The ReCAP Corpus: A Corpus of Complex Argument Graphs on German Education Politics

Lorik Dumani, Manuel Biertz, Alex Witry, Anna-Katharina Ludwig, Mirko Lenz, Stefan Ollinger,
Ralph Bergmann, Ralf Schenkel
Trier University
D-54286 Trier, Germany
{dumani, biertz, s4alwitr, s2aaludw, s4mtlenz, ollinger, bergmann, schenkel}@uni-trier.de

***Abstract*—The automatic extraction of arguments from natural language texts is a highly researched area and more important than ever today, as it is nearly impossible to manually capture all arguments on a controversial topic in a reasonable amount of time. For testing different algorithms such as the retrieval of the best arguments, which are still in their infancy, gold standards must exist. An argument consists of a claim or standpoint that is supported or opposed by at least one premise. The generic term for a claim or premise is Argumentative Discourse Unit (ADU). The relationships between ADUs can be specified by argument schemes and can lead to large graphs. This paper presents a corpus of 100 argument graphs with about 2,500 ADUs in German, which is unique in its size and the utilisation of argument schemes. The corpus is built from natural language texts like party press releases and parliamentary motions on education policies in the German federal states. Each high-quality text is presented by an argument graph and created by the use of a modified version of the annotation tool OVA. The final argument graphs resulted by merging two previously independently annotated graphs based on detailed discussions.**

I. Introduction

Argument mining (AM), i.e. the automatic extraction of arguments in natural language texts, is a nascent field in computational argumentation. It goes beyond topic identification, summarisation, and stance detection and identifies the reasons provided in a text to follow a course of action or to accept a judgement. An *argument* is generally understood to be a combination of a *claim* (or conclusion) and a *premise* (or reason), and an inference rule linking the two [1]. Inference within this micro-structure can further be explained as *argument scheme* which can support or limit the conclusion. Besides the inference, an argument scheme also describes the pattern of premises which are pertinent for the inference. A deeper explanation of these schemes is given in Section II of this paper.

Research on AM needs high-quality pre-annotated corpora in order to verify the validity of approaches. By now existing corpora either have serious issues with regard to their quality, or they are available only in English language: The Internet Argument Corpus (IAC) [2], which consists of 390,704 posts in 11,800 discussions that were extracted from the online debate site 4forums.com is created with “little argumentation theory sitting behind it” [3]. AIFdb [4] is a database that allows to store and to retrieve argument structures in the Argument Interchange Format (AIF) [5] and could provide a good source, but as nearly all corpora most of the arguments are in English language only. The Potsdam Microtexts Corpus [6], being the exception to the rule, consists of short texts that respond to a trigger question. It does not fulfil our requirements in three ways: First, it does not utilise argument schemes, which are helpful for both human annotators and machine learning algorithms [7]. Second, the corpus is artificially created so that every document or graph is composed of about five elements (conclusions and premises), thus it does not provide a real world scenario. Third, with only 112 graphs[1] and the corresponding number of premises and conclusions, the corpus is rather small [3].

Hence, we created a new corpus which avoids these pitfalls. In this paper, we present a high-quality corpus in German language which we make publicly available to the argument mining and argument retrieval community, built of texts ranging in their length from press releases to election manifestos, and including not only inferences between claims and their respective premises but also argumentation schemes and hence additional information about the premises’ interrelations and the inference rule (or warrant) utilised. Annotations were conducted on German texts only, but since Lenz et al. [8] already worked with our corpus and used DeepL to translate the argument graphs into English, the corpus is available in both languages.[2]

This work is part of the ReCAP project [9] which is part of the DFG priority program robust argumentation machines (RATIO)[3]. Bergmann et al [9] propose an architecture of an argumentation machine which should reason on a knowledge level formed by arguments and argumentation structures, e.g. the corpus presented in this paper.

The remainder of this paper is organised as follows: Section II gives an introduction to argumentation theory. In particular, it explains the inference schemes used in our corpus. Then, Section III explains our modified variant of the OVA tool [10], which was used to annotate texts, i.e. the manual transformation of texts into graphs. Section IV describes the annotation process and annotator training, Section V gives a brief insight into the corpus, and Section VI shows some

[1] http://angcl.ling.uni-potsdam.de/resources/argmicro.html

[2] Both corpora can be downloaded via the following link: https://basilika.uni-trier.de/nextcloud/s/JePuLMGdZNBJmUK

[3] www.spp-ratio.de

possible applications for the corpus in recent research.

II. Argumentation Theoretical Considerations

In this section we scratch the surface of argumentation theory and describe argumentation schemes which are important for the inferences between ADUs. We will need this as the corpus is based on this theory.

Beginning in the 1950's, argumentation theory experienced a great turn which is largely attributed to Toulmin [11]. He argued against a scholarship largely focused on the logical soundness of arguments and dismissing anything else as fallacious, and advocated, taking the example of legal argumentation, an argumentation theory which originates from empirical practice, thus looks at argumentation as it occurs in everyday life. While the classical approach looked at the premises and the conclusion and analysed, whether the latter followed logically from the former, Toulmin recognised that the step from the data to the conclusion is often only presumptive and not logically cogent. In the following decades, several approaches emerged which followed his example.

Our approach is built on the work of Walton [12]. As a dialectical theory the approach perceives argumentation as a rule-guided exchange, is not interested in logical correctness or rhetorical success of an argument, and has essentially a dialogue in mind. As Walton denotes: "the offering of an argument presupposes a dialogue between two sides" [12]. A dialogue is understood as a "type of goal-directed conversation" of at least two people which take turns in their contributions [12]. To think of argumentation as dialogue is intuitive as long as we analyse parliamentary debates, for instance. However, we intend to include only monologic texts into our corpus. Thus, we assume that they are written with a dialogic intention directed towards an anonymous, at least larger, audience, because the reader "should raise critical questions about the argument he or she has been presented for acceptance" [12].

An argument consists of a claim or standpoint that is supported or opposed by at least one premise [13]. The claim is the central component of an argument and often also controversial [14]. Its acceptance is either increased or decreased by premises [15]. A support or attack relation is marked by an inference from the premise to the claim. Premises and claims can be subsumed under the generic term *argumentative discourse unit* (ADU) [16]. Since ADUs can have both outgoing and incoming edges, large graphs (called *argument graphs* in the following) can emerge. The main point of an argumentative text is a claim, referred to as the *major claim* [17].

Walton's major contribution is a comprehensive catalogue of argumentation schemes found in natural language texts [18], hence not derived from normative ideals (unlike Pragma-Dialectics' *argumentative patterns* [19]). Walton's *argumentation schemes* have been empirically discovered in natural language, thus they are less normative than other approaches [20], and are promising for machine learning and artificial intelligence application.

"Argumentation schemes are stereotypical patterns of reasoning"[4] and stand in the tradition of Aristotelian *topoi* [21][5]. They are a combination of inference and material relations [3], and facilitate the detection of arguments for both human annotators and AI algorithms [7]. This sets them apart from other approaches, for instance the very general Toulmin model [11] or the likewise general, but (over-)complex, Argumentum Model of Topics as coherently presented by Rigotti and Greco [20]. Schemes do not only facilitate detection, they also help to avoid biases. Furthermore, too much would be lost in a simple pro/contra-dichotomisation of arguments [23]. Instead, argument schemes allow for greater differentiation, e.g. whether an author reasons about the consequences of an action or whether he supports his claim with an expert's opinion. Even some of Toulmin's examples show great resemblance to schemes [11].

The inference built in argumentation schemes can be presumptive and defeasible, i.e. in contrast to deductive and inductive inferences the conclusion does not follow necessarily or with statistical probability from the premise(s), but is better understood as an assumption [12].

TABLE I: The scheme for argument from positive consequences.

ADU	Description
Premise	*If A is brought about, good consequences will plausibly occur.*
Conclusion	*Therefore, A should be brought about.*

To provide an example for such a scheme, Table I shows the argument from positive consequences as identified in the compendium of Walton et al. [18]. As the corpus is available in German language and we intend to address a wider readership, all the following examples are in English language only. The argument from positive consequences consists of two elements, a premise assuming positive consequences of action A and a conclusion that A should, therefore, be conducted [18]. Thus, this scheme has a *course of action* as its conclusion, and is inherently presumptive since its premise tries to forecast future consequences [12]. In "real" texts however, the premise rarely takes this form and the conclusion is often omitted. For example, the premise "*Introducing a universal basic income would improve people's living conditions*" already implies through the word "*improve*" that the consequences are judged to be positive, thus the sentence will rarely be succeeded by an additional value judgement reading "*and improving people's living conditions is a positive thing*" to act as support for the premise. Likewise, it is not necessary to spell the conclusion "*a universal basic income should be brought about*", though maybe it is not quite as unusual. Schemes can grow more complex when they involve several steps of reasoning.

Yet, notwithstanding their usefulness, the vast number of 60 different schemes (subschemes not included) identified by

[4] Note here that the authors use argumentation in the sense of micro-level arguments. Their argumentation schemes are hence patterns of single arguments.

[5] Hannken-Illjes describes argument schemes as formal topoi [22].

Walton et al. [18] complicates annotator training (students often have issues in differentiating the schemes [7]) and is quite demanding in the usage of computational resources [3]. Two solutions are available, either building a hierarchical classification system (as e.g. proposed by Walton and Macagno [21]) or choosing a subset of schemes adapted to the research purpose (as done by Hansen and Walton [24]).

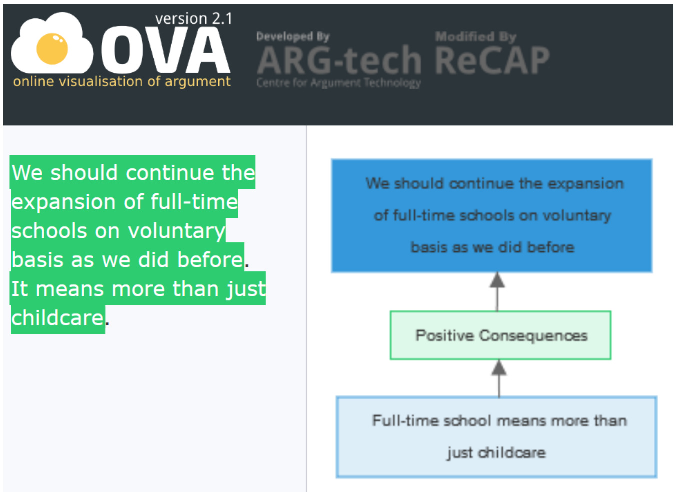
In order to store argumentation and respective schemes, the Argument Interchange Format (AIF) [5] and its argument graphs can be used. AIF represents an abstract model, which aims to represent and exchange arguments between various argumentation tools. Argument graphs store ADUs and the inference link (scheme) connecting them for multiple arguments. As they also store interrelations of arguments, they allow representation of the argumentative structure of a text (possibly concluding in one major claim of the text as a whole). The AIF ontology defines two types of nodes to build argument graphs: information nodes (*I-nodes*) and scheme nodes (*S-nodes*). I-nodes relate to content and represent ADUs that depend on the domain of discourse. Scheme nodes are divided further into subclasses: Rule application nodes (*RA-nodes*) denote specific inference relations, conflict application nodes (*CA-nodes*) denote specific inference relations. A rephrase of one proposition is marked with a rephrase relation (*MA-nodes*). Please note that argument schemes are only defined for RA-nodes. As the format is extendable there is potential for a theoretically unlimited number of scheme types. Nodes may have different attributes such as title, text, type (e.g. decision, action, goal, belief) and more. These attributes may vary upon the use case. A Node A supports a node B if and only if there is an edge connecting A to B. This means that the edges have an associated direction.

III. Modifications to OVA

Our annotations were performed with a modified version of the *Online Visualisation of Argument* tool (OVA) by Arg-Tech [10][6], which succeeds *Araucaria* [25] and provides a Web-based facility to annotate argumentation in text files in conformation with the AIF standard. Since argument schemes can be used independently from languages, OVA has already been used for annotations in a variety of languages, including Chinese, Hindi, and Ukrainian[7]. For our purposes, we set up an adapted version[8], while the use and the appearance remain largely identical.

Arg-Tech provides two versions of OVA, namely OVA and OVA+. The latter allows the user to specify locutions and transitions to the nodes according to Inference Anchoring Theory, and is designed in particular for the analysis of dialogues [10]. Since we intended to analyse large, monologic texts only, and OVA+ graphs grow utterly complex even with short texts, we decided for the use of OVA.

Figure 1 shows the graphical user interface of the modified tool. The content of a text file is taken as input on the left side.



We should continue the expansion of full-time schools on voluntary basis as we did before. It means more than just childcare.

We should continue the expansion of full-time schools on voluntary basis as we did before

Full-time school means more than just childcare

Fig. 1: Example of an annotated document with the modified OVA tool

The annotator creates an ADU (blue box) by selecting a text passage and clicking to a free space in the main frame on the right side. In the example, the argument constructed from the plain text consists of a claim (at the top) and a premise (at the bottom). The text in the generated nodes can also be modified and filled with new content, e.g. to resolve pronouns. These so-called "reconstructions" ensure that an ADU is self-contained and understandable without other ADUs. During annotation we tried to stay close to the text to create an ideal-typical argument. In the example the word "*It*" in the lower node has been modified to "*Full-time school*".

Edges are formed between multiple ADUs and specify the argument schemes. Those will be displayed as separate nodes (green box) with arrows depicting the direction of the edge. There are several types of edges using different colouring, while inference nodes (RA, green color) are the ones used most in our annotations. While in modified OVA inference nodes offer a list of possible schemes, all other edge types (e.g. Conflict and Rephrase) only provide one generic scheme. If a scheme has been chosen, the connected nodes can be marked according to their role, i.e. as conclusion or one of the several premises a scheme might contain. This is, however, not necessary, since otherwise it would not be possible to annotate enthymemes[9]. So-called descriptors – the "description"-column in Table I – help with the annotation to assign the roles, and have been adapted to further our annotators' understanding.

We added the *expert opinion descriptor* to each scheme, since we experienced that generally an argument from expert opinion – claiming the truth of a judgement by referring to the authority of a source – also includes other schemes, e.g. *argument from positive consequences*. This is especially the case when the author quotes the person at the end of a sentence, e.g. "*according to Kim Doe*". To account for this

[6] http://ova.arg-tech.org

[7] http://corpora.aifdb.org/araucaria

[8] http://ova.uni-trier.de

[9] Enthymemes are arguments where one or more premises or the conclusion is left implicit [18].

and to avoid complicated constructions with limited gain, it is now for example possible to add an expert assessment to an *argument from positive consequences* or any other scheme. The expert then supports the inference step included in the scheme.

In the example in Figure 1 the scheme *argument from positive consequences* was chosen as inference between premise and claim. The Positive Consequences scheme consists of three descriptors. The premise has the descriptor $desc_p$ "If A is brought about, good consequences will plausibly occur", the claim has the descriptor $desc_c$ "Therefore, A should be brought about". Optionally, a descriptor $desc_e$ can be assigned to an ADU if the statement comes from an expert. This has the role "Expert E asserts that proposition A is true/false". As illustrated in the example, the selection of this scheme and the assignment of the descriptors $desc_p$ and $desc_c$ as "*Full-time school means more than just childcare*" and "*We should continue the expansion of full-time schools on voluntary basis as we did before*" respectively is appropriate.

The modified OVA version also features new attributes to nodes. Many texts provide arguments to convince the reader of a core thesis. This core thesis is important for further steps and can be flagged as "major claim" [17] in modified OVA. Additionally, argument nodes now have attributes to indicate the start and end position in the origin text. This is important as annotators can alter the displayed text during reconstruction. The position attribute helps to keep track of the original text and to measure the inter-annotator agreement.

IV. Building the Corpus

A. Choice of Schemes

We decided not to build a classification system, but to determine an appropriate subset of schemes. Thus, we turned to the AIFdb database and searched for the most-used argumentation schemes[10]. Table II provides an overview of the most used schemes in the AIFdb in total, as well as the number of graphs, and the different data sets a scheme occurs in. Altogether, we crawled 102 data sets with a total of 10,622 graphs from the AIFdb database. Taking Table II into consideration, it is obvious that the graphs are relatively small. Despite the high number of uses of the *default inference* scheme (i.e. no scheme used at all), we found a large overlap between the schemes of the most-used schemes in AIFdb and Hansen and Walton [24], who as we do also studied political texts, namely the Ontario provincial election campaigns from 2011. Hence, we built a subset of 18 schemes and one residual scheme (*default inference*) based on the list of Hansen and Walton, which we later on expanded based on our annotators' feedback[11].

Table III illustrates the schemes used by Hansen and Walton in comparison with our scheme set. Please note that 95.3 % of the total of 256 arguments, which they collected, have been classified with the scheme set as depicted in Table III. The remaining 4.7 % could not be classified at all.

TABLE II: Analysis of the most used schemes on AIFdb. Schemes (same or variants) we used also in our scheme set are emphasised. Dataset from August, 2019.

Scheme	Total Occurrences	Occurrences in unique graphs	Occurrences in data sets in AIFdb
Default Inference	22984	5175	92
JP-Reason	713	148	5
Example	316	201	29
Argument To Moral Virtue	180	90	3
ERPractical Reasoning	164	56	3
Cause To Effect	97	40	7
Expert Opinion	94	83	20
Argument To Practical Wisdom	82	50	3
Positive Consequences	63	43	6
Evidence To Hypothesis	41	23	6
Practical Reasoning	40	16	7
Analogy	31	30	13
Argument To Good Will	30	26	3
Argument From Authority	26	22	10
Reason	26	9	1
Negative Consequences	21	15	6
Popular Opinion	20	12	11

TABLE III: Comparison of scheme sets. ✓ marks the use of a scheme set. (✓) denotes that it is unclear which version of the scheme is utilised. Walton et al. [18] is incomplete and serves only as illustration.

	Walton et al. [18]	Hansen & Walton [24]	ReCAP
Position to Know	✓	✓	✓
Expert Opinion	✓	✓ (Authority)	✓
Popular Opinion	✓		✓
Example	✓		✓
Analogy	✓	✓	✓
Alternatives (cognitive schemes)	✓	(✓)	✓
Alternatives (normative schemes)	✓	(✓)	
Values	✓	✓	
Practical Reasoning	✓		✓
Cause to Effect	✓	✓	✓
Correlation to Cause	✓	✓	
Sign	✓	✓	✓
Positive Consequences	✓	✓	✓
Negative Consequences	✓	✓	✓
Generic Ad Hominem	✓	✓	
Inconsistent Commitment	✓	✓	✓
Circumstantical Ad Hominem	✓	✓	✓
Rule	✓		✓
Fairness		✓	✓
Unfairness		✓	✓
Misplaced Priorities		✓	✓
Appeal to Sympathy		✓	
Explanation		✓	
Residual Category		✓ (Can't classify)	✓ (Default Inference)

[10] http://corpora.aifdb.org/

[11] We cannot discuss the use of every single scheme here, but see for example [26] on the importance of practical argumentation in politics.

B. Annotator Training and Process

Our annotations were conducted by two student assistants coming from political science, media studies, as well as law studies. Hence, they are educated and trained to work with argumentative texts written for public discourse.

The annotators completed a tutorial of about four hours in which we introduced them to the OVA-Tool, a pre-built set of scheme examples from our own annotations, and our extensive annotation guidelines[12] which are based on Stab and Gurevych [17] but also include argumentation schemes. Furthermore, theoretical fundamentals were built for a deeper understanding of the project. The first weeks of the annotating-process were supervised and guided more closely – as problems tend to occur during the actual process of annotating –, and we stayed in close contact in order to permanently receive and give feedback. For instance, this exchange produced rules to solve situations where more than one possible major claim could be identified and a hierarchy of schemes to apply when different schemes were conceivable.

After the training, annotation work was bipartite: in a first step the annotators worked on each text individually based on the experience they gained and the annotation guidelines. Secondly, both individual versions of the same graph are merged into a final one. This merging process was not an automated one but performed manually and discoursively by the annotators. It consisted of a three-step-system that was applied in every merging process: identifying the major claim and their concord, the overall structure of the individual graphs and finally the discursive merging based on one of the annotations. That guaranteed an ongoing exchange, served as a reliability check, and provided an opportunity to discuss minor problems, e.g. in the understanding of a text. Working in direct and frequent exchange with the annotators, unsolvable problems emerging at this stage were discussed in bi-weekly ReCAP plenums to decide for the best solution.

While only the merged versions are to be considered the reliable gold standards, this approach nonetheless produced three high-quality annotations for each document (in German language), which can be used for further tests and other purposes as there exist few German corpora.

C. Inter-Annotator Agreement

For the inter-annotator agreement, we evaluated two aspects: first, the segmentation into ADUs, and second, the agreement of decisions on which claim was selected as the major claim. We did not evaluate the choice of schemes because the texts were independently annotated and resulted in different graphs. For an inter-annotator agreement for schemes, graphs would need to be identical beforehand (apart from the chosen scheme).

We have measured the inter-annotator agreement for making the ADUs with Cohen's κ [27] which also takes coincidental agreements into account. Its formula is $\kappa = \frac{Pr(a)-Pr(e)}{1-Pr(e)}$ where $Pr(a)$ is the relatively observable agreement and $Pr(e)$ is the hypothetical probability of a random agreement [28]. An agreement of $\kappa = 1$ is perfect, an agreement of $\kappa = 0$ equals coincidence. We built a set of possible ADUs that the annotators could have annotated as follows: since ADUs are generally separated by punctuation marks, we split all 100 texts into a total of 18,920 clauses by using separators ".", ",", "?", "!", ":", ";". As some ADUs also cross sentence boundaries, we also included the ADUs that were actually set by the two annotators. For each of the obtained ADUs to be checked we then examined whether the annotators tagged their start and/or end positions. For beginning positions we reached $\kappa = 0.5595$, for ending $\kappa = 0.5989$, and for both together, i.e. equal ADUs, we obtained $\kappa = 0.4982$. We can observe that the agreement for ending positions of ADUs is higher than for starting positions. Both of these values are higher than the value for complete ADUs. According to Landis and Koch [29], $\kappa \in [0.41, 0.60]$ implies a fair agreement. Hence, it is very important that the annotators first annotated the graphs independently of each other and then came to an agreement on how the gold standard should look like.

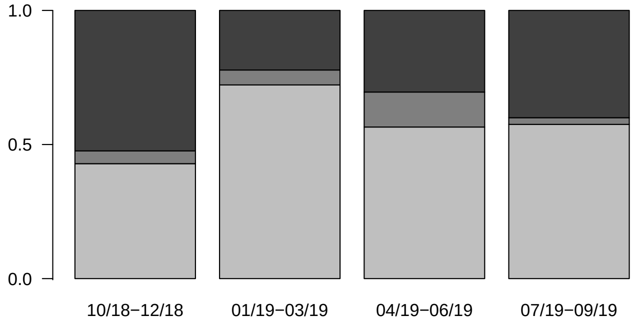


Fig. 2: Proportions of identical (light grey), intersecting (medium grey), and diverging (dark grey) major claims found in the merging process, grouped by quarters.

Figure 2 shows the proportions of identical and diverging major claims found in the individual phase of the annotation process over time. Besides detecting identical and divergent major claims, we also noticed a hybrid state of intersecting major claims. Whilst both annotators identified the major claim as two different I-nodes, their content and declaration were similar in some way. We used to describe them in the term of "semantically identical". Those intersecting major claims could differ regarding their topic, coherence and how they were reconstructed by the annotator, yet it was clear in the merging process that they meant the same.

D. Reflexions on Annotation Process

Two aspects are noteworthy: first, each of the annotators developed an individual style of annotating which caused huge differences in the first quarter annotations. This is also reflected in the high proportion of diverging major claims. This margin got reduced over time as both annotators drew advantage from the discursive merging process and individual

[12] The link to the annotation guidelines can be found here: https://basilika.uni-trier.de/nextcloud/s/BHg3vyQwYxDIrpA

errors continuously started to disappear. However, while the annotations as a whole became more similar, a noteworthy divergence of identified major claims remained. This second trend demonstrates that identifying the main intention of a text is far more interpretative than its component arguments. It is not uncommon for the annotators to identify the same premises and conclusions and find different major claims nonetheless. We have found that a major claim has an average length of 135 characters and occurs after an average of 32 % of the characters in the text. We investigated this appearance more closely and found that the largest fraction of major claims (58 out of 100) occurred in the first 25 % of the text. In 13 cases they occur between 25 and 50 % of the characters, in 11 cases between 50 and 75 % and in 16 cases in the last 25 % of the text. Two cases could not be identified. For the distribution across text types, see Figure 4. In order to resolve divergences we developed two merging rules with regard to major claims: normativity trumps description, more content trumps less content (as long as it still constitutes a conclusion).

V. The Corpus

Our final corpus consists of 100 argument graphs and is publicly available.[13] Used text sources are German state parliamentary motions, press releases, position papers, and party manifestos which deal with education issues in Bavaria (BY), Hamburg (HH), and Rhineland Palatinate (RLP). Topics covered are reforms of school systems in general, the integrative comprehensive school (RLP), the Stadtteilschule (a Hamburgian alternative to the Gymnasium), and digitisation as a general challenge (particularly BY). Figure 3 shows a relatively large argumentation graph in our corpus (176 nodes, 154 edges) and illustrates how complex these graphs can be.

Documents were either available as Web page or as portable document file, so we decided to save Web pages as PDFs as well. Due to the text positions which were saved for the mapping of the nodes, the PDFs had to be transformed to text files in a standardised manner, which also removed hyphenations, kept paragraphs together, and the more. For this purpose, we wrote a text extractor for the PDF files, whose outcomes needed slight manual editing only. In total, the 100 documents contain 2,479 premises and conclusions (argumentative discourse units (ADUs) in terms of Peldszus and Stede [16]). The total number of edges in our corpus is 2,281, whereof the large majority (91.1 %) are represented as inference nodes. The remaining part is split between conflict nodes (4.3 %) and rephrase nodes (4.6 %). Please note that these two types are not enriched with schemes and descriptors. The average numbers of nodes and edges are 25.33 and 20.78, median numbers of nodes and edges are 17 and 15, respectively, while they include 1.112 sentences and 6.789 words on average, 1 and 6 on median, respectively.

The selection of the documents followed three main considerations: (1) a rough thematic coherence per state without neglection of topical diversity, (2) an experience-based intuition regarding the argumentative quality of the text, and (3) public availability. Education as a topic has been chosen since it is one of the few domains where the German *federal states* enjoy an exclusive legislative competence in the Federal Republic of Germany (article 70 (1) in conjunction with articles 73 and 74 of the German federal constitution). We can thus profit from different education systems, since the diverging education discourses, which nonetheless share a language and general concepts (like *Abitur* (comparable to the A-levels), *Grundschule* (comparable to primary school), and the more), enable us to test for the transferability of arguments by case-based reasoning (CBR) as described by Bergmann et al. [9].

TABLE IV: Analysis of used schemes in ReCAP Corpus.

Scheme	Total Occurrences	Occurrences in unique graphs
Positive Consequences	471	76
Practical Reasoning	353	61
Negative Consequences	297	62
Sign	280	78
Cause to Effect	195	50
Example	173	49
Unfairness	48	23
Misplaced Priorities	47	31
Rule	41	22
Fairness	31	20
Alternatives (Cognitive Schemes)	29	16
Position to Know	28	22
Popular Opinion	28	23
Circumstantial Ad Hominem	13	12
Expert Opinion	13	8
Danger Appeal	12	10
Analogy	6	6
Inconsistent Commitment	6	5
Default Inference	7	6
Default Rephrase	106	61
Default Conflict	97	36

Table IV shows the most used schemes on inference nodes in our corpus. In comparison to AIFdb (Table II) our corpus shows a different proportion of scheme usage. *Default inference* has only been used 5 times in our corpus, while it is the most common inference scheme in the AIFdb collection. Our corpus specifies the inferences between claim and premises more precisely which results in higher quality graphs. However, schemes that we also use in our annotations are highlighted in Table II. One common pattern which appeared in the argument structures is that the major claim forms a central premise from which multiple other claims are derived using the *practical reasoning* argumentation scheme.

VI. Possible Applications

There is a plethora of literature on argument mining, and the “post-processing” of mined arguments, some of them have already been discussed in the introduction.

In the combined approach of Lawrence and Reed [30], they use argument scheme structure as one of three layers. The other layers are identification using discourse indicators and topical similarity. In the argument scheme layer, they

[13] Due to German copyright law, we can share the corresponding plain texts only on request from the authors.

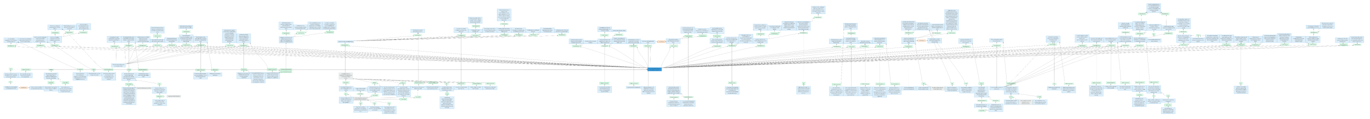


Fig. 3: Example of a large argumentation graph, illustrating argumentation complexity.

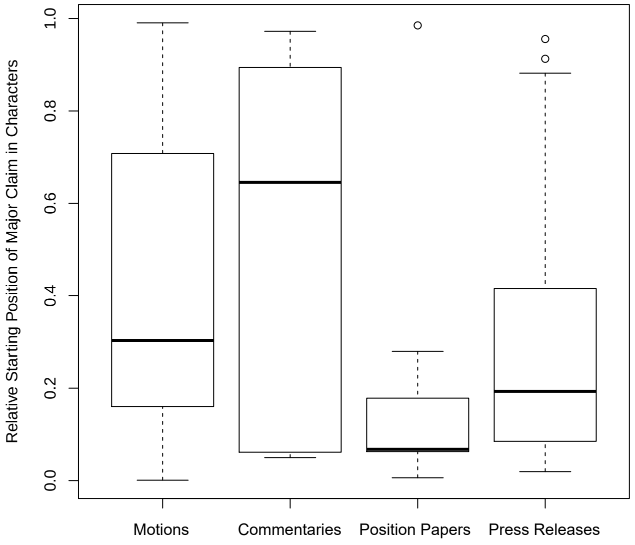


Fig. 4: Distribution of Relative Major Claim Positions Across Text Types. Total numbers: 17 motions, 9 position papers, 66 press releases, 5 newspaper commentaries, and 1 manifesto (not shown).

use argument scheme structure information obtained with a Naïve Bayes classifier of the *scikit-learn*[14] Python package, trained on the AIFdb data and fed with a manually crafted list of keywords for each descriptor in the scheme, enlarged by similar words from WordNet. Yet, they do this only for the two schemes *argument from expert opinion* and *argument from positive consequences*. Together with the other layers, they achieve impressive results of 0.91 precision, and 0.77 recall (F_1, 0.83), yet they test only on 36 preconnected propositions with two schemes. If applied to our corpus, the test could be repeated on a larger scale, with more schemes, and would thus achieve greater validity.

Lenz et al. [31] use the Potsdam Microtexts Corpus [6] to evaluate graph similarity measurements, but they suffer all the disadvantages of the corpus as mentioned in Section I. With the new ReCAP corpus, they would profit from argument graphs of greater complexity which are based on non-artificial texts.

ArgumenText is a system presented by Stab et al. [32] "for topic-relevant argument search in heterogeneous texts"[15]. It follows a similar idea as the ReCAP argumentation machine in that it mines for arguments in large amounts of Web-scraped texts, though they do not make use of argument schemes. The system is able to output ranked arguments for a given topic, i.e. user-given query. Results for German language are poor with virtually no directly relevant argument and a lot of absolutely unrelated items in the output. Performance in English is significantly better, yet the share of unrelated items is still large[16]. In their own evaluation, they report a high recall of 0.89 compared to expert annotators but a low precision of only 0.47, which explains the high output of non-related items. Results of Lawrence and Reed [30] indicate that precision could be optimised with the help of argument schemes[17].

We also provide a python framework[18] to ensure that the graphs can be loaded, used, or processed. It is also at PyPI[19].

VII. Concluding Remarks

We presented a new gold standard corpus for the training and testing of argument mining approaches which is made publicly available. Its feature set is unique: It is built from natural language texts, which are drawn from three similar yet different discourses, annotated with a carefully crafted list of argument schemes, and, finally, it is one of the few corpora in German language. There are already three findings worth reporting here:

First, political science often works with party manifestos, e.g. in spatial analysis of party positions. Party manifestos are aggregated political beliefs of party members weighed by their importance and hence in general an important object of study. An argument analysis yet reveals that – though voluminous – manifestos are more like demand catalogues. While certainly explainable it goes against a political science intuition of manifestos as high quality sources[20]. Second, annotator training is key for reliable results. Ambiguities can produce great differences in human output, as our annotators know to report. This issue is widely neglected in argument mining literature, but should experience greater attention to achieve comparability between studies, and reliability of

[14] https://scikit-learn.org/

[15] Available on https://www.argumentsearch.com.

[16] These impressions are based on a number of queries entered into ArgumenText.

[17] Discourse indicators help as well, but they occur only infrequent and, hence, are not sufficient. Lawrence and Reed [30] report 0.04 recall for discourse indicators alone.

[18] https://github.com/ReCAP-UTR/Argument-Graph

[19] https://pypi.org/project/recap-argument-graph/

[20] See e.g. the extensive work of the Manifesto Project, https://manifestoproject.wzb.eu. This impression is not only build on the manifesto in the corpus but as well on the many more considered for analysis.

checks against gold standards. Third, a new annotator training based on a classification of argumentation schemes as often discussed [21], [3], [7], [20], could facilitate scheme recognition by annotators and hence allow to build a more comprehensive subset of schemes. In particular, it would be useful to include the "Explanation" category from Hansen and Walton [24] in future research, which would both help with the differentiation between arguments and explanations, and provide contextual information in the final argument graph.

In future work we will find and build arbitrary queries to develop and enhance tools for information retrieval and case-based reasoning which use our corpus. In previous work, we investigated textual similarity methods to find similar claims for query claims [33] as well as similarity methods for graph retrieval [31], [34]. Lenz et al. [8] already made first use of our corpus to test an end-to-end argumentation mining pipeline and used DeepL[21] to obtain the argument graphs presented in this paper in English language.

Acknowledgment

We would like to thank Lasse Cronqvist and Simon Jakobs for their invaluable help in finding a subject for the corpus that is perfectly suited to our task and, amongst other things, allows the synthesis of arguments. We would also like to thank them for finding very competent student assistants, who have created this corpus.

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project ReCAP, Grant Number 375342983 - 2018-2020, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

References

[1] M. Moens, "Argumentation mining: How can a machine acquire common sense and world knowledge?" Argument & Computation, vol. 9, no. 1, pp. 1–14, 2018.

[2] M. A. Walker, J. E. F. Tree, P. Anand, R. Abbott, and J. King, "A corpus for research on deliberation and debate," in LREC, 2012, pp. 812–817.

[3] F. Macagno, D. Walton, and C. Reed, "Argumentation schemes. history, classifications, and computational applications," FLAP, vol. 4, no. 8, 2017.

[4] J. Lawrence, "About AIFdb," 2017, last access: November 08, 2019.

[5] C. I. Chesñevar, J. McGinnis, S. Modgil, I. Rahwan, C. Reed, G. R. Simari, M. South, G. Vreeswijk, and S. Willmott, "Towards an argument interchange format," Knowledge Eng. Review, vol. 21, no. 4, pp. 293–316, 2006.

[6] A. Peldszus and M. Stede, "Rhetorical structure and argumentation structure in monologue text," in ArgMining@ACL, 2016.

[7] F. Macagno and D. Walton, "Classifying the Patterns of Natural Arguments," Philosophy & Rhetoric, vol. 48, no. 1, p. 26–53, 2015.

[8] M. Lenz, P. Sahitaj, S. Kallenberg, C. Coors, L. Dumani, R. Schenkel, and R. Bergmann, "Towards an argument mining pipeline transforming texts to argument graphs," in Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020. IOS Press, 2020, pp. 263–270.

[9] R. Bergmann, R. Schenkel, L. Dumani, and S. Ollinger, "Recap - information retrieval and case-based reasoning for robust deliberation and synthesis of arguments in the political discourse," in LWDA, 2018, pp. 49–60.

[10] M. Janier, J. Lawrence, and C. Reed, "OVA+: an argument analysis interface," in COMMA, 2014, pp. 463–464.

[11] S. E. Toulmin, The Uses of Argument, updated ed. Cambridge University Press, 2003.

[12] D. Walton, Fundamentals of Critical Argumentation, ser. Critical Reasoning and Argumentation. Cambridge University Press, 2006.

[13] M. Stede, S. D. Afantenos, A. Peldszus, N. Asher, and J. Perret, "Parallel discourse annotations on a corpus of short texts," in LREC, 2016.

[14] C. Stab and I. Gurevych, "Identifying argumentative discourse structures in persuasive essays," in EMNLP, 2014, pp. 46–56.

[15] F. H. van Eemeren, B. Garssen, E. C. W. Krabbe, A. F. S. Henkemans, B. Verheij, and J. H. M. Wagemans, Eds., Handbook of Argumentation Theory. Springer, 2014.

[16] A. Peldszus and M. Stede, "From argument diagrams to argumentation mining in texts: A survey," IJCINI, vol. 7, no. 1, pp. 1–31, 2013.

[17] C. Stab and I. Gurevych, "Guidelines for annotating argumentation structures in persuasive essays," 2016, last access: November 08, 2019.

[18] D. Walton, C. Reed, and F. Macagno, Argumentation Schemes. Cambridge University Press, 2008.

[19] F. H. van Eemeren, "Identifying Argumentative Patterns," Argumentation, vol. 30, no. 1, p. 1–23, Mar 2016.

[20] E. Rigotti and S. Greco, Inference in Argumentation, ser. Argumentation Library. Springer Nature, 2019, no. 34.

[21] D. Walton and F. Macagno, "A classification system for argumentation schemes," Argument & Computation, vol. 6, no. 3, pp. 219–245, 2015.

[22] K. Hannken-Illjes, Argumentation, ser. Narr Studienbücher. Narr Francke Attempto, 2018.

[23] T. Niehr and K. Böke, "Diskursanalyse unter linguistischer Perspektive," in Forschungspraxis, 4th ed., ser. Interdisziplinäre Diskursforschung, R. Keller, A. Hirseland, W. Schneider, and W. Viehöver, Eds. VS Verlag für Sozialwissenschaften, 2010, vol. 2, p. 359–385.

[24] H. V. Hansen and D. Walton, "Argument kinds and argument roles in the Ontario provincial election, 2011," Journal of Argumentation in Context, vol. 2, no. 2, p. 226–258, 2013.

[25] C. Reed and G. Rowe, "Araucaria: Software for argument analysis, diagramming and representation," International Journal on Artificial Intelligence Tools, vol. 13, no. 4, p. 983, 2004.

[26] I. Fairclough and N. Fairclough, Political Discourse Analysis. Routledge, 2012.

[27] J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, vol. 20, no. 1, p. 37, 1960.

[28] E. Cabrio and S. Villata, "Five years of argument mining: a data-driven analysis," in IJCAI, 2018, pp. 5427–5433.

[29] J. R. Landis and G. G. Koch, "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers," Biometrics, vol. 33, no. 2, pp. 363–374, 1977.

[30] J. Lawrence and C. Reed, "Combining argument mining techniques," in ArgMining@HLT-NAACL, 2015, pp. 127–136.

[31] M. Lenz, S. Ollinger, P. Sahitaj, and R. Bergmann, "Semantic textual similarity measures for case-based retrieval of argument graphs," in ICCBR, ser. Lecture Notes in Computer Science, K. Bach and C. Marling, Eds., vol. 11680. Springer, 2019, pp. 219–234.

[32] C. Stab, J. Daxenberger, C. Stahlhut, T. Miller, B. Schiller, C. Tauchmann, S. Eger, and I. Gurevych, "ArgumenText: Searching for Arguments in Heterogeneous Sources," in NAACL. Association for Computational Linguistics, 2018, pp. 21–25.

[33] L. Dumani and R. Schenkel, "A systematic comparison of methods for finding good premises for claims," in SIGIR, B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, and F. Scholer, Eds. ACM, 2019, pp. 957–960.

[34] R. Bergmann, M. Lenz, S. Ollinger, and M. Pfister, "Similarity measures for case-based retrieval of natural language argument graphs in argumentation machines," in FLAIRS, R. Barták and K. W. Brawner, Eds. AAAI Press, 2019, pp. 329–334.

[21] https://www.deepl.com/