}{1 - A_e},$$

[1] For the sake of brevity, we only refer to Fleiss’ κ in the following.

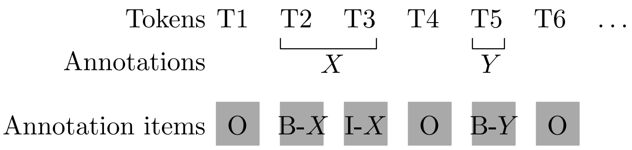


Figure 5.1.: Example of mapping to token-level annotation items (IOB scheme)

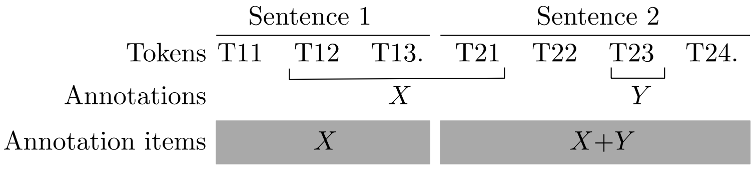


Figure 5.2.: Mapping of annotations X and Y to combined sentence-level annotation items. Annotation X is the only annotation that overlaps the first sentence (*T1 ... T3*), while the second sentence (*T21 ... T24*) is overlapped by X and Y.

where A_o is the observed agreement and A_e is the expected agreement. Therefore, if A_o equals A_e, κ becomes zero.

5.1.3 Agreement on Token-level

Mapping every token to an annotation item is a straight-forward representation of free annotation boundaries. The category of each token is the label of its (unique) covering AU.

Additionally, we prefix the first tokens in each AU with „B-" (*begin*) and any other covered token with „I-" (*inside*). Tokens outside of AUs are labeled with „O" (*outside*). This so-called *IOB scheme* was first introduced in the CoNLL 2000 shared task on chunking [Tjong Kim Sang and Buchholz, 2000].:

After the mapping, κ may be applied on the generated annotation items. The token-level mapping is depicted in Figure 5.1. We denote the observed token-level agreement with $A_{o,t}$ and the chance-corrected agreement with κ_t.

5.1.4 Agreement on Sentence-level

When using sentences as annotation items, we need to take into account that multiple AUs may intersect with one sentence. We decided to use the concatenated labels of all intersecting AUs as the category of the sentence. For example, if a sentence intersects with a claim and a post-claim support, the resulting sentence-level label resembles *claim+post-support*.

The sentence-level mapping is depicted in Figure 5.2. We denote the observed sentence-level agreement with $A_{o,s}$ and the chance-corrected agreement with κ_s.

5.1.5 Krippendorff's Unitized Alpha

Krippendorff's unitizing approach is a versatile IAA metric for arbitrary span annotations [Krippendorff, 1995]. It can also deal with multiple annotators out-of-the-box and belongs to the

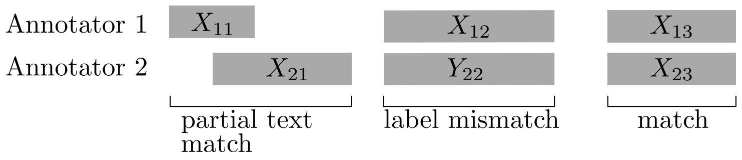


Figure 5.3.: Example of calculating the Jaccard agreement. This figure illustrates three typical constellations of overlapping annotations. Only X_{13} and X_{23} match in label *and* covered text. In this situation, we have 5 distinct annotations and an intersection of 1 annotation, and therefore, $j = 1/5 = 20\,\%$.

family of so-called α-agreement metrics. In general, α is calculated from (observed and expected) *disagreement* as follows:

$$\alpha = 1 - \frac{D_o}{D_e},$$

where D_o is the observed and D_e is the expected disagreement.

The disagreement is usually calculated from a user-defined distance function; thus, any property of the annotations can be taken into account. Yalçinkaya [2010] showed that α may be a suitable IAA metric for evaluating discourse treebanks.

The *unitized* α agreement α_u uses a three-dimensional space to represent annotations: Annotators are the first dimension, the different annotation categories/labels are the second dimension, and the text is the third dimension. Given a pair of annotator and category, the third dimension describes the sequence of annotations and unannotated passages (also called gaps) within the document.

Krippendorff [2004a] describes the precise calculation of α_u. Roughly speaking, the distance of two annotation spans depends on the length of their overlapping and non-overlapping parts.

5.1.6 Jaccard-based Agreement

Miltsakaki et al. [2004] used a relatively strict metric to evaluate the IAA for particular explicit DMs in the Penn Discourse Treebank. Their approach builds on the Jaccard similarity, which compares two sets, A_1 and A_2, by dividing the cardinality of their intersection by the cardinality of their union:

$$j(A_1,A_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}$$

We consider the annotations of each annotator as one set. Two AUs are equal if and only if they have the same label and the same covered text (*exact matching*). The Jaccard agreement is the average of the Jaccard similarity over all unordered pairs of annotators.

Obviously, the Jaccard agreement is conservative, since partial overlaps of annotations do not contribute to the score at all, which could be refined in future implementations.

An example is illustrated in Figure 5.3.

5.1.7 Wilson and Wiebe

Wilson and Wiebe [2003, Sec. 5.1] propose a custom IAA metric for an Opinion Mining annotation study with three annotators. Their IAA metric uses pairs of overlapping annotations as annotation items. The metric returns two values: one that signifies the amount of overlapping annotations (*overlap score*), and another that evaluates κ based on the pairs of overlapping annotations.

Step 1: Overlap score

The first step maps pairs of overlapping annotations to annotation items. Let's consider two sets of annotations A_1, A_2. The function $\text{agr}(A_1 \rightarrow A_2)$ is defined as the fraction of annotations in A_1 that overlap with at least one annotation in A_2; more precisely:

$$\text{agr}(a_1 \rightarrow a_2) = \frac{\left|\{a_1 \in A_1 \mid \exists a_2 \in A_2 : a_1 \text{ and } a_2 \text{ overlap } \}\right|}{\left|A_1\right|}$$

Note that the *agr* function is not symmetric. To calculate an aggregated value, we average *agr* over all ordered pairs of annotators. This aggregated value is called the *overlap score* and is denoted with ω_o (The number of all ordered pairs of n annotators is $n^2 - n$):

$$\omega_o = \frac{1}{n(n-1)} \sum_{i \in \{1,\ldots,n\}} \sum_{j \in \{1,\ldots,n\}, i \neq j} \text{agr}(A_i \rightarrow A_j)$$

Step 2: Label agreement

In the second step, the overlapping annotations of each annotator pair serve as annotation items for an evaluation of the annotation labels with Fleiss' κ. As in the first step, we average over all pairs of annotators and denote the result with ω_l.

Note that this time we iterate over all *unordered* pairs of annotators, as the definition of the annotation items is symmetric. With n annotators and $N = \frac{1}{2}n(n+1)$ unordered annotator pairs, we obtain:

$$\omega_l = \frac{1}{N} \sum_{i \in \{1,\ldots,n\}} \sum_{j \in \{i+1,\ldots,n\}} \kappa(A_i, A_j)$$

Figure 5.4 illustrates the algorithm.

5.1.8 Why do we Need Six IAA metrics?

It is not common to evaluate a corpus with six IAA metrics in parallel and Krippendorff's α_u was our method of choice because it is tailored to the task. However, we know of only few published works that use Krippendorff's unitized alpha (e.g., [Kolhatkar et al., 2013; Kolhatkar and Hirst, 2012]), what makes our results difficult to compare. Therefore, we included several established metrics based on token- and sentence-level annotation items.

The metrics adopted from Wilson-Wiebe and the PDTB are rather experimental and we were interested in how far their values compare to the other metrics.

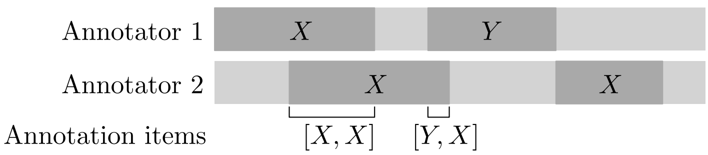


Figure 5.4.: Evaluating two annotators' annotations with the two-step approach by Wilson and Wiebe.
Step 1: Both of Annotator 1's annotations overlap Annotator 2's first annotation, producing the two annotation items $[X,X]$ and $[X,Y]$.
Therefore, we find an overlapping annotation for any of Annotator 1's annotations, i.e., $\text{agr}(A_1,A_2) = 1$. Vice versa, only one of Annotator 2's annotations is intersects with Annotator 1's annotations, i.e., $\text{agr}(A_2,A_1) = 0.5$. This yields $\omega_o = 0.75$.
Step 2: According to Fleiss' κ, the label agreement for the annotation items $[X,X],[X,Y]$ is $\omega_l = -0.33$.

Metric	Implementation
$A_{o,t}$,$A_{o,s}$	d.t.u.d.s.agreement.MultiRaterObservedAgreement
κ_t,κ_s	d.t.u.d.s.agreement.MultiRaterPiAgreement
α_u	d.t.u.d.s.unitizing.AlphaUnitizedAgreement
ω_o,ω_l	de.tudarmstadt.ukp.rk.mt.iaa.WilsonWiebeAgreement
j	de.tudarmstadt.ukp.rk.mt.iaa.JaccardAgreement

Table 5.1.: Implementation of the inter-annotator agreement metrics in Java (*d.t.u.d.s = de.tudarmstadt.ukp.dkpro.statistics*)

5.1.9 Implementation

Our implementation builds on DKPro Statistics[2], an open-source implementation of various IAA metrics. Table 5.1 shows the mapping between IAA metrics and implementing classes. Of course, we implemented the non-standard Wilson-Wiebe and Jaccard metrics.

5.2 Agreement vs. Confidence

This section demonstrates the influence of filtering by confidence scores on the IAA scores. Table 5.2 presents the average agreement resulting when AUs below a certain minimum confidence are filtered out.

Due to their small contribution, low-confidence annotations do not affect the agreement much. In contrast, the IAA *declines* when we keep only the high-confidence AUs: Only ω_l increases (by 0.79 %), while the other metrics drop by 0.25 pp to 1.12 pp. This may well be a result of the different per-annotator distributions of confidence scores (see Section 4.9).

[2] *Darmstadt Knowledge Processing Repository*, `https://code.google.com/p/dkpro-statistics/`

Even though keeping low-confidence annotations would not affect the IAA much, we excluded them from the further analysis, which is in accordance with the annotation guidelines [Kluge, 2013, p. 6].

Distributional Statistics

In the following sections, we always average the IAA scores over all documents in the corpus. While this allows us to work with one score per IAA metric, averaging hides virtually all distributional characteristics of the annotations.

Table 5.3 describes the dataset in terms of common statistical assessment metrics. The metrics standard deviation, inter-quartile range (IQR), and range signify the „width“ of the distribution. IQR is the least sensitive metric with respect to outliers. The largest variance can be observed for α_u (between −22.8 % and 96.1 %), while ω_o appears to be relatively uniform with an IQR of 9.5 %. The distribution of sentence-based, token-based, and the Jaccard metric show relatively homogeneous width in terms of standard deviation and IQR.

Min con.	**#AUs**	$A_{o,t}$ **[%]**	κ_t **[%]**	$A_{o,s}$ **[%]**	κ_s **[%]**	α_u **[%]**	j **[%]**	ω_o **[%]**	ω_l **[%]**
low	5,126	60.96	44.16	60.83	45.17	40.19	27.15	86.42	43.13
medium	5,123	61.00	44.23	60.87	45.24	40.20	27.16	86.42	43.15
high	4,975	60.31	43.00	60.37	44.12	39.53	26.91	85.13	43.94
m. → h.	−148	−0.69	−1.23	−0.50	−1.12	−0.67	−0.25	−1.29	0.79

Table 5.2.: Inter-annotator agreement by minimal confidence score. **Min con.**: Minimal confidence score of the AUs. **#AUs**: Number of AUs with at least the minimal confidence. **m. → h.**: Difference between *high* and *medium* minimal confidence.

Metric	$A_{o,t}$ **[%]**	κ_t **[%]**	$A_{o,s}$ **[%]**	κ_s **[%]**	α_u **[%]**	j **[%]**	ω_o **[%]**	ω_l **[%]**
Min	34.2	13.4	32.1	11.9	−22.8	5.3	64.2	1.9
Q_1	54.1	34.9	53.8	37.8	23.4	18.4	82.2	30.7
Median	60.8	43.5	60.7	45.0	40.3	27.2	87.1	41.2
Mean	61.0	44.2	60.9	45.2	40.2	27.2	86.4	43.2
Q_3	68.4	53.1	67.0	52.2	57.7	33.8	91.7	54.3
Max	86.5	79.3	87.9	80.8	96.1	60.1	100.0	93.1
Std. dev.	10.2	13.1	10.7	13.3	23.9	11.6	7.5	17.1
IQR	14.3	18.2	13.2	14.3	34.3	15.4	9.5	23.6
Range	52.3	65.9	55.7	68.9	118.9	54.8	35.8	91.1

Table 5.3.: Distributional statistics of the IAA scores. Q_1, Q_3: 25 %-/75 %-quartile. **IQR** = $Q_3 - Q_1$. **Range** = Max - Min.

5.3 Agreement vs. Generalization Level

This section compares how IAA varies depending on the generalization level. Our analysis strategy in this section resembles the *category distinction test* proposed by Krippendorff [2012][3]: By abstracting from certain characteristics of the annotations, we may deduce, which aspects of the annotation task the annotators understood better or worse.

5.3.1 Characteristics of Generalization Levels

The analysis applies four generalization levels: `PR`, `NO-R`, `AU`, and `ARG`. The following paragraphs explain, what a change in IAA on each of these generalization levels may indicate. We list the number of labels for each level because a reduced number of labels generally increases scores for non-chance-corrected metrics due to higher probability of agreement by chance.

Level *Premises* (4 labels)

This level drops the distinction between support and attack. An increased IAA may indicate that distinguishing support from attack was difficult. Yet, we hypothesize that the distinction should be mostly clear and that no relevant improvement occurs.

Another reason for increased agreement could be that the choice for support or attack depends on the polarity of the corresponding claim.

Level *No Relations* (2 labels)

This level ignores the attributes indicating polarity, direction, and restatement relation. Higher agreement on this level indicates that the annotators had problems to identify inter-AU relations.

Furthermore, the results for this level could be compared to an IAA evaluation of an annotation scheme with a graph-based relation representation.

Levels *Argument Units* and *Arguments* (1 label)

Generalizing to one of these levels yields a segmentation of the documents into argumentative or non-argumentative passages. A high IAA score on this level indicates that the annotators agree on the argumentative segmentation, either on the level of AUs or on the (coarse-grained) level of arguments.

5.3.2 Results

Table 5.4 shows the IAA per generalization level. It is not surprising that the overlap score ω_o is the same for `PR`, `NO-R`, and `AU` because the AU spans remain the same – only their labels change. Note that calculating ω_o agreement for `ARG` and `AU` is not possible, as only one label exists and each overlapping pair is guaranteed to agree on that label.

As expected, `PR` yields the least improvement among the generalization level. Dropping the relational attributes (`NO-R`) resulted in slight improvements, also for the chance-corrected metrics.

[3] [Peldszus and Stede, 2013] applied this test in their argumentation annotation study.

Level	$A_{o,t}$ [%]	κ_t [%]	$A_{o,s}$ [%]	κ_s [%]	α_u [%]	j [%]	ω_o [%]	ω_l [%]
ORIG	61.0	44.2	60.9	45.2	40.2	27.2	86.4	43.2
PR	62.7	45.1	62.6	46.3	41.7	27.5	86.4	44.4
	(+1.7)	(+0.9)	(+1.7)	(+1.1)	(+1.5)	(+0.3)	(± 0)	(+1.2)
NO-R	69.0	49.1	68.2	49.7	49.6	28.8	86.4	49.4
	(+8.0)	(+4.9)	(+7.3)	(+4.5)	(+9.4)	(+1.6)	(± 0)	(+6.2)
AU	79.0	50.0	77.0	50.0	56.6	31.6	86.4	–
	(+18.0)	(+5.8)	(+16.1)	(+4.8)	(+16.4)	(+4.4)	(± 0)	
ARG	79.4	46.1	80.3	49.2	60.0	23.6	90.5	–
	(+18.4)	(+1.9)	(+19.4)	(+4.0)	(+19.8)	(−3.6)	(+4.1)	

Table 5.4.: Agreement on different generalization levels. The differences compared to the original AUs are enclosed in parentheses and measured in percentage points.

For ARG and AU, we recognize that $A_{o,t}$, $A_{o,s}$, and α_u increase considerably, while the changes to κ_t and κ_s are at most moderately positive, so the improvement probably only resulted from the reduced number of categories.

For the label-generalizing levels PR, NO-R, and AU, we could expect that the Jaccard agreement increases because the annotation spans remain the same, but the number of categories decreases compared to ORIG. Therefore, the improvement may probably result from agreement by chance. The drop in Jaccard agreement for ARG is sensible because it is difficult to find arguments with the exact same text span. The same cause may be responsible for the increased overlap score.

As a result, we could not observe dramatic changes in the IAA on any generalization level. The results of the chance-corrected metrics may be the most reliable ones because they take into account the reduced number of categories.

5.4 Agreement vs. Topic Order

This section presents an analysis of IAA by topic and of IAA (since time because the annotators processed the topics in a predefined order). The order of topics was: *master*, *g8*, *lehrer*, *promovieren*, *sport*, and *sitzenbleiben*.

Figure 5.5 shows no obvious trend over time. The topic *g8* exhibits a relatively low agreement for most metrics, and *promovieren* and *sport* perform best for the sentence- and token-based metrics, but also for α_u and Jaccard.

For a quantitative analysis, we applied two common rank correlation coefficients: Spearman's ρ_s and Kendall's τ. The correlation scores were calculated with the *cor* function from the *stats* package in R[4]. The correlation scores may not be reliable if the sample is too small, so we additionally calculated confidence values for ρ_s using the *rcorr* function in the *Hmisc* package[5]. Table 5.5 shows the results. The sentence-level agreement correlates best with the order of topics, but none of the correlation scores is significant (α=5 %). This means that in most cases a correlation different from 0 is improbable.

[4] http://stat.ethz.ch/R-manual/R-patched/library/stats/html/cor.html

[5] http://cran.r-project.org/web/packages/Hmisc/

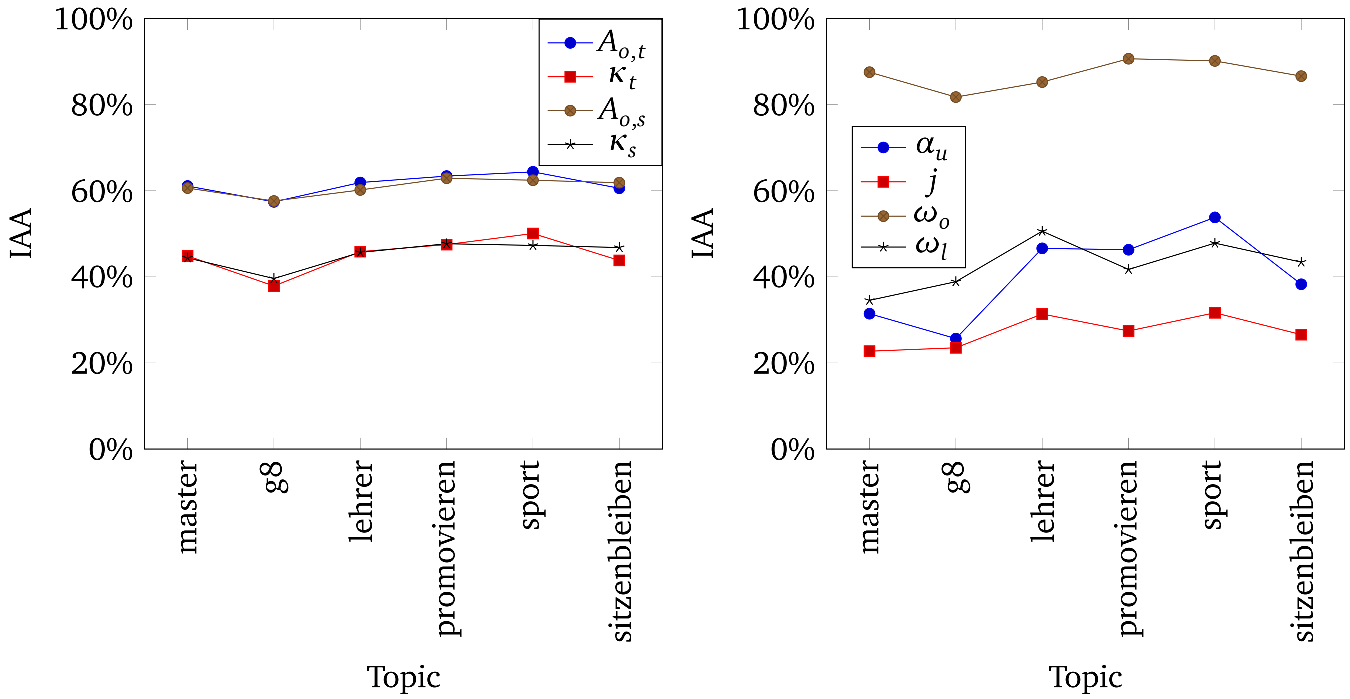


Figure 5.5.: Inter-annotator agreement by topic. The topics were annotated in the displayed order.

Coefficient	$A_{o,t}$	κ_t	$A_{o,s}$	κ_s	α_u	j	ω_o	ω_l
ρ_s	0.31	0.31	0.60	0.71	0.54	0.60	0.31	0.60
p-value for ρ_s	0.54	0.54	0.21	0.11	0.27	0.21	0.54	0.21
τ	0.33	0.33	0.33	0.47	0.33	0.47	0.20	0.47

Table 5.5.: Correlation of inter-annotator agreement with topic order, evaluated with Spearman's ρ_s and Kendall's τ.

Due to the annotators' weekly consultations, we hoped to find a (significant) positive correlation between IAA and the passed time, but the qualitative and quantitative analyses did not support this hypothesis.

5.5 Agreement vs. Annotation Time Demand

In the previous section, we showed that no correlation between IAA and the order of topics exists. This section presents a similar analysis for the annotation time demand. For our further analysis, we focus on α_u as single IAA metric.

The scatter plot in Figure 5.6 hints at a negative correlation between IAA and annotation time demand. The quantitative correlation analysis in Table 5.6 shows that there is indeed a slight and almost significant negative correlation between annotation time and α_u if we accept a higher significance level of α=7%.

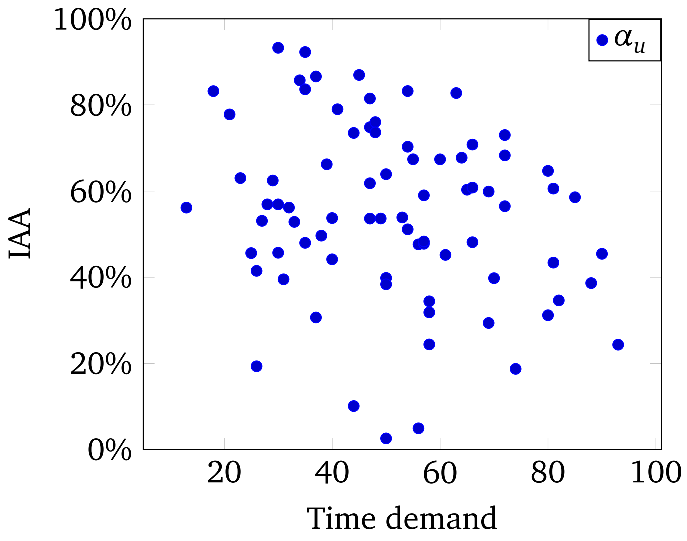


Figure 5.6.: Inter-annotator agreement in terms of α_u vs. annotation time demand. Every data point corresponds to one document.

Metric	Total time	Token rate	Sentence rate
ρ_s	−0.21	0.11	0.16
p-value for ρ_s	6.7 %	35.3 %	14.7 %
τ	−0.15	0.07	0.11

Table 5.6.: Correlation of inter-annotator agreement with annotation time demand, evaluated with Spearman's ρ_s and Kendall's τ. **Total time:** Sum of annotation time per document. **Token/Sentence rate:** Number of tokens/sentences divided by the total time.

We expected that documents that need more time to be annotated, may receive a smaller IAA. Indeed, we observed a slight negative correlation between annotation time demand per document and IAA in terms of α_u.

5.6 Pairwise Agreement

This section compares the IAA performance of annotators pairs with the combined IAA scores.

Table 5.7 shows that, on average, the annotator pair JEK/GW has the worst average IAA (Δ=−1.3 pp), while RK/GW perform best (Δ=+1.1 pp). The results of the annotator pair RK/JEK were identical to the combined results. However, the differences are not large in magnitude: The maximum (absolute) deviation is 3.2 pp, 6 difference entries in the table are below 2.0 pp, and the remaining 11 entries are below 1.0 pp.

We may conclude that there is no relevant difference between the performance of any annotator pair and the combined IAA scores.

Annotators	$A_{o,t}$ [%]	κ_t [%]	$A_{o,s}$ [%]	κ_s [%]	α_u [%]	j [%]	ω_o [%]	ω_l [%]
All	61.0	44.2	60.9	45.2	40.2	27.2	86.4	43.2
Δ RK/GW	+1.1	+1.4	+0.9	+1.0	+3.2	−0.2	0.2	1.5
Δ RK/JEK	+0.2	+0.2	+0.2	−0.1	−0.9	+0.3	+0.2	+0.0
Δ JEK/GW	−1.3	−2.5	−1.0	−1.8	−1.8	−0.1	−0.4	−1.4

Table 5.7.: Inter-annotator agreement by annotator pairs. The first row contains the average agreement scores. All other rows contain the difference in agreement for the annotator pair in comparison to the combined IAA, measured in percentage points. A negative number indicates worse performance.

5.7 The Maximum Overlap Normalization Algorithm

Our annotation guidelines did not enforce strict annotation boundaries, but we instructed the annotators to annotate only clauses or sentences, whenever possible. This section describes an approach to „normalize“ differences in the annotation lengths, which could be suitable as a pre-processing step for creating a gold standard.

Example 5.1 shows annotations in the corpus where both annotators have identified the same claim, but premises of different length. Even though this disagreement did not result from unclear annotation boundaries (all annotations start and end at sentence boundaries), a sensible normalization strategy could help to merge both annotations into a common gold standard. In this case, we would accept GW's annotations as they are subsumed by JEK's annotations.

Example 5.1: Different annotation granularity (*g80*)
(Claims appear in **boldface** type and supports are *italicized*.)
JEK:
Doch erhöht ein Jahr Altersunterschied wirklich die Chancen auf dem Jobmarkt? *Zumal die Buben wegen der Abschaffung der Wehrpflicht jetzt früher fertig sind. Im G 9 haben zudem mehr Schüler Zeit im Ausland verbracht, was von Arbeitgebern positiv bewertet wurde. Firmen achten ja nicht nur auf Noten und Alter, sondern auch auf die Persönlichkeit.*

GW:
Doch erhöht ein Jahr Altersunterschied wirklich die Chancen auf dem Jobmarkt? Zumal die Buben wegen der Abschaffung der Wehrpflicht jetzt früher fertig sind. *Im G 9 haben zudem mehr Schüler Zeit im Ausland verbracht, was von Arbeitgebern positiv bewertet wurde. Firmen achten ja nicht nur auf Noten und Alter, sondern auch auf die Persönlichkeit.*

5.7.1 Algorithm

The normalization algorithm works as follows: Given an annotator pair, we look for pairs of overlapping annotations and reduce the span of both annotations to their maximum overlap. If one annotation overlaps with multiple other annotations, each overlap is handled separately.

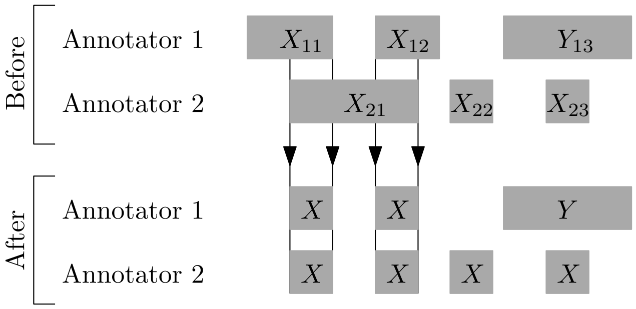


Figure 5.7.: Example of calculating maximum overlap for 2 annotators. X and Y indicate annotation labels. In case of multiple overlapping annotations, each pair is treated separately (X_{11},X_{12},X_{21}). Annotations without overlapping partners (X_{12}) and overlapping annotations with different labels (Y_{13},X_{23}) are not touched.

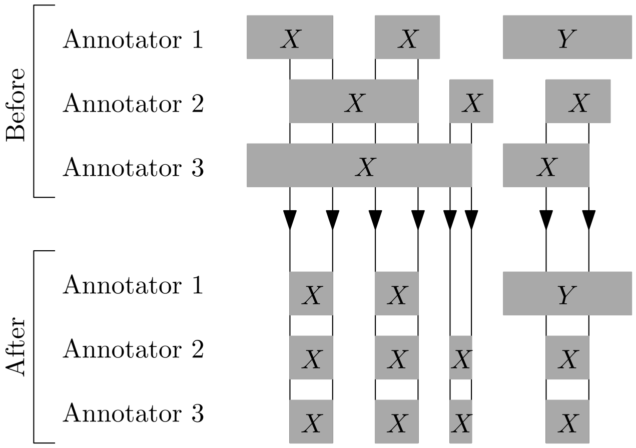


Figure 5.8.: Example of calculating maximum overlap for 3 annotators. Legend: see Figure 5.7.

Thus, long original AUs may produce several smaller AUs. Figure 5.7 showcases how the algorithm works for two annotators.

For more than two annotators, the same procedure can be applied iteratively for all unordered pairs of annotators. An example with three annotators is depicted in Figure 5.8.

The algorithm is implemented in the class `MaximumOverlappingSpanAnnotator`.

5.7.2 Results

Since the algorithm splits AUs that overlap multiple AUs by other annotators, the total number of AUs increased by 125 from 5,123 to 5,248.

The effects of the maximum overlap normalization on IAA are shown in Table 5.8. We expected the decrease in ω_o because the number of AUs increases but the number of overlaps remains the same. The increase in j is also sensible because the algorithm artificially adjusted the annotation spans.

Level	$A_{o,t}$ [%]	κ_t [%]	$A_{o,s}$ [%]	κ_s [%]	α_u [%]	j [%]	ω_o [%]	ω_l [%]
ORIG	61.0	44.2	60.9	45.2	40.2	27.2	86.4	43.2
MON	66.7	51.7	65.3	50.7	53.5	46.9	78.0	65.6
	(+5.7)	(+7.4)	(+4.5)	(+5.4)	(+13.3)	(+19.7)	(−8.4)	(+22.5)

Table 5.8.: Effect of maximum overlap normalizer (MON) on inter-annotator agreement. The differences compared to the original AUs are enclosed in parentheses and measured in percentage points.

5.8 Discussion

This chapter investigated IAA from various perspectives. We analyzed IAA with respect to confidence scores, generalization levels, the change over time, annotation time demand, and annotator pairs. Additionally, we proposed a method to prepare a gold standard, called the maximum overlap normalization algorithm.

In the following paragraphs, we put our results on IAA into context. We found general scales for Fleiss' κ and Krippendorff's α_u and also related work in AM. We should note, however, that their comparability is limited because our annotation scheme differs or the authors do not report on the precise calculation of their IAA scores.

According to Kundel and Polansky [2003, Table 2], the chance-corrected token- and sentence-level agreement scores are on the border between fair and moderate agreement. Regarding unitized alpha, Krippendorff [2004b] suggests that agreement above 67 % is considered acceptable, and agreement above 80 % is considered perfect, so that our results would be considerably less than acceptable.

Miltsakaki et al. [2004] report excellent IAA scores for the Jaccard metric: For 10 connectives of type *subordinating conjunction* and *adverbial*, they achieved average IAA scores of 90.2 % (if relation arguments were considered separately) and 82.8 % (if relation arguments were counted together). In comparison, our results for this metric are at ca. 30 %. However, the position of the annotations in their task was already relatively „fixed" in the text (due to the presence of the particular DMs) and the annotators had to select the two relation arguments around the DM. In our case, no such „assistance" existed.

Wilson and Wiebe [2003] report overlap scores (ω_o) of 83 % to 91 % and agreement scores (ω_l) of 67 % to 84 %. These results are considerably better than ours: While the scores for ω_o are comparable (ca. 86 %), ω_l is around 43 % in our dataset. However, the agreement scores are hardly comparable because both studies were carried out under entirely different conditions: Wilson and Wiebe [2003] annotated subjective and objective speech acts in English news documents (13 documents, 210 sentences) and we annotated interlinked AUs in German Web documents (79 documents, ca. 3000 sentences).

Peldszus and Stede [2013] report $\kappa = 38.4\%$ and $\alpha = 42.5\%$[6] for a classroom annotation study with 23 annotators, which is comparable to our results. However, their documents were artificial data and the annotation items were pre-defined to be sentences.

[6] Note that this is *not* unitized alpha!

Bal and Dizier [2010] obtained $\kappa = 80\,\%$ for the task of marking claims and premises in newspaper editorials with two annotators. Still, they do not report how they determined this value.

White [2010] presents a sentence-level annotation study of scientific documents with 3 annotators, who annotated 400 sentences according to two annotation models, a custom model by White and an adapted version of Toulmin's model. The annotators agreed perfectly for 60.5 %/39.25 % of the sentences and disagreed completely for 3.75 %/8.25 % of the sentences. In the remaining 35.75 %/52.5 % cases, exactly two annotators agreed. In our dataset, 46.0 % of the sentences achieve perfect agreement, in 43.4 % of the cases, two annotators agree, and in 10.6 % of the cases, all annotators disagree. Thus, the fractions of cases with perfect agreement and majority agreement are smaller (46 % vs. 49.9 %[7] and 43.4 % vs. 44.1 %), while the fraction of cases with complete disagreement is larger in our case (10.6 % vs. 6.0 %), but in general, the results appear to be comparable.

[7] For simplification, we use the mean of both scores.

6 Experiments

This chapter describes a series of experiments that we performed on the corpus. These experiments investigate how DMs serve to discriminate claims and premises in our corpus. A significance analysis and a feature selection experiment evaluate the DMs in isolation, and several Machine Learning experiments evaluate the DMs in combination and give a taste of their usefulness as features for Argument Extraction.

6.1 Discourse Marker Resources

We built three lists of DMs for our experiments originate including both discourse connectives and discourse particles.

Particles list

A total of 28 semantically categorized particles from a large German grammar constitute the first list [Helbig, 1996, pp. 481–484]. These particles are also called intensifiers [Quirk et al., 1980] and include the subgroups of amplifiers (e.g., ganz („*quite*")) and downtoners (e.g., nur („*only*")).

PDTB-DM

We compiled the second list with 51 discourse connectives, called PDTB-DM, based on a manual translation of the DMs listed in Appendix B of the PDTB annotation guidelines, which is a frequency distribution of all DRs that are expressed by *explicit* DMs in the PDTB [Prasad et al., 2007]. To reduce the manual translation effort, we only considered DMs expressing a particular DR sense at least 10 times. Furthermore, we merged *contrast* and *concession* into the more general DR *comparison*, since both DRs are expressed by almost the same set of DMs[1]. Based on the remaining 6 DR senses, we translated the English DMs into German, using semantic groups of adverbs, prepositions, and connectives listed in Helbig [1996, pp. 453–474].

DiMLex

The third resource is a large lexicon of German connectives called DiMLex [Berger et al., 2002; Stede and Heintze, 2004][2]. DiMLex lists ca. 170 DMs along with syntactic and semantic information including their part of speech (e.g., subordinating conjunction, preposition) and DR annotations from 13 DR categories (e.g., *contrast*, *cause*, *concession*).

We identified differences regarding the granularity and names of DRs in PDTB-DM and in DiMLex, as summarized in Table 6.1.

[1] listed at least 10 times in the Appendix B of the PDTB annotation guidelines

[2] *Discourse Marker Lexicon*; DiMLex is available as part of the annotation tool *ConAno* [Stede and Heintze, 2004], see also: `http://www.ling.uni-potsdam.de/acl-lab/Forsch/pcc/pcc.html`

PDTB	DiMLex
comparison	*concession, contrast*
reason, result	*cause*
conjunction, specification	*elaboration*
alternative	*contrast*

Table 6.1.: Correspondence between DRs in PDTB-DM and DiMLex.

6.2 Significance Testing Experiment

The first experiment is a two-sample statistical test to find out whether the two classes claim and premise are significantly different regarding the number of occurrences of the DMs and semantic groups of DMs in our three resources.

We considered the two classes as random samples of word form tokens. For each DM (or semantic group of DMs), we tested the null hypothesis that there is no difference between the proportion of DM tokens in the sample of type claim and the proportion of DM tokens in the sample of type premise. If the computed significance value p was below a significance level α, we would reject the null hypothesis, meaning that there was indeed a difference between claims and premises regarding the occurrence of DMs.

By applying a statistical significance test to compare two language samples, we do not claim that language is random in any way. Instead, it is a common approach in statistics to artificially introduce randomness in order to model deterministic, but very complex phenomena, such as natural language [Breiman, 2001].

6.2.1 Fisher's Exact Test

As test we chose Fisher's exact test, a non-parametric randomization test that makes no assumptions about the underlying probability distribution of the DMs [Fisher, 1932].

The two classes claim and premise yield a 2x2 contingency table. Table 6.2 describes a general contingency table.

	Class A	Class B	Row totals
Positives	a	b	a + b
Negatives	c	d	c + d
Column totals	a + c	b + d	a + b + c + d

Table 6.2.: General shape of a contingency table. A contingency table summarizes how often a particular result is observed (positives) or not observed (negatives) in the populations of the classes A and B. The column totals describe the sizes of the populations in class A and B ($a+c$ and $b+d$) and the row totals summarize the ratio of positive observations ($a+b$) to the negative observations ($c+d$) in the combined population $A \cup B$, which is of size $a+b+c+d$.

It is out of scope of this thesis to describe how Fisher's exact test works in detail[3]. Roughly speaking, the test evaluates how probable the given contingency table is among all the contingency tables with the same marginal totals (row and column totals).

In Example 6.1, the marginal totals plus the upper-left entry (matching tokens for class premise) unambiguously determine the remaining three cells; therefore 113 (in general $a + 1$) contingency tables with the same marginal totals exist. Under the null hypothesis that there is no association between the distribution of *wie* in the classes claim and premise, the probability of obtaining this contingency table is given as:

$$\begin{aligned} p_{\text{ref}} &= \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d!)}{(a+b+c+d)!\,a!\,b!\,c!\,d!} \\ &= \frac{(112+39)!\,(37467+13094)!\,(112+37467)!\,(39+13094)!}{(112+39+37467+13094)!\,112!\,39!\,37467!\,13094!} \\ &= 0.074099 \end{aligned}$$

In the two-sided version of Fisher's exact test, the confidence value is the sum over all probabilities (of contingency tables) equal to or lower than p_{ref}.

6.2.2 Experimental Setup

Lacking semantic discourse relation annotations, we counted surface word forms of the lexical items listed in our three DM resources. We only counted single word DMs and continuous multi-word DMs (n-grams). Sentence initial DMs were counted separately to capture DRs being signaled by a sentence initial position. We considered it the best to perform the tests for each annotator's dataset separately to avoid interference's.

We performed all statistical analyses with R and used the implementation of Fisher's exact test from the *stats* package[4]. For each DM, we computed a contingency table containing the number of observed occurrences in the two samples as illustrated in Example 6.1. For each semantic groups of DMs given in DiMLex and PDTB-DM, we calculated the per-group contingency table by adding up the contingency tables for each DM in that group.

Example 6.1: Contingency table for DM wie („*how/as*“)

The following contingency table shows the distribution of the DM *wie* for annotator GW. The DM makes up 39 (0.298 %) of the 13,094 tokens covered by claims and 112 (0.299 %) of the 37,467 tokens covered by premises.

	claim	premise	Row totals
Matching tokens	39	112	151
Non-matching tokens	13,055	37,355	50,410
Column totals	13,094	37,467	50,561

[3] For more detail, see for example `http://www.sheffield.ac.uk/polopoly_fs/1.43998!/file/tutorial-9-fishers.pdf`

[4] `http://www.r-project.org/`

DM	Discourse semantics	AU	Rank	Sign.	Pred.
also („*therefore*“)	D-cause, P-result	claim	36	3	–
Doch („*however*“)	D-contrast, D-elaboration, P-comparison	claim	20	3	–
jedoch („*though*“)	D-contrast, D-elaboration, P-comparison	claim	35	3	–
sondern („*but*“)	D-contrast, P-comparison	claim	23.5	2	–
ganz („*quite*“)	HB-amplifier	claim	27	3	1
Denn („*as*“)	D-cause, P-reason	premise	24	3	3
weil („*because*“)	D-cause, P-reason	premise	21	–	1
oder („*or*“)	D-contrast, P-Alternative	premise	4	2	3
und („*and*“)	D-elaboration	premise	1	2	3
auch („*also*“)	D-elaboration	premise	2	–	3
etwa („*roughly*“)	HB-downtoner	premise	15	3	3
nur („*only*“)	HB-downtoner	premise	6	–	2
um („*in order to*“)	D-purpose	premise	9	–	3
als („*when*“)	D-circumstance, D-elaboration	premise	3	–	3
wie („*as*“)	D-circumstance, D-elaboration	premise	7	–	3

Table 6.3.: Overview of most significant and most predictive (in terms of IG) DMs. **Disc. Semantics**: *D* – DiMLex; *P* – PDTB; *HB* – Helbig-Buscha. **AU**: Predicted AU according to odds ratio (Sign.) and/or percentage of presence (Pred.). **Rank**: Median of the three ranks according to frequency. **Significance (Sign.)**: Number of annotators, for which a DM is significant (α=5 %), '–' means less than 2. **Predictiveness (Pred.)**: Number of annotators, for which a DM appears within the top 10 most predictive DMs; '–' means 0.

6.2.3 Results

Table 6.3 shows the DMs that occur with significantly different frequency in the two classes of claims and premises in at least two of the annotated three datasets (column *Sign.*). We also included the discourse semantics based on PDTB-DM and DiMLex; for highly ambiguous DMs, we only list the DRs relevant for the genre of argumentative texts.

The table shows that claims and premises are significantly different regarding the occurrences of DMs expressing *result*, *reason*, *concession*, and *contrast*. Furthermore, particular discourse particles are indicative of claims and premises. For premises, these are the downtoners etwa („*roughly*“) and nur („*only*“), and for claims the amplifier ganz („*quite*“).

We found several groups of DMs (given by DRs in PDTB-DM (*P*-) and DiMLex (*D*-)) to be significant as a whole (α=5 %). For claims, these are the groups *P-comparison* (e.g., gleichwohl („*however*“)) and *P-result* (e.g., also („*therefore*“)). For premises, the groups *P-alternative* (e.g., oder („*or*“)), *P-reason* (e.g., weil („*because*“)), and *D-sequence* (e.g., dann („*then*“)) are significant.

6.3 Feature Selection Experiment

The preceding significance tests show that particular DMs occur with significant difference in claims and premises, and could therefore qualify as distinctive features for telling apart claims from premises.

An alternative approach to identify distinctive features is to rank the DMs according to their Information Gain (IG), a method that is often used in Machine Learning to select the most predictive features and to avoid overfitting. In our second experiment, we considered all the DMs from the three resources as features and ranked them by IG separately for each annotator using the Weka data mining framework [Hall et al., 2009].

We extracted a set of binary features for each AU: Every DM maps to one feature that is 1 if the DM can be found in the AU and 0 if not. The results are summarized in column *Pred.* of Table 6.3.

6.4 Classification Experiments

Significance analysis and the feature selection experiment evaluated the DMs in isolation, so we performed a third experiment to examine the predictiveness of all DMs *in combination* by training several Machine Learning algorithms on the annotated datasets; therefore, DMs are the *only* features in this setup.

We specified two classification problems that test whether the presence of DMs can be used (i) to identify argumentative sentences, and (ii) to distinguish claims from premises. The experiments were carried out as 10-fold cross validation using Weka.

It is well-known in statistical classification that there is no single best classifier that is superior to all other classifiers in all situations (cf. Problem 7.3 in [Devroye and Lugosi, 1996]). Therefore we applied four common algorithms for text classification [Aggarwal and Zhai, 2012]: Naive Bayes (NB), Random Forests (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MP). We compare the performance of these classification algorithms to the majority class baseline (MC). The classifiers were configured with the default parameters given in Weka 3.6.10, except for MP, where the number of hidden layers was 20.

6.4.1 Features

For both experiments, we used the same features as for the feature selection experiment: Each of the 360 DMs in our resources corresponds to one binary feature, which is 1 if the DM can be found in the classification instance, and 0 if not.

6.4.2 Experiment 1: Sentence-level Classification

In this experiment, we classified sentences according to whether they contain a claim, a premise or no AU.

One challenge is the handling of sentence that intersect with multiple AUs – we evaluated the following three solutions: (i) We could ignore multi-AU sentence, or (ii) we could create multiple classification instances from them, each with the same features, but different labels, or

(iii) we could accept the AU with the longest intersection. The second strategy will probably confuse the classifiers, especially the SVM that expects linearly separable classes.

Another question was whether the features of a single sentence would be distinctive enough, or whether we should include context features of the previous and following sentences.

Of course, these are not the only parameter dimensions of the classification task. A fundamental question is whether it is a good idea to classify on sentence level. Probably, it would make more sense to use tokens or clauses (identified by a parser) as classification instances. But for now, we stick to our sentence-level problem definition.

To evaluate the two parameter dimensions – how to handle multi-AU sentences and whether to use context features –, we ran 6 sentence-level experiments in total, which are outlined in Table 6.4.

Experiment	Description
Exp. 1:C-Ign	with context features, ignore multi-AU sentences
Exp. 1:C-Max	with context features, classification instance for AU with max. overlap
Exp. 1:C-Mul	with context features, classification instance for each intersecting AU
Exp. 1:NoC-Ign	no context features, ignore multi-AU sentences
Exp. 1:NoC-Max	no context features, classification instance for AU with max. overlap
Exp. 1:NoC-Mul	no context features, classification instance for each intersecting AU

Table 6.4.: Overview of sentence-level classification experiments (Experiment 1).

Results

Table 6.5 shows the results of Experiment 1. The results are rather negative: When using context information, no classifier performs better than the majority-class baseline. While the SVM is at least able to reproduce the baseline results, all other classifiers perform even worse. The situation changes slightly when we ignore context features: Naive Bayes beats the baseline by 1.98 pp, 1.47 pp, and 0.26 pp, but the improvement is not significant. Again, the accuracy of SVM and MC baseline are equal. RF and MP consistently perform worse than the baseline.

There are several possible reasons for these unsatisfactory results: First, we ran the classifiers without feature selection, so overfitting might be a problem. Second, context features may be problematic because we have no information about the context of each classification instance. Since claims are on average 1.1 sentences and premises 2.2 sentences long, the labels of the surrounding sentences are unpredictable. It is questionable whether a sentence-level classification task is appropriate at all: Perhaps, a sequence tagging classifier on tokens or clauses may produce better results.

6.4.3 Experiment 2: Argument Unit Classification

Since the negative results from Experiment 1 may partly result from misleading context features, Experiment 2 uses AUs as classification instances that are classified either as claim or as premise. For Experiment 2, we excluded context features of surrounding AUs.

Experiment	MC[%]	NB[%]	RF[%]	SVM[%]	MP[%]
Exp1:C-Ign	**46.90**	46.06	42.74	**46.90**	41.30
Exp1:C-Max	**48.09**	46.70	42.38	**48.09**	45.07
Exp1:C-Mul	**47.44**	45.48	39.80	**47.44**	40.70
Exp1:NoC-Ign	46.90	**48.88**	44.51	46.90	40.45
Exp1:NoC-Max	48.09	**49.56**	45.43	48.09	45.93
Exp1:NoC-Mul	47.44	**47.70**	41.47	47.44	40.27

Table 6.5.: Results of Experiment 1: Average accuracy scores for Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MP) in comparison with the majority class (MC) baseline. The **best-performing classifier** per experiment is marked in bold.

Anno.	MC[%]	NB[%]	RF[%]	SVM[%]	MP[%]
GW	53.39	65.04*	63.75*	**67.17***	62.86*
JEK	52.61	**64.91***	64.48*	63.52*	58.85*
RK	52.16	**64.99***	66.17*	64.87*	65.64*

Table 6.6.: Results of Experiment 2: Accuracy scores for Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MP), with the majority class (MC) baseline. x^* means a significant difference from baseline ($\alpha = 5\,\%$). Scores marked in bold denote the **best-performing algorithm** for the dataset of the row.

Results

Table 6.6 summarizes the results of Experiment 2. All classifiers are able to significantly improve upon the baseline. The best improvements of 13.78 pp, 12.30 pp, and 12.83 pp (per dataset) are remarkable, in view of the fact that DMs were the only features in this experiment.

6.5 Discussion

From the three experiments, we learned that DMs are valuable features for discriminating claims from premises. Our results support previous findings in linguistic research on the role of DMs in argumentative discourse [Grote et al., 1997]. We found that certain DMs are significant either for claims or for premises. Interestingly, DMs signaling *result* tend to indicate claims, whereas DMs that express *alternative*, *reason*, or *sequence* indicate premises. This is in accordance (i) to the ability of a claim to act as conclusion or result of an argument, and (ii) to the role of premises as providing support for a claim. Our experiments also show that particular intensifying discourse particles play an important role in discriminating claims and premises: Downtoners seem to be significant for premises, amplifiers for claims.

Discourse particles that either restrict – e.g., nur („*only*“) – or intensify – e.g., ganz („*quite*“) – turned out to be significant as well: Restricting particles rather indicate premises, while intensifying particles indicate premises.

With respect to the classification experiments, our results are ambivalent: In our experiments, DMs were not suitable as exclusive features for classifying sentences into claims, premises, and non-argumentative sentences. However, the second experiment, which classified AUs instead of sentences, showed that DMs can be highly predictive in a proper classification setup.

Qualitative analysis

At least two questions arise from Table 6.3: (i) Why are some of the predictive DMs not significant (and vice versa), and (ii) why are some of the most frequent DMs neither significant nor predictive?

We found that the answer to the first question lies in the nature of the binary features. For instance, the fraction of the DM wie („*how*“) for annotator GW was almost identical for claims and premises, while there are more than twice as much claims than premises containing at least one *wie*; this explains why wie („*how*“) appears to be predictive, but not significant.

To answer the second question, we investigated the DMs that appear at least 40 times in at least one annotator's dataset, 10 of which are listed in Table 6.3. Only the DMs also („*therefore*“) and jedoch („*though*“) listed in Table 6.3 occur less than 40 times. The remaining set of 10 non-significant and non-predictive DMs consists of 5 particles and 5 connectives.

The non-significant and non-predictive, but highly frequent particles are the amplifiers (with median rank) noch („*still*“) (rank 5), immer („*always*“) (rank 11), schon („*already*“) (rank 14), selbst („*even*“) (rank 17), and the downtoner etwas („*some*“) (rank 19). All these intensifiers are highly ambiguous and may carry many different functions within a sentence or within the discourse. For instance, immer („*always*“) and schon („*already*“) can be used in temporal adjuncts, *noch* can be part of the multi-word expression noch einmal („*once more*“), *selbst* can be part of an emphasized reflexive pronoun sich selbst („*oneself*“), and *etwas* often appears as modifier of a nominalization, e.g., etwas Neues („*something new*“).

The most frequent non-significant and non-predictive, but highly frequent connectives fall into two classes of DRs: (i) *D-elaboration* (Auch („*Also*“) (rank 12); Und („*And*“) (rank 13); So („*therefore*“) (rank 28.5)) and *D-sequence* (dann („*then*“)), and (ii) *D-concession* or *D-contrast* (aber („*but*“) (rank 10) and Aber („*But*“) (rank 20)). The former group occurs in claims and premises alike, since they are used in claims to move on to the next argument, whereas in premises they signal a move to the next supporting or attacking premise. Regarding aber/Aber („*but/But*“), the similar frequencies in claims and premises were against our expectation drawn from the literature. We performed a qualitative analysis of the annotated sentences and found that *aber/Aber* signals *concession* in premises as well, as the following example of a premise illustrates:

> **Example 6.2: Fine-grained argumentation structures (*sitzenbleiben11*)**
> **German:**
>
> (1) Es gibt kaum ein wissenschaftliches Indiz dafür, dass das Wiederholen den Schülern etwas bringt.
> (2) Bestenfalls zeigen sie in dem Jahr, das sie wiederholen, bessere Leistungen.
> (3) *Aber* im Vergleich zu anderen ähnlich schwachen Schülern, die versetzt wurden, hängen sie in der nächsten Klassenstufe trotzdem zurück. [. . .]

> (4) Außerdem ist das Sitzenbleiben peinlich und demütigend.
>
> **English:**
> (1) There is few scientific evidence that repeating a class helps students.
> (2) In the best case, they perform better during the year they repeat.
> (3) *But* in subsequent years, they perform worse in comparison with their former classmates with similar weak grades who proceeded to the next class.
> (4) Furthermore, staying down is embarrassing and abasing.

This example demonstrates that it is possible to identify embedded arguments (claims and premises) within a premise itself: Within the premise, (1) – (4) form an argument consisting of the claim (1) and the premises (2) – (4). Our corpus does not cover these embedded arguments, since our pre-study showed that much more time is required to annotate arguments at such a fine-grained level.

7 Conclusion

In this chapter, we summarize our work and give an overview of potential next steps.

7.1 Summary

The field of Argumentation Mining lacks publicly available annotated corpora in German. In this thesis, we describe our efforts in building a new German corpus annotated with argumentative structures according to the claim-premise argumentation model. Contrary to graph-based argumentation schemes, our model represents relations as attributes of the argument units, which simplified the annotation task.

Our annotation study consisted of a pre-study and a main study. During the pre-study, we created detailed annotation guidelines and developed an annotation tool. During the main study, the annotators produced ca. 5,000 argument units in 79 documents, consisting of ca. 70,000 tokens.

After the annotation study, we compiled a number of corpus statistics, which showed that the annotators performed similarly with respect to argument unit length and label distributions. We put special emphasis on the inter-annotator agreement analysis. We applied six inter-annotator agreement metrics, including several implementations of Fleiss' κ (token-based/sentence-based/overlap-based) and Krippendorff's unitized alpha.

A series of three experiments evaluates the role of discourse markers in telling apart claims from premises: A significance analysis and a feature selection experiment showed that particular discourse particles, discourse connectives, and groups of discourse markers are indicative of either claims or premises. In a classification experiment we learned that a sentence-level classification into claim, premise, and non-argumentative sentences was problematic. In contrast, if argument units are used as classification instance, DMs proved to be highly predictive features for distinguishing claims and premises.

7.2 Future Work

This section describes potential subsequent steps.

7.2.1 Additional Corpus Analysis

While we extensively analyzed the corpus in this work, there is still more interesting information to extract from it. We observed that the inter-annotator agreement scores show a wide distribution (especially α_u). Documents with particularly high and low agreement scores should be analyzed for other particularities (e.g., concerning confidence, structure), particularly in view of the pending gold standard creation. Furthermore, the annotators' notes have not been thoroughly analyzed, yet.

7.2.2 Publishing the Corpus

Two theses – this work and Vovk's Master's thesis – build on the described dataset of Web documents. We invested considerable work in cleaning and organizing the data, and in implementing appropriate tools (such as the annotation tool and the UIMA readers). Therefore, it seems reasonable to make the corpus and the tools available to the research community.

We are already in contact with the responsible persons and could publish 55 of the original 89 documents. A license fee would be due for another 19 documents. Presently, we have not received responses for 14 documents. At least 20 documents from *welt.de*, *spiegel.de*, and *zeit.de* could be made accessible by means of customized download scripts.

7.2.3 Refining the Annotation Scheme

We recognized that there is no straight-forward way to establish a gold standard from our annotations. We believe that this partly results from our simplified annotation scheme: By annotating two concepts – the AUs and their relations – in one step, we missed the opportunity to consolidate annotation boundaries in between.

A future annotation scheme could build on the annotated AUs and refine them. The necessary data can be generated by generalizing the annotations to the level of claims and premises (`NO-R`) and then exporting them using the *JsonAnnotatedCorpusWriter* component.

Despite the problems with creating the gold standard, we found that the combined annotation scheme was convenient to annotate. Also, we suggest to keep the annotation of claims and premises in separate phases (rounds 2 and 3 in the per-document annotation process) as this allows the annotator to first identify the argumentation line..

7.2.4 Building a Gold Standard

Even though the focus of this task was not on Argument Extraction, it would still be desirable to create a gold standard sooner or later. Annotations by three annotators allow to use majority voting on several levels of generalization. An automatic procedure would be favorable in light of the ca. 5,000 annotated AUs.

An obvious approach would be to work on token level: We could represent the AUs in the same IOB scheme that has been used for calculating inter-annotator agreement (see Section 5.1) and perform a majority vote per token. Tie breaking strategies need to be developed if all three annotators have assigned different labels.

A similar algorithm is also apt for a sentence-level majority voting. Here, an additional degree of freedom is to decide how to treat sentences with multiple intersecting AUs. In every case, a sentence-level abstraction eradicates information on the exact boundaries of the AUs.

The problem with these two strategies is that they may yield invalid AU patterns. For instance, a pre-claim support that is directly followed by a post-claim support is not valid because of the missing intermediate claim. The token-level majority voting produces 14 invalid arguments, which is moderate compared to the ca. 2300 arguments per annotator.

Therefore, a third option would be to do a majority voting on the arguments (`ARG`). While this preserves the argument boundaries, we can expect that only few arguments endure this procedure, since the exact match criterion is strict and only a single differing token causes two arguments to count as different.

For instance, our preliminary experiments showed that only 383 of the 2,349 arguments (ca. 16 %) would remain after a selection by majority voting. Analogously, 791 of the 5,123 AUs (ca. 15 %) would remain, if we applied majority voting on the original AUs. Clearly, such high losses are not acceptable.

To alleviate this issue, we may reduce the AUs to their maximum overlap beforehand (see Section 5.7) and then derive the arguments. Or we could introduce another matching criterion for the majority voting that considers partial matches.

7.2.5 Argument Extraction

To build the system that we envisioned in the introduction, it is essential to find methods that generalize from the manually annotated data and extract arguments automatically. The moderate inter-annotator agreement shows that even *manually* identifying arguments is challenging and will certainly need more investigation.

Besides classical Machine Learning algorithms such as Support Vector Machine or Random Forests, a plethora of state-of-the-art algorithms for discourse analysis exists and could be exploited for Argument Extraction. Promising examples are deep learning, graph-based methods, or joint inference (e.g., [Collobert and Weston, 2008; Somasundaran et al., 2009; Rahman and Ng, 2009]).

7.3 Acknowledgments

Now it's time to switch to the first person. :-)

I am thankful to my supervisors Dr. Judith Eckle-Kohler and Prof. Dr. Iryna Gurevych for providing me with the opportunity to work on this interesting and challenging project.

Especially, I would like to thank Judith for her steady support and for her intensive involvement in the experiments chapter. With her undisputed expertise and engagement, she helped me to find and keep the right direction throughout the project. I really enjoyed working closely with her and regret that we cannot extend this cooperation into the future.

Also, I would like to thank the other members of the special interest group on Argumentation Mining – Dr. Ivan Habernal, Christian Kirschner, Krish Perumal, Piyush Paliwal, and Christian Stab – for the vivid and insightful discussions.

By providing his annotation tool, the pre-selected documents, and the insights in this thesis, Artem Vovk considerably supported this work.

A Implementation Notes

This chapter presents implementation notes such as the storage format of the corpus.

A.1 Important Classes

The following lists contain important components that may be worth reusing in the future. I also want to shortly acknowledge at this place the user-friendly open-source bibliography management tool JabRef, which I used during my work [JabRef Development Team, 2013].

A.1.1 Data Model

UIMA types

`Annotator` is a document-level annotation that marks an annotator of a document.

`ArgumentUnit` is a span annotation, that represents an AU and has a *label*, *confidence*, and *annotator* property.

`DocumentMetadata` contains document-level information such as the source URL and the file-name/document ID.

`SentenceLevelArgUnit` is used to represent aggregated AUs on a sentence-level, e.g., for sentence-level IAA analysis or gold standard creation.

Non-UIMA classes

`ArgumentUnitLabel` contains the possible AU labels for all generalization levels.

`ArgumentUnits` provides several methods for selecting and filtering `ArgumentUnits` in a JCas.

`GeneralizationLevel` models the generalization levels.

`Topic` models the topics in the corpus.

A.1.2 Corpus Input and Output

`HtmlCorpusReader` reads the manually preprocessed corpus files as one HTML file per document.

`JsonAnnotatedCorpusReader` reads the annotated corpus from JSON (see Appendix A.2).

`JsonAnnotatedCorpusWriter` writes the corpus to JSON (see Appendix A.2).

`JsonCorpusUtil` contains constants and utility methods concerning the corpus.

`JsonPreprocessedCorpusWriter` writes the automatically pre-processed documents to JSON, HTML, and plain text.

A.1.3 Corpus Processing

`ArgumentAnnotator` annotates whole arguments.

`ConfidenceFilteringAnnotator` filters the AUs in a JCas according to confidence.

`HeadingAnnotator` annotates the DKPro type `Heading` and corrects the sentence segmentation.

`LabelGeneralizingAnnotator` generalizes the labels of AUs for the generalization levels `PR`, `NO-R`, and `AU`.

`OriginalArgumentUnitsRestorer` restores the original AUs, reverting `ArgumentAnnotator` and `LabelGeneralizingAnnotator`.

`ParagraphAnnotator` annotates the DKPro type `Paragraphs`.

A.1.4 Corpus Evaluation

We developed a number of components for corpus analysis. The class `CorpusEvaluationApplication` and the contained `*Runner` are good entry points to explore the corpus analysis components.

Inter-annotator agreement

`JaccardAgreementCalculator` evaluates the Jaccard-based IAA metric on a document.

`SentenceLevelStudyFactory` produces an `AnnotationStudy` instance with sentence-level annotation items.

`TokenLevelStudyFactory` produces an `AnnotationStudy` instance with token-level annotation items.

`WilsonWiebeAgreementCalculator` evaluates Wilson-Wiebe's IAA metric on a document.

`MaximumOverlapNormalizer` implements the Maximum Overlap Normalization algorithm.

A.2 Corpus Format

The corpus is stored in a single JSON file. The following listing contains the JSON schema definition (according to JSON Schema draft (version 4)[1]) of the file format. The corpus dumps have been validated against this scheme with the *json-schema-validator* toolkit[2].

```
{
    "type" : "array",
    "items" : {
        "type" : "object",
        "properties" : {
          "corpus_metadata" : {"type": "string"},
```

[1] http://tools.ietf.org/html/draft-zyp-json-schema-04

[2] https://json-schema-validator.herokuapp.com/

```
          "preprocessing_date" : {"type" : "string"},
          "postprocessing_date" : {"type" : "string"},
          "segmenter" : {"type" : "string"},

          "num_sentences" : {"type" : "number"},
          "num_tokens" : {"type" : "number"},
          "url" : {"type" : "string"},
          "file" : {"type" : "string"},
          "text" : {"type" : "string"},
          "user_annotations" : {
            "type" : "array",
            "items" : {
              "type" : "object",
              "properties" : {
                "arg_units" : {
                  "type" : "array",
                  "items" : {"type" : "string"}
                },
                "notes" : {"type" : "string"},
                "annotator" : {"type" : "string"},
                "approved" : {"type" : "string"}
              }
            }
          }
        }
    }
}
```

Corpus metadata

The file consists of an array of JSON objects. The first top-level object contains metadata with properties *corpus_metadata*, *preprocessing_date*, *segmenter*:

- `copus_metadata`: The property is either `"true"` or `"false"`. If it is true, the current object is treated as corpus metadata; if it is false, the current object is treated as document metadata.

- `segmenter`: The fully qualified class name of the segmenter that was used to segment the document, e.g., `de.tudarmstadt.ukp.dkpro.core.languagetool.LanguageToolSegmenter`.

- `preprocessing_date,postprocessing_date`: Indicates the date and time, when the corpus was pre- and post-processed.

Document metadata

All top-level objects but the first describe a document. The following list describes the metadata entries per document:

- `num_sentences, num_tokens`: The number of tokens and sentences. Note that numbers are stored without surrounding quotes, e.g., `"num_sentences":74`.

- `url`: The source URL of the document, e.g., `"url":"http://www.spiegel.de"`.

- `file`: The filename in the corpus, e.g., `"file":"g80.json"`.

- `text`: The pre-rendered HTML text as displayed in the annotation tool. The text contains span tags for every token as well as paragraph and heading tags. The annotation tool expects a token to look as follows. For technical reasons, whitespaces between tokens are marked with the *gap* class:

  ```
  <span id="t1" class="token" idx="1">Turbo-Abiturienten</span>
  <span class="gap"> </span>
  <span id="t2" class="token" idx="2">sind</span>
  ```

- `user_annotations`: A JSON array with one entry per annotator, holding the annotator's annotations (see below).

User annotations

The `user_annotations` property of a document is a list of objects, each representing one annotator's annotations:

- `annotator`: Name of the annotator. The annotation tool uses the registered mail address, e.g., `"annotator" : "xy@gmail.com"`.

- `notes`: The annotator's notes for this document. Notes that allude to a particular paragraph look like this: `"notes":"p13: Indirekte Evidenz"`

- `approved`: A flag that indicates whether the annotator has permanently approved his annotations in this document, e.g., `"approved":"True"`.

- `arg_units`: A JSON array of strings. Each string represents a single annotation in form of a serialized JavaScript list:

  ```
  ["<label>", "<confidence>", <firstTokenIndex>, <lastTokenIndex>]
  ```

  An example with two annotations follows:

  ```
  "arg_units":[
    "[\"support-pre\",\"medium\",225,253]",
    "[\"claim\",\"high\",255,271]"
  ]
  ```

B Document Statistics

Table B.1.: Per-document statistics. **#P/#S/#T:** The number of paragraphs, sentences, and tokens in the document. **Date:** The date is in the format *dd.mm.yy* or *yyyy* if only the year is known. **Notes:** *more pages* means that missing pages were added

File	#P	#S	#T	Date	Notes
g80.json	17	78	1149	29.08.11	
g810.json	12	43	760	2013	
g81.json	12	29	845	15.05.13	
g82.json	11	74	1344	27.11.12	
g83.json	14	40	906	05.09.12	
g84.json	12	40	716	07.01.11	
g85.json	8	29	509	20.08.10	
g86.json	10	38	672	10.10.12	
g87.json	8	42	816	10.10.12	more pages
g88.json	13	42	992	15.11.12	
g89.json	14	38	676	21.08.11	
g811.json	19	108	1764	10.09.12	prev.: g98.json
inklusion0.json	12	34	668	09.10.12	more pages
inklusion1.json	19	78	1351	18.03.13	
inklusion2.json	4	18	457	09.04.13	
inklusion3.json	45	216	4307	31.05.10	
inklusion4.json	26	103	1818	05.02.12	
inklusion5.json	16	37	693	01.05.12	
inklusion6.json	15	64	1214	19.08.11	
inklusion7.json	9	61	959	01.12.11	
lehrer0.json	18	42	872	27.03.13	
lehrer10.json	12	68	1126	28.02.09	
lehrer11.json	19	78	1429	12.06.07	
lehrer12.json	13	54	1002	13.11.12	
lehrer13.json	16	44	1031	26.03.09	
lehrer14.json	15	46	887	10.04.03	
lehrer15.json	9	24	571	21.03.05	
lehrer16.json	8	27	485	30.11.09	
lehrer17.json	11	47	921	16.01.13	
lehrer18.json	21	69	1044	07.09.12	
lehrer19.json	8	24	464	11.09.09	
lehrer1.json	4	19	296	09.03.13	
lehrer2.json	11	45	756	26.03.13	
lehrer3.json	10	34	712	28.10.02	
lehrer4.json	17	72	1326	07.08.08	

Continued on next page. . .

Table B.1: General statistics of the corpus documents (continued)

File	#P	#S	#T	Date	Notes
lehrer6.json	10	33	754	08.05.03	
lehrer7.json	27	84	1307	24.04.12	
lehrer8.json	11	38	636	16.06.09	
lehrer9.json	8	28	455	19.06.02	
master0.json	12	45	714	19.06.12	more pages
master1.json	24	56	1044	30.03.10	more pages
master2.json	21	64	1061	09.04.13	more pages
master3.json	11	74	1578	05.02.12	
master4.json	14	94	1888	12.11.12	
master5.json	8	39	673	01.09.12	
master6.json	13	35	825	08.10.10	
promovieren0.json	19	83	1449	05.03.11	more pages
promovieren10.json	17	70	1532	23.04.13	
promovieren11.json	8	46	780	23.06.03	
promovieren12.json	10	34	582	15.02.07	
promovieren1.json	10	63	1108	03.04.13	more pages
promovieren2.json	17	62	1060	30.08.12	
promovieren3.json	7	53	875	2001	
promovieren4.json	18	62	895	07.09.12	
promovieren5.json	12	54	834	26.07.12	
promovieren6.json	22	76	1562	01.11.13	
promovieren7.json	20	86	1187	08.02.13	
promovieren8.json	12	52	975	09.08.09	
promovieren9.json	8	44	703	29.11.10	
sitzenbleiben0.json	17	48	1083	08.07.05	
sitzenbleiben10.json	10	50	881	04.02.13	
sitzenbleiben11.json	8	29	502	20.02.13	
sitzenbleiben12.json	13	38	631	19.02.13	
sitzenbleiben13.json	18	61	1196	15.04.08	
sitzenbleiben14.json	9	27	668	16.05.13	
sitzenbleiben15.json	-	-	-	18.03.13	dropped
sitzenbleiben16.json	16	67	1083	17.02.13	
sitzenbleiben17.json	6	23	440	05.05.13	
sitzenbleiben18.json	5	16	232	12.03.13	
sitzenbleiben19.json	23	98	1951	03.07.11	more pages
sitzenbleiben1.json	31	73	1106	05.04.13	
sitzenbleiben20.json	12	48	866	16.02.13	
sitzenbleiben21.json	14	45	639	16.03.13	
sitzenbleiben2.json	28	84	1810	08.02.02	
sitzenbleiben3.json	8	30	597	06.03.13	
sitzenbleiben4.json	6	22	476	04.09.09	
sitzenbleiben5.json	13	41	906	03.09.09	
sitzenbleiben6.json	23	62	1036	06.03.13	

Continued on next page...

Table B.1: General statistics of the corpus documents (continued)

File	#P	#S	#T	Date	Notes
sitzenbleiben7.json	7	19	373	06.09.04	
sitzenbleiben8.json	10	40	659	17.01.12	
sitzenbleiben9.json	20	55	1053	06.05.08	
sport0.json	8	29	628	09.04.13	
sport1.json	8	33	447	18.05.10	
sport2.json	19	34	723	10.04.13	
sport3.json	7	22	391	30.07.13	
sport4.json	3	29	412	12.04.13	
sport5.json	7	28	516	30.07.13	
sport6.json	3	21	310	05.04.13	
sport7.json	4	22	301	19.06.12	

C Document Source URLs

Table C.1.: List of document sources URLs

File	URL
g80.json	http://www.spiegel.de/schulspiegel/turbo-abiturienten-nach-klasse-12-wir-versuchskaninchen-a-781215.html
g810.json	https://www.change.org/de/Petitionen/wiedereinf\%C3\%BChrung-des-g9-an-hamburger-gymnasien-mit-wahlfreiheit-zwischen-g8-und-g9
g81.json	http://www.spiegel.de/schulspiegel/wissen/g9-jetzt-hamburger-eltern-starten-ini-fuer-neunjaehriges-gymnasium-a-900009.html
g82.json	http://www.spiegel.de/schulspiegel/wissen/kess-studie-zu-g8-und-g9-acht-jahre-gymnasium-reichen-aus-a-869483.html
g83.json	http://www.spiegel.de/schulspiegel/wissen/g8-eltern-lehnen-turbo-abitur-ab-a-854096.html
g84.json	http://www.spiegel.de/schulspiegel/wissen/schulen-in-nrw-einige-gymnasien-wollen-turbo-abi-kippen-a-738375.html
g85.json	http://www.spiegel.de/schulspiegel/wissen/hessen-schueler-klagt-gegen-ungerechtigkeit-bei-turbo-abi-a-712926.html
g86.json	http://www.sueddeutsche.de/bildung/debatte-um-gymnasialreform-mehr-zeit-weniger-stress-1.1301724
g87.json	http://www.sueddeutsche.de/bildung/folgen-der-verkuerzten-schulzeit-setzen-sechs-1.1400905
g88.json	http://www.welt.de/regionales/frankfurt/article111169857/Wahlfreiheit-bei-G8-G9-traegt-zu-Unruhe-bei.html
g89.json	http://www.welt.de/print/wams/vermischtes/article13556539/Ein-gutes-Abitur-braucht-seine-Zeit.html
g811.json	http://www.welt.de/print/die_welt/politik/article109114444/Das-doppelte-Schul-Lottchen.html
inklusion0.json	http://www.sueddeutsche.de/bildung/inklusion-statt-foerderschule-wann-gemeinsames-lernen-sinnvoll-ist-1.1482320
inklusion1.json	http://www.welt.de/politik/deutschland/article114538623/Wie-Eltern-das-Projekt-Inklusion-torpedieren.html

Continued on next page...

Table C.1: Document source URLs (continued)

File	URL
inklusion2.json	`http://www.nsfkn.info/?p=1464`
inklusion3.json	`http://www.bpb.de/apuz/32713/ueber-widersacher-der-inklusion-und-ihre-gegenreden-essay?p=all`
inklusion4.json	`http://www.welt.de/politik/deutschland/article13850739/Funktioniert-die-Schule-mit-der-vollen-Inklusion.html`
inklusion5.json	`http://www.myhandicap.de/behinderte-kinder-schule-inklusiv.html`
inklusion6.json	`http://www.spiegel.de/schulspiegel/wissen/behinderte-schueler-na-bitte-es-geht-doch-a-769821.html`
inklusion7.json	`http://www.focus.de/schule/schule/unterricht/inklusion/inklusion-eine-schule-fuer-alle\_aid\_684442.html`
lehrer0.json	`http://www.spiegel.de/schulspiegel/wissen/beamtenstatus-und-gehalt-ob-es-sich-lohnt-lehrer-zu-werden-a-877467-druck.html`
lehrer10.json	`http://www.spiegel.de/schulspiegel/arme-lehrer-notfalls-gehe-ich-putzen-a-608995.html`
lehrer11.json	`http://www.spiegel.de/schulspiegel/leben/cyber-mobbing-gegen-lehrer-pornomontagen-und-hinrichtungsvideos-a-488062-druck.html`
lehrer12.json	`http://www.spiegel.de/schulspiegel/wissen/studie-jeder-sechste-lehrer-fuehlt-sich-gemobbt-a-866808.html`
lehrer13.json	`http://www.spiegel.de/schulspiegel/wissen/imageproblem-mehrheit-der-deutschen-haelt-lehrer-fuer-ueberfordert-und-unfaehig-a-615636.html`
lehrer14.json	`http://www.spiegel.de/schulspiegel/stress-im-klassenzimmer-jeder-dritte-lehrer-ist-ausgebrannt-a-244095.html`
lehrer15.json	`http://www.spiegel.de/schulspiegel/deutschlands-lehrer-raus-aus-der-schmollecke-a-347012.html`
lehrer16.json	`http://www.focus.de/schule/lernen/bildung-praemien-und-boni-fuer-gute-paedagogen\_aid\_459248.html`
lehrer17.json	`http://www.spiegel.de/schulspiegel/lehrer-lehnt-verbeamtung-ab-und-moechte-als-angestellter-arbeiten-a-877431-druck.html`
lehrer18.json	`http://www.focus.de/schule/schule/9000-lehrer-gehen-fuer-die-bildung-auf-die-strasse-lehrer-warnstreik-in-sachsen\_aid\_814848.html`
lehrer19.json	`http://www.focus.de/schule/lehrerzimmer/schulpraxis/angst-wenn-schule-lehrer-krank-macht\_aid\_434812.html`
lehrer1.json	`http://www.welt.de/print/die\_welt/finanzen/article114280562/Einige-Privilegien-wenig-Dank.html`

Continued on next page...

Table C.1: Document source URLs (continued)

File	URL
lehrer2.json	http://www.spiegel.de/schulspiegel/wissen/faktencheck-wie-viel-arbeiten-lehrer-und-wie-viel-freizeit-haben-sie-a-874089-druck.html
lehrer3.json	http://www.spiegel.de/schulspiegel/lehrer-arbeitszeit-keine-fleisskaertchen-fuer-paedagogen-a-219833-druck.html
lehrer4.json	http://www.spiegel.de/unispiegel/jobundberuf/20-000-freie-stellen-deutschland-gehen-die-lehrer-aus-a-570627.html
lehrer6.json	http://www.spiegel.de/schulspiegel/lehrer-als-schulschwaenzer-protest-paedagogen-muessen-attest-vorlegen-a-247821.html
lehrer7.json	http://www.welt.de/politik/deutschland/article106221096/Disziplinlose-Schueler-ueberfordern-deutsche-Lehrer.html
lehrer8.json	http://www.spiegel.de/schulspiegel/wissen/lehrer-studie-weltweite-klagen-ueber-ruepel-schueler-a-630741-druck.html
lehrer9.json	http://www.spiegel.de/schulspiegel/lachseminare-fuer-lehrer-lernziel-witzischkeit-a-193916-druck.html
master0.json	http://www.sueddeutsche.de/bildung/nach-dem-bachelor-warum-der-master-keine-entscheidenden-vorteile-bringt-1.1414318
master1.json	http://www.zeit.de/campus/2010/s1/was-studieren-interview
master2.json	http://www.zeit.de/campus/2013/s2/master-studium-entscheidung
master3.json	http://jetzt.sueddeutsche.de/texte/anzeigen/538752/Muss-ich-wirklich-noch-den-Master-machen
master4.json	http://abi.de/studium/studiengaenge/weiterfuehrende/master09530.htm?zg=schueler
master5.json	http://www.fr-online.de/berufsrundschau/abschluss-der-master-ist-kein-muss,4599958,17009424.html
master6.json	http://www.ingenieur.de/Arbeit-Beruf/Ausbildung-Studium/Der-Master-lohnt-fuer-erfahrene-Kollegen
promovieren0.json	http://www.zeit.de/2011/10/Ueberfluessige-Dissertationen
promovieren10.json	http://www.ksta.de/job-und-karriere/promotion-mehr-geld-mit-doktortitel,20063080,22559908.html
promovieren11.json	http://www.spiegel.de/unispiegel/jobundberuf/promotion-was-tun-dr-arbeitslos-a-252315.html

Continued on next page...

Table C.1: Document source URLs (continued)

File	URL
promovieren12.json	http://www.faz.net/aktuell/beruf-chance/campus/karriere-persoenlichkeit-statt-promotion-1407397.html
promovieren1.json	http://www.zeit.de/studium/hochschule/2013-04/promotionen-anstieg-studentenzahlen
promovieren2.json	http://www.spiegel.de/unispiegel/studium/promovieren-doktortitel-kann-die-jobsuche-erschweren-a-843999.html
promovieren3.json	http://www.haus-der-sprache.de/lektor.php/redaktion/lesen-karriere/promovieren_oder_nicht_was_bringt_der_doktorhut/
promovieren4.json	http://www.bildung-news.com/bildung-und-karriere/erfahrungsberichte/10-gute-grunde-nicht-zu-promovieren/
promovieren5.json	http://www.bildung-news.com/bildung-und-karriere/erfahrungsberichte/10-grunde-fur-eine-promotion/
promovieren6.json	http://www.jobvector.de/journal/bewerbung/soll_ich_promovieren/index_ger.html
promovieren7.json	http://www.christophburger.de/?p=1180
promovieren8.json	http://www.welt.de/welt_print/wissen/article4285459/Lohnt-sich-eine-Doktorarbeit.html
promovieren9.json	http://www.sueddeutsche.de/karriere/phd-statt-promotion-auswandern-fuer-den-doktortitel-1.1029549
sitzenbleiben0.json	http://www.spiegel.de/schulspiegel/sitzenbleiben-nichts-als-verplemperte-zeit-a-364198-druck.html
sitzenbleiben10.json	http://www.sueddeutsche.de/bildung/niedersachsen-will-sitzenbleiben-abschaffen-aus-fuer-die-unruehmliche-ehrenrunde-1.1591350
sitzenbleiben11.json	http://www.spiegel.de/schulspiegel/wissen/bildungsforscher-sitzenbleiben-bringt-schuelern-kaum-vorteile-a-884286-druck.html
sitzenbleiben12.json	http://www.welt.de/geschichte/article113734891/Viele-Sitzenbleiber-machten-doch-noch-Karriere.html
sitzenbleiben13.json	http://www.focus.de/schule/schule/recht/schule-sitzenbleiben-wird-abgeschafft_aid_295316.html
sitzenbleiben14.json	http://www.spiegel.de/schulspiegel/leben/schueler-berichten-ueber-sitzenbleiben-a-899607.html
sitzenbleiben15.json	http://www.erstenachhilfe.de/blog/Sitzenbleiben-abschaffen-Schueler-und-Studenten-sagen-Nein
sitzenbleiben16.json	http://www.badische-zeitung.de/suedwest-1/auch-baden-wuerttemberg-ist-gegen-das-sitzenbleiben--69201353.html
sitzenbleiben17.json	http://daserste.ndr.de/guentherjauch/rueckblick/schulreform439.html

Continued on next page...

Table C.1: Document source URLs (continued)

File	URL
sitzenbleiben18.json	http://www.faz.net/aktuell/wissen/faktencheck/faktencheck-hilft-das-sitzenbleiben-in-der-schule-12111532.html
sitzenbleiben19.json	http://www.zeit.de/2011/27/C-Interview-Prenzel
sitzenbleiben1.json	http://www.dw.de/streit-ums-sitzenbleiben/a-16692803
sitzenbleiben20.json	http://www.sueddeutsche.de/bildung/bildungssenator-in-hamburg-sitzenbleiben-nuetzt-nichts-und-verschwendet-viel-geld-1.1601356
sitzenbleiben21.json	http://www.heise.de/tp/artikel/38/38752/1.html
sitzenbleiben2.json	http://www.spiegel.de/schulspiegel/ehrenrunden-debatte-deutsche-sind-fuer-das-sitzenbleiben-in-der-schule-a-181217.html
sitzenbleiben3.json	http://www.spiegel.de/schulspiegel/laut-umfrage-halten-deutsche-schueler-das-sitzenbleiben-fuer-richtig-a-887150-druck.html
sitzenbleiben4.json	http://www.tagesspiegel.de/wissen/bildungsforschung-foerdern-statt-frustrieren/1593774.html
sitzenbleiben5.json	http://www.spiegel.de/schulspiegel/wissen/neue-bildungsstudie-sitzenbleiben-ist-nutzlos-und-teuer-a-646709.html
sitzenbleiben6.json	http://www.welt.de/politik/deutschland/article114159103/Deutsche-Schueler-wollen-das-Sitzenbleiben-retten.html?config=print
sitzenbleiben7.json	http://www.spiegel.de/schulspiegel/ehrenrunde-sitzenbleiber-bringen-bessere-leistungen-a-316824-druck.html
sitzenbleiben8.json	http://www.focus.de/schule/schule/bildungspolitik/tid-24802/abschied-auf-raten-sitzenbleiben-kommt-aus-der-mode_aid_684417.html
sitzenbleiben9.json	http://www.spiegel.de/schulspiegel/wissen/teuer-sinnlos-frustrierend-weg-mit-der-ehrenrunde-a-551743.html
sport0.json	http://www.tagesspiegel.de/meinung/jungen-und-maedchen-im-sportunterricht-getrennt-turnt-es-sich-besser/8035878.html
sport1.json	http://www.derwesten.de/zeusmedienwelten/zeus/fuer-schueler/zeus-regional/gladbeck/geschlechtertrennung-im-sportunterricht-id3518467.html
sport2.json	http://blog.initiativgruppe.de/2013/04/10/sportunterricht-und-integration-jungen-und-madchen-zusammen-oder-getrennt/

Continued on next page...

Table C.1: Document source URLs (continued)

File	URL
sport3.json	`http://www.rbb-online.de/politik/beitrag/2013/07/maedchen\_und\_jungen\_duerfen\_getrennt\_unterrichtet\_werden.html`
sport4.json	`http://www.neues-deutschland.de/artikel/818437.streitfall-getrennter-sportunterricht.html`
sport5.json	`http://www.berliner-zeitung.de/berlin/getrennter-sportunterricht-maedchen-muessen-auf-den-schwebebalken,10809148,23868898.html`
sport6.json	`http://www.ismail-tipi.de/inhalte/2/aktuelles/35355/getrennter-sportunterricht-schadet-der-integration-und-ist-eine-steilvorlage-fuer-extremisten/index.html`
sport7.json	`http://www.derwesten.de/staedte/luenen/jugend/getrennt-statt-im-team-im-sport-ein-volltreffer-id6785196.html`

D Argumentation Patterns

D.1 All Argumentation Patterns

Table D.1 lists all argumentation patterns in the corpus.

Table D.1.: List of all argumentation patterns. Legend: *C* – claim, *C-Re* – restatement, *S-Pr/Po* – pre-/post-claim support, *A-Pr/Po* – pre-/post-claim attack

Rank	Pattern	Frequency	Fraction
1	C,S-Po	1398	59.51 %
2	S-Pr,C	272	11.58 %
3	S-Pr,C,S-Po	141	6.00 %
4	C	130	5.53 %
5	C,S-Po,A-Po	66	2.81 %
6	A-Pr,C,S-Po	51	2.17 %
7	C,A-Po	49	2.09 %
8	C,S-Po,C-Re	41	1.75 %
9	C,A-Po,S-Po	36	1.53 %
10	A-Pr,C	31	1.32 %
11	C,S-Po,S-Po	23	0.98 %
12	S-Pr,C,A-Po	14	0.60 %
13	A-Pr,S-Pr,C	12	0.51 %
14	C,S-Po,C-Re,S-Po	11	0.47 %
15	S-Pr,C,S-Po,A-Po	9	0.38 %
16	C,S-Po,A-Po,S-Po	8	0.34 %
17	A-Pr,C,S-Po,A-Po	3	0.13 %
18	S-Pr,S-Pr,C	3	0.13 %
19	C,A-Po,S-Po,A-Po	3	0.13 %
20	A-Pr,S-Pr,C,S-Po	3	0.13 %
21	C,A-Po,C-Re	3	0.13 %
22	S-Pr,A-Pr,C	3	0.13 %
23	S-Pr,C,C-Re	3	0.13 %
24	C,A-Po,S-Po,C-Re	2	0.09 %
25	C,S-Po,A-Po,C-Re	2	0.09 %
26	C,S-Po,S-Po,A-Po	2	0.09 %
27	S-Pr,C,S-Po,C-Re	2	0.09 %
28	C,A-Po,S-Po,S-Po	2	0.09 %
29	A-Pr,C,S-Po,C-Re,S-Po	2	0.09 %
30	S-Pr,C,S-Po,C-Re,S-Po	2	0.09 %
31	C,C-Re,S-Po	2	0.09 %
32	S-Pr,C,S-Po,S-Po	1	0.04 %

Continued on next page...

Table D.1: List of all annotation patterns (continued)

Rank	Pattern	Frequency	Fraction
33	C,A-Po,S-Po,S-Po,A-Po	1	0.04 %
34	C,S-Po,A-Po,S-Po,A-Po,S-Po	1	0.04 %
35	S-Pr,C,S-Po,S-Po,S-Po,S-Po	1	0.04 %
36	S-Pr,C,A-Po,S-Po	1	0.04 %
37	A-Pr,C,A-Po	1	0.04 %
38	A-Pr,A-Pr,C,S-Po	1	0.04 %
39	S-Pr,S-Pr,A-Pr,C,S-Po	1	0.04 %
40	C,S-Po,A-Po,A-Po	1	0.04 %
41	S-Pr,A-Pr,S-Pr,C	1	0.04 %
42	C,C-Re	1	0.04 %
43	S-Pr,S-Pr,C,S-Po,S-Po	1	0.04 %
44	A-Pr,C,S-Po,A-Po,S-Po	1	0.04 %
45	C,S-Po,A-Po,A-Po,S-Po,A-Po	1	0.04 %
46	C,A-Po,C-Re,S-Po	1	0.04 %
47	S-Pr,C,C-Re,S-Po	1	0.04 %
48	S-Pr,A-Pr,C,S-Po	1	0.04 %
49	C,A-Po,A-Po,S-Po,A-Po	1	0.04 %
50	C,S-Po,A-Po,S-Po,S-Po,A-Po,S-Po,A-Po,A-Po	1	0.04 %
51	C,S-Po,C-Re,A-Po,S-Po	1	0.04 %

D.2 Argumentation Patterns in Introduction and Conclusion

Table D.2 and Table D.3 list all argumentation patterns in the first and last paragraph.

Table D.2.: List of argumentation patterns in the first paragraph.

Rank	Pattern	Frequency	Percentage
1	C→S-Po	56	31.28 %
2	S-Pr→C	42	23.46 %
3	C	39	21.79 %
4	A-Pr→C	9	5.03 %
5	C→S-Po→A-Po	6	3.35 %
6	A-Pr→C→S-Po	5	2.79 %
7	S-Pr→C→S-Po	5	2.79 %
8	C→A-Po→S-Po	4	2.23 %
9	A-Pr→S-Pr→C	3	1.68 %
10	S-Pr→C→A-Po	3	1.68 %
11	A-Pr→C→S-Po→A-Po→S-Po	1	0.56 %
12	C→S-Po→C-Re→A-Po→S-Po	1	0.56 %
13	C→S-Po→C-Re→S-Po	1	0.56 %
14	A-Pr→C→S-Po→C-Re→S-Po	1	0.56 %
15	C→C-Re	1	0.56 %

Continued on next page…

Table D.2: List of all annotation patterns in the first paragraph (continued)

Rank	Pattern	Frequency	Percentage
16	S-Pr→C→A-Po→S-Po	1	0.56 %
17	C→A-Po	1	0.56 %

Table D.3.: List of argumentation patterns in the last paragraph.

Rank	Pattern	Frequency	Percentage
1	C→S-Po	137	54.15 %
2	C	29	11.46 %
3	S-Pr→C	24	9.49 %
4	S-Pr→C→S-Po	14	5.53 %
5	C→S-Po→C-Re	10	3.95 %
6	C→A-Po	6	2.37 %
7	C→S-Po→A-Po	5	1.98 %
8	C→A-Po→S-Po	4	1.58 %
9	A-Pr→C	4	1.58 %
10	A-Pr→C→S-Po	3	1.19 %
11	A-Pr→S-Pr→C	3	1.19 %
12	C→S-Po→C-Re→S-Po	3	1.19 %
13	C→A-Po→C-Re	3	1.19 %
14	C→S-Po→A-Po→S-Po	1	0.40 %
15	S-Pr→C→A-Po	1	0.40 %
16	S-Pr→S-Pr→A-Pr→C→S-Po	1	0.40 %
17	S-Pr→C→S-Po→C-Re	1	0.40 %
18	S-Pr→C→S-Po→C-Re→S-Po	1	0.40 %
19	C→S-Po→A-Po→C-Re	1	0.40 %
20	S-Pr→A-Pr→C	1	0.40 %
21	A-Pr→S-Pr→C→S-Po	1	0.40 %

E Pairwise Overlap Distributions

This part of the appendix lists the frequency distributions of pairwise overlap.

Overlap	#AUs
0	238 (14.4 %)
1	1182 (71.6 %)
2	160 (9.7 %)
3	49 (3.0 %)
4	8 (0.5 %)
5	5 (0.3 %)
6	4 (0.2 %)
7	2 (0.1 %)
8	1 (0.1 %)
10	1 (0.1 %)

JEK vs. GW

Overlap	#AUs
0	255 (14.3 %)
1	1355 (75.9 %)
2	133 (7.5 %)
3	32 (1.8 %)
4	8 (0.4 %)
5	1 (0.1 %)
8	1 (0.1 %)

GW vs. JEK

Overlap	#AUs
0	220 (13.0 %)
1	1242 (73.4 %)
2	169 (10.0 %)
3	44 (2.6 %)
4	10 (0.6 %)
5	2 (0.1 %)
6	2 (0.1 %)
7	2 (0.1 %)

RK vs. GW

Overlap	#AUs
0	243 (13.6 %)
1	1353 (75.8 %)
2	149 (8.3 %)
3	33 (1.8 %)
4	4 (0.2 %)
5	1 (0.1 %)
8	1 (0.1 %)
9	1 (0.1 %)

GW vs. RK

Overlap	#AUs
0	211 (14.3 %)
1	1301 (87.9 %)
2	140 (9.5 %)
3	30 (2.0 %)
4	6 (0.4 %)
5	2 (0.1 %)
10	1 (0.1 %)

RK vs. JEK

Overlap	#AUs
0	218 (13.2 %)
1	1219 (73.9 %)
2	165 (10.0 %)
3	35 (2.1 %)
4	8 (0.5 %)
5	2 (0.1 %)
6	2 (0.1 %)
7	1 (0.1 %)

JEK vs. RK

Bibliography

Safia Abbas and Hajime Sawamura. Argument mining using highly structured argument repertoire. In *First International Conference on Educational Data Mining (EDM)*, pages 202–210, Montreal, Qubec, Canada, 2008.

Charu C. Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer, 2012.

Shameem Ahmed, Catherine Blake, Kate Williams, Noah Lenstra, and Qiyuan Liu. Identifying Claims in Social Science Literature. In *Proceedings of iConference 2013*, pages 942–946, Fort Worth, TX, USA, 2013. doi: 10.9776/13485.

Amal Al-Saif and Katja Markert. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2046–2053, Valletta, Malta, 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Bal Krishna Bal and Patrick Saint Dizier. Towards Building Annotated Resources for Analyzing Opinions and Argumentation in News Editorials. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Bal Krishna Bal and Patrick Saint-Dizier. Who speaks for whom? Towards analyzing opinions in news editorials. In *Eighth International Symposium on Natural Language Processing, 2009 (SNLP'09)*, pages 227–232, Bangkok, Thailand, 2009.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259, 2010. ISSN 0269-2821. doi: 10.1007/s10462-010-9154-1.

Daniela Berger, David Reitter, and Manfred Stede. XML/XSL in the Dictionary: The Case of Discourse Markers. In *Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2002) held at the 19th International Conference on Computational Linguistics*, pages 1–8, Taipei, Taiwan, 2002. ACL. doi: 10.3115/1118808.1118812.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

Fiona Browne, Yan Jin, Colm Higgins, David Bell, Niall Rooney, Hui Wang, Fergal Monaghan, Zhiwei Lin, Jann Mueller, Alan Sergeant, et al. The Application of a Natural Language Argumentation Based Approach within Project Life Cycle Management. In *Proceedings of 22nd Irish conference on Artificial Intelligence and Cognitive Science*, page 10, Magee, 2011. ISRC.

Katarzyna Budzynska. Araucaria-PL: software for teaching argumentation theory. In *Tools for Teaching Logic*, pages 30–37. Springer, 2011.

Elena Cabrio and Serena Villata. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 208–212. Association for Computational Linguistics, 2012.

Elena Cabrio and Serena Villata. Detecting Bipolar Semantic Relations among Natural Language Arguments with Textual Entailment: a Study. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, 2013.

Elena Cabrio, Sara Tonelli, and Serena Villata. From Discourse Analysis to Argumentation Schemes and Back: Relations and Differences. In *Computational Logic in Multi-Agent Systems*, pages 1–17. Springer, 2013.

Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. *Building a discourse-tagged corpus in the framework of rhetorical structure theory*, chapter 5, pages 85–112. Springer, 2003.

Ronan Collobert and Jason Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390177.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, 6(4): 1–220, 2013. doi: 10.2200/S00509ED1V01Y201305HLT023.

László Györfi Devroye, Luc and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability. Springer, New York, 1996.

Vanessa Wei Feng. Classifying Arguments by Scheme. Master's thesis, Department of Computer Science, University of Toronto, 2010.

Vanessa Wei Feng and Graeme Hirst. Classifying Arguments by Scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

David Ferrucci and Adam Lally. Building an example application with the unstructured information management architecture. *IBM Systems Journal*, 43(3):455–475, 2004.

Ronald Aylmer Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, London, 1932.

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

James B Freeman. *Dialectics and the macrostructure of arguments: A theory of argument structure*, volume 10. Walter de Gruyter, 1991.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-88-6.

Thomas F Gordon and Douglas Walton. The Carneades argumentation framework. *Frontiers in Artificial Intelligence and Applications*, 144:195, 2006.

Brigitte Grote, Nils Lenke, and Manfred Stede. Ma(r)king concessions in English and German. *Discourse Processes*, 24(1):87–118, 1997.

Adrian Groza and Sergiu Indrie. Enacting Social Argumentative Machines in Semantic Wikipedia. *arXiv preprint arXiv:1304.5554*, 2013.

R. Guha, Rob McCool, and Eric Miller. Semantic Search. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 700–709, New York, NY, USA, 2003. ACM. ISBN 1-58113-680-3. doi: 10.1145/775152.775250.

Yufan Guo, Ilona Silins, Ulla Stenius, and Anna Korhonen. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. *Bioinformatics*, 29(11):1440–1447, 2013.

Iryna Gurevych, Max Mühlhäuser, Christof Müller, Jürgen Steimle, Markus Weimer, and Torsten Zesch. Darmstadt knowledge processing repository based on uima. In *Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the Society for Computational Linguistics and Language Technology*, Tübingen, Germany, April 2007.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

Gerhard und Joachim Buscha Helbig. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht.* Langenscheidt, 1996.

Stella Heras, Katie Atkinson, Vicente J Botti, Floriana Grasso, Vicente Julián, and Peter McBurney. How Argumentation can Enhance Dialogues in Social Networks. *Computational Models of Argument (COMMA 2010)*, 216:267–274, 2010.

Alexander Hogenboom, Frederik Hogenboom, Uzay Kaymak, Paul Wouters, and Franciska De Jong. Mining economic sentiment using argumentation structures. In *Advances in Conceptual Modeling–Applications and Challenges*, pages 200–209. Springer, 2010.

Constantin Houy, Tim Niesen, Peter Fettke, and Peter Loos. Towards automated identification and analysis of argumentation structures in the decision corpus of the German Federal Constitutional Court. In *Digital Ecosystems and Technologies (DEST), 2013 7th IEEE International Conference on*, pages 72–77. IEEE, 2013.

Rick Iedema, S. Feez, and P.R.R. White. *Media Literacy*. AMES (Adult Migrant English Service) NSW Publications, Sydney, 1994. ISBN: 978-1-921075-47-6.

Rick Iedema, Susan Feez, and Peter White. Appraisal and Journalistic Discourse. *extracted from Ibid., Media Literacy, Disadvantaged Schools Program, NSW Department of School Education, Sydney (available at www. grammatics. com/appraisal/appraisalandmediadiscourse/section2, accessed July 2006)*, 2003.

JabRef Development Team. *JabRef*, 2013.

Joel Katzav and Chris Reed. A Classification System for Arguments. *Division of Applied Computing, University of Dundee Technical Report*, 2004.

Paul A Kirschner, Simon J Buckingham-Shum, and Chad S Carr. *Visualizing argumentation: Software tools for collaborative and educational sense-making*. Springer, 2003.

Roland Kluge. Identifying Argumentation Structures in Controversial Educational Web Documents – Annotation Guidelines. Internal document, 2013.

Varada Kolhatkar and Graeme Hirst. Resolving "This-issue" Anaphora. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1255–1265, Jeju Island, Korea, 2012. ACL.

Varada Kolhatkar, Heike Zinsmeister, and Graeme Hirst. Annotating Anaphoric Shell Nouns with their Antecedents. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 112–121, Sofia, Bulgaria, 2013. ACL.

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

Klaus Krippendorff. Measuring the reliability of qualitative text analysis data. *Quality & quantity*, 38:787–800, 2004a.

Klaus Krippendorff. Reliability in content analysis. *Human Communication Research*, 30(3): 411–433, 2004b.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage, 2012.

Harold L. Kundel and Marcia Polansky. Measurement of observer agreement. *Radiology – Statistical Concepts Series*, 228:303–308, 2003.

John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. AIFdb:Infrastructure for the Argument Web. In *COMMA*, pages 515–516, 2012.

Xiaoqing Frank Liu, Ekta Khudkhudia, Lei Wen, Vamshi Sajja, and M Leu. An intelligent computational argumentation system for supporting collaborative software development decision making. *Artificial Intelligence Applications for Improved Software Engineering Development, Farid Meziane and Sunil Vadera, Hershey, PA: IGI Global*, pages 167–180, 2009.

Clare Llewellyn. Using Argument Analysis to Define a Structure for User Generated Content (Ph.D. Thesis Proposal), 2012.

Ronald P. Loui, Jeff Norman, Joe Altepeter, Dan Pinkard, Dan Craven, Jessica Linsday, and Mark Foltz. Progress on Room 5: A Testbed for Public Interactive Semi-formal Legal Argumentation. In *Proceedings of the 6th International Conference on Artificial Intelligence and Law*, ICAIL '97, pages 207–214, New York, NY, USA, 1997. ACM. ISBN 0-89791-924-6. doi: 10.1145/261618.261655.

Manuel Maarek. On the extraction of decisions and contributions from summaries of French legal IT contract cases. In *Proceedings of LREC 2010 - Workshop 23*, page 30, Malta, 2010.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. Identifying High-Level Organizational Elements in Argumentative Discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–28, Montréal, Canada, 2012. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Annotating discourse connectives and their arguments. In *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, Boston, Massachusetts, USA, 2004.

Raquel Mochales and Marie-Francine Moens. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22, 2011.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. Automatic Detection of Arguments in Legal Texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL 2007, pages 225–230, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-680-6. doi: 10.1145/1276318.1276362.

Emanuela Moreale and Maria Vargas-Vera. Genre analysis and the automated extraction of arguments from student essays. In *The Seventh International Computer Assisted Assessment Conference (CAA-2003)*, pages 8–9, 2003.

Subhabrata Mukherjee and Pushpak Bhattacharyya. Sentiment Analysis in Twitter with Lightweight Discourse Analysis. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1847–1864, Mumbai, India, 2012.

Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with MMAX2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214, 2006.

Susan Newman and Catherine Marshall. Pushing Toulmin too far: Learning from an argument representation scheme. *Xerox PARC Tech Rpt SSL-92-45*, 1992.

Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107, New York, NY, USA, 2009. ACM.

Dae Hoon Park and Catherine Blake. Identifying comparative claim sentences in full-text scientific articles. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 1–9. Association for Computational Linguistics, 2012.

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, 2013.

Lucie Poláková, Jiří Mírovskỳ, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, Japan, 2013.

John L Pollock. *Cognitive carpentry: A blueprint for how to build a person*. The MIT Press, 1995.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. The Penn Discourse Treebank 2.0 Annotation Manual. Technical report, University of Pennsylvania, 2007.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 28–30, Marrakech, Morocco, 2008.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartig. *A Grammar of Contemporary English*. Longman, 1980.

Altaf Rahman and Vincent Ng. Supervised Models for Coreference Resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, EMNLP '09, pages 968–977, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-62-6.

John D. Ramage, John C. Bean, and June Johnson. *Writing Arguments: A Rhetoric with Readings, Concise Edition*. Longman, 2008.

Chris Reed. Preliminary results from an argument corpus. *Linguistics in the twenty-first century*, pages 185–196, 2006.

Chris Reed and Glenn Rowe. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979, 2004.

Jodi Schneider, Brian Davis, and Adam Wyner. Dimensions of argumentation in social media. In *Knowledge Engineering and Knowledge Management*, pages 21–25. Springer, 2012.

William A Scott. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325, 1955.

Alla Vitaljevna Smirnova. Reported speech as an element of argumentative newspaper discourse. *Discourse & Communication*, 3:79–103, Febuary 2009.

Robin Smith. Aristotle's logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition, 2014.

Swapna Somasundaran, Galileo Namata, Lise Getoor, and Janyce Wiebe. Opinion Graphs for Polarity and Discourse Classification. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, TextGraphs-4, pages 66–74, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-54-1.

Jobien Sombekke, Tom van Engers, and Henry Prakken. Argumentation structures in legal dossiers. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 277–281. ACM, 2007.

Manfred Stede and Silvan Heintze. Machine-assisted Rhetorical Structure Annotation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, pages 425–431, Geneva, Switzerland, 2004. Association for Computational Linguistics. doi: 10.3115/1220355.1220416.

Manfred Stede and Antje Sauermann. Linearization of arguments in commentary text. In *Proceedings of the Workshop on Multidisciplinary Approaches to Discourse*, Oslo, 2008.

Amber Stubbs. MAE and MAI: lightweight annotation and adjudication tools. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 129–133. Association for Computational Linguistics, 2011.

John M. Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge Univ. Press, 1990. ISBN: 978-0-521-32869-2.

Maite Taboada. Discourse Markers as Signals (or Not) of Rhetorical Relations. *Journal of Pragmatics*, 38(4):567–592, 2006.

Maite Taboada and María de los Ángeles Gómez-González. Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences*, 6(1-3):17–41, 2012.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.

Imad Tbahriti, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. Using argumentation to retrieve articles with similar citations: an inquiry into improving related articles search in the MEDLINE digital library. *International Journal of Medical Informatics*, 75(6):488–495, 2006.

Simone Teufel. *Argumentative zoning: Information extraction from scientific text*. PhD thesis, University of Edinburgh, 1999.

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*, pages 110–117. Association for Computational Linguistics, 1999.

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, 2009. Association for Computational Linguistics.

Erik F Tjong Kim Sang and Sabine Buchholz. Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 127–132, Lisbon, Portugal, 2000. ACL.

Fatemeh Torabi Asr and Vera Demberg. On the Information Conveyed by Discourse Markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 84–93, Sofia, Bulgaria, 2013. ACL.

Stephen E. Toulmin. *The uses of argument*. Cambridge University Press, 1958.

Tim Van Gelder. The rationale for Rationale(TM). *Law, probability and risk*, 6(1-4):23–42, 2007.

Artem Vovk. Discovery and Analysis of Public Opinions on Controversial Topics in the Educational Domain. Master's thesis, Ubiquitious Knowledge Processing, TU Darmstadt, 2013.

Douglas Walton. The three bases for the enthymeme: A dialogical theory. *Journal of Applied Logic*, 6(3):361–379, 2008.

Douglas Walton. Argumentation theory: A very short introduction. In *Argumentation in artificial intelligence*, pages 1–22. Springer, 2009.

Douglas Walton. Argument mining by applying argumentation schemes. *Studies in Logic*, 4(1): 38–64, 2011.

Douglas Walton. Using Argumentation Schemes for Argument Extraction: A Bottom-Up Method. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 6(3):33–61, 2012. doi: 10.4018/jcini.2012070103.

Douglas Walton, Professor Christopher Reed, and Dr Fabrizio Macagno and. *Argumentation Schemes*, volume 1st edition. Cambridge University Press, 2008.

Douglas N Walton. *Argumentation schemes for presumptive reasoning*. Routledge, 1996.

Moshe Wasserblat, Ezra Daya, Eyal Hurvitz, Maya Gorodetsky, Dmitri Volsky, Ido Dagan, Meni Adler, Asher Steren, Sebastian Pado, Tae-Gil Noh, et al. Introduction to the EXCITEMENT project: towards an open platform for EXploring Customer Interactions through Textual entailMENT. In *Afeka-AVIOS Speech Processing Conference 2012*, 2012.

Barbara White. Identifying Sources of Inter-annotator Variation: Evaluating Two Models of Argument Analysis. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 132–136, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 978-1-932432-72-5.

Theresa Wilson and Janyce Wiebe. Annotating opinions in the world press. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 13–22, Hannover, Germany, 2003.

Adam Wyner and Tom van Engers. A framework for enriched, controlled on-line discussion forums for e-government policy-making. In *Proceedings of eGov 2010*, 2010.

Şaban Hsan Yalçinkaya. An inter-annotator agreement measurement methodology for the Turkish discourse bank (TDB). Master's thesis, Middle East Technical University, Ankara, 2010.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. WebAnno: A Flexible,Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 1–6, Stroudsburg, PA, USA, August 2013. Association for Computational Linguistics.

List of Figures

List of Tables

Index

Erklärung zur Master-Thesis

Hiermit versichere ich, die vorliegende Master-Thesis ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den April 10, 2014

____

(Roland Kluge)