DARIUS: A Comprehensive Learner Corpus for Argument Mining in German-Language Essays

Nils-Jonathan Schaller[1], Andrea Horbach[2,3], Lars Ingver Höft[1], Yuning Ding[2], Jan Luca Bahr[1], Jennifer Meyer[1], Thorben Jansen[1]
[1]Leibniz Institute for Science and Mathematics Education at the University of Kiel
[2]CATALPA, FernUniversität in Hagen, Germany
[3]Hildesheim University, Germany
schaller@leibniz-ipn.de

Abstract

In this paper, we present the DARIUS (Digital Argumentation Instruction for Science) corpus for argumentation quality on 4589 essays written by 1839 German secondary school students. The corpus is annotated according to a fine-grained annotation scheme, ranging from a broader perspective like content zones, to more granular features like argumentation coverage/reach and argumentative discourse units like claims and warrants. The features have inter-annotator agreements up to 0.83 Krippendorff's α. The corpus and dataset are publicly available for further research in argument mining.

Keywords: corpus, discourse annotation, argument mining

1. Introduction

Scientific argumentation competencies are essential to form well-founded opinions and engage in socio-scientific discourses, such as those on climate change. Therefore, the development of such competencies is a central aim of secondary schooling (NGSS, 2013), and part of upcoming standardized school tests like the Programme for International Student Assessment (OECD, 2023). Students need to learn how to formulate well-structured argumentations (Osborne et al., 2016), including scientifically correct arguments considering multiple options for action and a comprehensible decision strategy (Eggert and Bögeholz, 2009). However, especially high school students struggle to form high-quality arguments (Dawson and Venville, 2009; Eggert and Bögeholz, 2009; Kelly et al., 1998).

Feedback can be a helpful instructional tool to support students' development of argumentation competencies (Ferretti and Graham, 2019). To provide such feedback, it is necessary to assess the structural and content qualities of students' written argumentation, which is very time-consuming for teachers. Automated assessment can support the creation of feedback. However, to develop and test algorithms that allow highlighting these argumentative structures in student writing and providing feedback accordingly, there is a need for creating large and diverse corpora with human annotations from different contexts that include both structural and content elements (Reed et al., 2008; Feng and Hirst, 2011).

Existing corpora (see Stede and Schneider (2019) for an overview) focus mainly on specific out-of-school domains, such as legal documents. Only few corpora contain argumentations written by students, with the Persuasive Essay Corpus (Stab and Gurevych, 2014, 2017) and the Persuade Corpus (Crossley et al., 2022) being the most prominent ones.

The Persuasive Essay Corpus contains 402 essays in English language. The annotations identified supporting claims, and premises, as well as their interrelationships. The Persuade corpus contains over 25,000 essays annotated for structural argumentation elements in English language.

There is a need for more corpora, especially in languages other than English, on student argumentations with high-quality annotations, including both structural and content-related aspects of argumentation quality, to better understand linguistic features of learners' argumentations from different contexts. Detailed annotations of additional large corpora including a large number of student essays would allow us to evaluate the nuanced quality of students' arguments.

We address these needs for further annotated corpora by introducing *DARIUS (Digital Argumentation Instruction for Science)* - a corpus of over 4,500 essays from German secondary school students, focused on two integrated writing prompts on climate change topics. Notably, our annotations include both structural elements and content-related dimensions, along with quality ratings for the arguments. These annotations have shown substantial inter-annotator agreement.

Thus, our corpus substantially broadens the resources available for training and validating machine learning models offering argumentations

LREC-COLING 2024, pages 4356–4367
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

from a diverse sample including students from multiple grades and school types, as well as their demographic data. Further, our corpus holds considerable utility beyond academic research as the basis for automated support systems for both teachers and students.

2. Related Work

Automatic scoring of argumentative texts can be seen as a subtask in automatic essay scoring (Klebanov and Madnani, 2022). However, educational essay datasets not only dealing with argumentative text but also containing detailed annotations about argumentative units are scarce, especially for languages other than English (see also Wang et al., 2022). Furthermore, existing essay datasets target argumentation competencies from a language perspective, but few focus on the measurement of argumentation competencies within a specific subject domain or topic, that is, tasks with a curricular validity beyond language competencies. This is a gap we aim to close with our new dataset. In this section, we first discuss options of how argumentative structures on learner texts can be annotated and then give an overview of other datasets with argumentation annotations from the educational domain.

2.1. Annotating Argumentation

Argumentation annotations aim at making the argumentative structure in texts visible and can be annotated with different levels of granularity.

On an very coarse-grained level, and probably out of the narrower scope of argument mining per se, content zoning aims at segmenting texts into structuring elements such as introduction, main part and conclusion (Stede et al., 2015). Such segmentations have been applied succesfully to scientific papers (Teufel and Moens, 2002; Mizuta and Collier, 2004) and abstracts (Hirohata et al., 2008), but to the best of our knowledge not for argumentative learner essays.

On a more fine-grained level, most argumentation models aim at capturing the argumentative structure of the whole text. One such model that inspired many annotation schemes of argumentative datasets in educational settings is the *microtext scheme* by Peldszus and Stede (2013) that represents argumentative texts as a tree structure. In such an argumentative tree, claims can be supported, i.e. justified, or attacked, i.e. rebutted or undercut, by other discourse units lower in the tree, which themselves can be new claims.

The early but still influential *Toulmin argumentation model* (Toulmin, 1958), in contrast, focuses on the internal structure of individual arguments. In this model, the central part is a *claim*, which is supported by *data*, according to some *warrant*. A warrant can be further supported by a *backing* and restricted in its scope by *qualifiers* or *rebuttals*.

2.2. Datasets

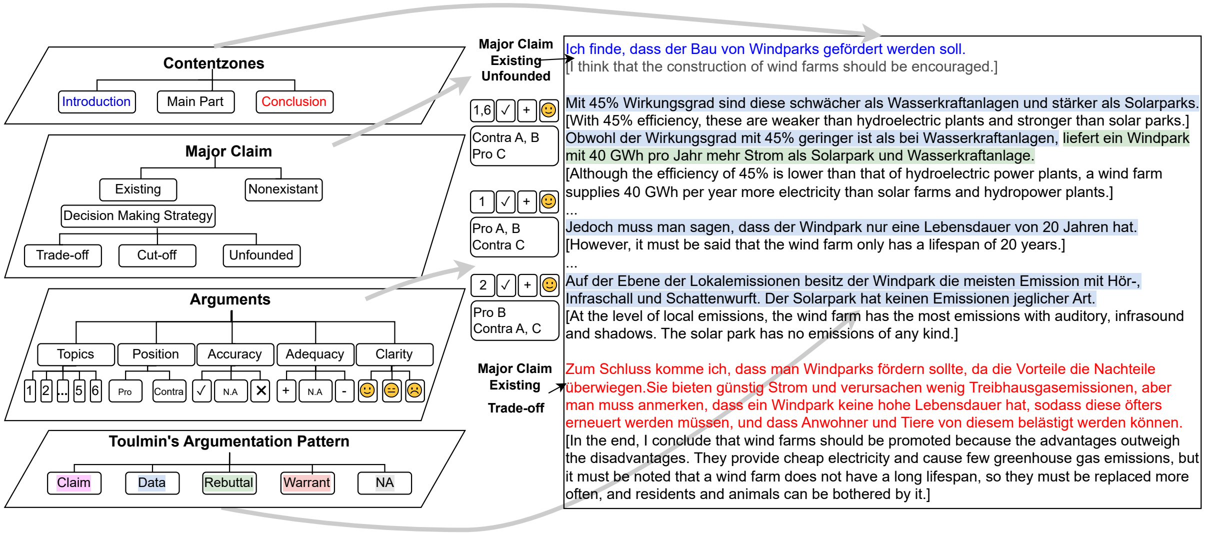
In one of the earliest approaches to annotate learner essays, Stab and Gurevych (2014) use an annotation scheme derived from the microtext scheme, called the *persuasive essay scheme*. Their annotation scheme consists of three explicitly named components: typically one *major claim* per essay and potentially several *claims* supporting or attacking the major claim and *premises*, i.e., reasons that should convince the reader of the validity of a claim (or another premise) or its invalidity if the premise attacks the claim. They originally annotated a set of 90 essays written by users of an online writing platform (presumably both native and non-native). This corpus was later extended (Stab and Gurevych, 2017) to cover more essays from the same online forum source.

This scheme is also followed by Alhindi and Ghosh (2021), who annotated claims and premises in a set of argumentative essays written by school students on the writing platform *WritingMentor*.

Putra et al. (2021) also adopt a similar approach in annotating tree structures of argumentative units but decided against labeling them as claims or premises because they noted that a unit serving as the premise on one level of the tree can be another claim at a lower level. They annotated part of the ICNALE corpus (Ishikawa, 2013) containing essays written by EFL learners.

Another dataset following a similar scheme, and the only German dataset we are aware of, is a set of 1,000 texts written by German university students in order to give each other peer feedback on business models (Wambsganss et al., 2020). They follow an annotation scheme very similar to the one used in Stab and Gurevych (2014) with the main difference that – due to the genre of their data – typically no *major claims* can be found in their texts and are thus not annotated.

The by far largest collection of annotated argumentative essays is the PERSUADE corpus by Crossley et al. (2022) as part of several Kaggle challenges. It contains more than 25,000 argumentative texts annotated by US high school students and has been annotated with an annotation scheme inspired by the Toulmin scheme or rather simplified versions thereof by Nussbaum et al. (2005) and Stapleton and Wu (2015). In this scheme, seven annotation labels are used to describe the argumentative structure of a text: *Lead*, *Position*, *Claim*, *Counterclaim*, *Rebuttal*, *Evidence* and *Concluding Statement*.



Ich finde, dass der Bau von Windparks gefördert werden soll.
[I think that the construction of wind farms should be encouraged.]

Mit 45% Wirkungsgrad sind diese schwächer als Wasserkraftanlagen und stärker als Solarparks.
[With 45% efficiency, these are weaker than hydroelectric plants and stronger than solar parks.]
Obwohl der Wirkungsgrad mit 45% geringer ist als bei Wasserkraftanlagen, liefert ein Windpark mit 40 GWh pro Jahr mehr Strom als Solarpark und Wasserkraftanlage.
[Although the efficiency of 45% is lower than that of hydroelectric power plants, a wind farm supplies 40 GWh per year more electricity than solar farms and hydropower plants.]
...
Jedoch muss man sagen, dass der Windpark nur eine Lebensdauer von 20 Jahren hat.
[However, it must be said that the wind farm only has a lifespan of 20 years.]
...
Auf der Ebene der Lokalemissionen besitz der Windpark die meisten Emission mit Hör-, Infraschall und Schattenwurft. Der Solarpark hat keinen Emissionen jeglicher Art.
[At the level of local emissions, the wind farm has the most emissions with auditory, infrasound and shadows. The solar park has no emissions of any kind.]

Zum Schluss komme ich, dass man Windparks fördern sollte, da die Vorteile die Nachteile überwiegen.Sie bieten günstig Strom und verursachen wenig Treibhausgasemissionen, aber man muss anmerken, dass ein Windpark keine hohe Lebensdauer hat, sodass diese öfters erneuert werden müssen, und dass Anwohner und Tiere von diesem belästigt werden können.
[In the end, I conclude that wind farms should be promoted because the advantages outweigh the disadvantages. They provide cheap electricity and cause few greenhouse gas emissions, but it must be noted that a wind farm does not have a long lifespan, so they must be replaced more often, and residents and animals can be bothered by it.]

Figure 1: Annotation schema and an example essay from the DARIUS corpus.

In our own annotations, we follow different schemes on different granularity levels. Inspired by the persuasive essay scheme, we annotate a *major claim* in each essay together with arguments in favor of or against this claim. I.e., every subtree of a major claim is considered an argument in our annotations without further annotating sub-claims and premises. Instead, we explore the inner structure of an argument following yet another operationalization of Toulmin's argumentation model (Toulmin, 1958), namely that by Riemeier et al. (2012). Out of this model we adopted the annotation of the elements *claim*, *data*, *rebuttal* and *warrant* used to annotate the components of individual arguments.

Table 1 provides an overview of the datasets discussed above and our newly collected DARIUS corpus.

Some further datasets provide trait scores associated with argumentation without explicitly addressing where in the text certain argumentative units appear. This is the case for an extension of the ASAP dataset [1], called *ASAP++* (Mathias and Bhattacharyya, 2018), where the two argumentative prompts are annotated with, among others, organization scores pertaining to the structure of an essay. Similarly, the German SkaLa dataset (Horbach et al., 2017) provides trait scores for argumentation quality.

3. Data

3.1. DARIUS Corpus

The DARIUS dataset initially contains 5,225 argumentative essays, written by 1,839 German high school students across 114 classrooms from 33 schools (see also Table 2 for an overview). It is divided into two writing tasks tackling socio-scientific issues, ”energy” and ”automotive”, with 2,598 and 2,627 texts, respectively. The writing tasks, distributed via the tool LimeSurvey[2], required students to discuss either which power plants or car engines should be supported out of three options presented in the task material. The students were randomly assigned to one of two writing tasks, wrote a first draft, and then received feedback, which was either general feedback or automated feedback based on keywords and various text features. Afterwards, they had to revise their texts based on the feedback they received. Finally, they wrote a text in response to the other writing task as a transfer task.

The corpus contains 1,827 drafts, 1,812 revisions, and 1,586 transfer texts. The lower numbers at each stage are attributed to technical errors or a lack of willingness to continue. 67 texts from a first pilot were additionally annotated. We removed essays that were off-topic, shorter than two sentences, empty, or contained names or other data relevant to data protection, leaving us with a final dataset of 4,589 texts, on which all further analyses took place.

Students also voluntarily provided additional data such as:

- age in years: average = 16.36, standard devi-

[1] https://www.kaggle.com/competitions/asap-aes

[2] https://www.limesurvey.org/

Dataset	Language	# Texts	Learner Population/Text Type	Annotations
Stab and Gurevych (2014)	English	90	essays from online writing forum	PES
Stab and Gurevych (2017)	English	402	essays from online writing forum	PES
Wambsganss et al. (2020)	German	1000	peer feedback by university students	PES
Putra et al. (2021)	English	434	argumentative essays by Asian EFL learners	PES
Alhindi and Ghosh (2021)	English	145	students on the WritingMentor platform	PES
Crossley et al. (2022)	English	> 25.000	US high-school students	TM
DARIUS	German	> 4500	argumentative science essays by secondary students	PES + TM

Table 1: Argument mining datasets with argumentative units annotated in the text. Annotation are categorized whether they adapt the persuasive essay scheme (PES) or the Toulmin model (TM).

ation = 1.47, range = [14, 21], 17.5% missing

- gender: female: 42.6%, male: 35.5%, diverse: 4.4%, 17.5% missing
- type of school and grade (9–13)
- cognitive abilities measured with KFT-R (Heller and Perleth, 2000)
- recent grades in German, Maths, Chemistry, and Physics
- highest degree of parents
- family language

Track	9th	10th	11th	12th	13th
academic	369	304	275	259	27
general	62	89	114	45	37
vocational	121	100	37	-	-

Table 2: Distribution of participating students on school type/grade

3.2. Annotation Scheme

All texts have been annotated on four levels, content zones, major claim, arguments and Toulmin´s Argumentation Pattern, which each investigate different levels of granularity for the overall goal of estimating argumentation quality. We explain the annotation labels in the following. See Figure 1 for an example essay, including the visualized annotation.

3.2.1. Content Zone

In the content zone layer, annotators are asked to label the framing and structuring elements of a text, i.e., the *introduction*, the *main part* and the *conclusion*. Annotations can span one or multiple full sentences. Each sentence can belong to at most one content zone. The annotation follows the guidelines for content areas by Stede et al. (2015). as well as the guidelines for writing argumentative texts in the subject requirements for German in upper high school levels (Fac, 2014).

Labels: *introduction*, *main part*, *conclusion*

3.2.2. Major Claim

Major claim refers to the author's final position on the given topic, upon which the author bases their decision. Our definition differs from the one by Stab and Gurevych (2014) in that the major claim may not only occur in the introduction but also in the conclusion of an essay. This label is annotated on the sentence-level.

A major claim can but does not have to be marked as such linguistically by using reporting verbs like *I think*, or *I believe* and equally can, but does not have to be introduced by discourse connectives such as *For these reasons*.

Example 1 shows an instance of a major claim.

(1) *Aus diesen Gründen denke ich, dass Wasserkraftwerke gezielt gefördert werden sollten.* (*For these reasons, I believe that hydroelectric power plants should be specifically promoted.*)

Labels: *major claim*

Each major claim is further annotated with the Decision-Making-Strategy.

Decision-Making Strategy

The Decision-Making-Strategy indicates which strategy is taken to justify the major claim. We follow Eggert and Bögeholz (2009) in our definition of possible strategies: We define *trade-off* as a compensatory consideration of all arguments that weighs off supporting and opposing arguments to come to a decision, whereas *cut-off* denotes a decision based on a single argument, considered important by the author while disregarding all other arguments. *unfounded* is annotated if no strategy can be identified. Example 2 shows a cut-off and Example 3 the more deliberative trade-off.

(2) *Wasserkraftanlagen sind die deutlich bessere Wahl, da sie in Wirkungskraft überwiegen.* (*Hydropower plants are the significantly better choice as they outweigh in terms of effectiveness.*)

(3) *Auch wenn der Preis deutlich höher ist, denke ich, dass der Wirkungsgrad der erneubaren Energien um einiges bedeutsamer ist, somit sind die Wasserkraftanlagen am Besten einzusetzen in Norddeutschland.*
(*Even if the price is significantly higher, I think the efficiency of renewable energies is much more significant, therefore hydroelectric power plants are best used in Northern Germany.*)

Labels: *trade-off*, *cut-off*, *unfounded*

3.2.3. Argument

Arguments should consist of at least one complete sentence, although they can span multiple sentences We define the arguments by the topics, described in the following paragraph, e.g., marking a sequence that discusses one or more topics. An argument can be marked for multiple topics at once.

The beginning of a new sentence, accompanied by a change in topic, signifies the start of a new argument. For each argument, annotators additionally determine five features, namely **Topic**, **Position**, **Accuracy**, **Adequacy**, and **Clarity**.

The following argument, Example 4, serves as an illustration for the five argument features which we discuss in more detail afterwards.

(4) *Als erstes Kriterium betrachten wir den Gesamtwirkungsgrad. Dabei geht es darum, dass elektrische Energie während des Autosfahrens an die Umwelt abgegeben wird. Die Akkumulatoren sind diesbezüglich sehr geeignet, denn sie haben einen Gesamtwirkungsgrad von 70% und im Vergleich zu den anderen Alternativen ist das der höchste. Die E-Fuels schneidet beim Gesamtwirkungsgrad am schlechtesten ab, das heißt es muss bei E-Fuels am meisten elektrische Energie in die Umwelt abgegeben werden.*
(*The first criterion we consider is the overall efficiency. This refers to the fact that electrical energy is released into the environment while driving a car. In this regard, accumulators are very suitable as they have an overall efficiency of 70% and compared to the other alternatives, this is the highest. The E-Fuels perform worst in terms of overall efficiency, which means that the most electrical energy has to be released into the environment with E-Fuels.*)

Topic

The two tasks contain information on six topics each, meant to guide the student in constructing an argumentation. Arguments are annotated based on the specific topic they address, with those veering off-topic marked as *other*. For example, an argument can discuss the topics of efficiency vs cost of a powerplant. It is possible that multiple topics are marked in an argument. We display only the labels of the automotive task here. Example 4 would be labeled with *efficiency*.

Labels: *operation*, *noise emissions*, *energy density*, *availability*, *greenhouse gas emissions* , *efficiency*, *other*

Position

The position reveals if the argument supports or opposes a specific position out of the three positions provided per task, such as the use of cars powered by e-fuels or batteries. The annotation includes general details about the argument's direction and task-specific information regarding which position it supports or refutes. If no clear preference is evident, the argument is labeled as *unclear*. An argument can support or oppose multiple positions at once, e.g., be supporting hydrogen while refuting e-fuels and electric cars.

The argument in example 4 would be labeled as *Pro C* and simultaneously *Contra B* as it supports battery electric vehicles while opposing e-fuels. The individual positions corresponding to A, B and C change for the two tasks.

Labels: *Pro A/B/C*, *Contra A/B/C*, *unclear*

Accuracy

Accuracy, as defined by Heitmann et al. (2014), indicates whether an argument is scientifically correct based on the material provided to students. It reflects students' comprehension and correct use of the material in their argumentation. Arguments unrelated to the material are marked as *not applicable*.

The argument in Example 4 refers to the data from the material and would be labeled as *correct*.

Labels: *correct*, *false*, *not applicable*

Adequacy

Adequacy is also conceptualized following the definition by Heitmann et al. (2014), indicating if a student's argument is relevant to the assignment. We developed it further and differentiated between three labels.

If the argument is relevant to the task, it is marked as *adequate*. If it is contextually relevant but leads to an incorrect or illogical conclusion, it is labeled as *inadequate*. Arguments unrelated to the material are marked as *not applicable*.

The argument in Example 4 is labeled as *adequate*, as it makes logical assumptions and follows the assignment.

Labels: *adequate*, *inadequate*, *not applicable*

Clarity
Clarity denotes the argument's understandability. If the annotator understands the argument immediately, it is marked as *understandable*. If it is necessary to read it more than twice, it is marked as *difficult to understand*. If the argument seems illogical or is barely readable due to poor orthography or grammar, it is marked as *unintelligible*. We found this differentiation important, as many students would already profit by simply rewriting or restructuring their arguments.

Example 4 would be annotated as *understandable*.

Labels: *understandable*, *difficult to understand*, *unintelligible*

3.2.4. Toulmin's Argumentation Pattern

Toulmin's Argumentation Pattern (TAP) describes a structural framework for constructing logical and compelling arguments by including a claim, providing supporting evidence (data), explaining the connection between the claim and data (warrant), reinforcing the warrant (backing), expressing the claim's limitations (qualifier), and addressing counterarguments (rebuttal). We based our interpretation of each element on a paper by Riemeier et al. (2012). We annotate only *claim*, *data*, *warrant*, and *rebuttal* as we found insufficient amounts of other TAP-elements while inspecting a data sample before the annotation. The annotation of TAP is the only one made on token-level. A sentence can contain multiple elements, e.g., a claim and a data, separated into phrases or divided by conjunctions.

- *claim*: Assertion that characterizes the position taken. *"Hydrogen is the best option, ..."*
- *data*: Fact that provides the basis for a claim: *"...because accumulators have too long a charge time..."*. It substantiates or explains a claim.
- *warrant*: Aspect that explains to what extent a data supports a claim: *"...and are therefore much more impractical in everyday life..."*.
- *rebuttal*: An objection to a presented data and/or warrant: *"Although the price is a counter-argument, if you calculate the price for a wind farm or solar park over its lifetime, you almost come out the same."*
- *not applicable*: the sequence cannot be recognized as part of an argument or does not fulfill the purpose of claim/data etc. Example: *"Let's move on to the next criterion."*

Labels: *claim*, *data*, *warrant*, *rebuttal*, *not applicable*

4. Annotation Process

4.1. Annotator Training

Ten annotators and one super-annotator participated in the annotation and curation process. The super-annotator took part in creating and testing the annotation scheme. Her responsibilities were co-training the other annotators as well as curating the annotations. Each annotator underwent training for approximately 40 hours, including reading the 29-page annotation manual and training on a sample of 30 texts. Every annotator completed a test set of an additional 30 texts to measure their inter-annotator agreement with the super-annotator and research team before working on the entire corpus. This was done to ensure a high agreement score and understanding of the manual. Each text was annotated by two randomly assigned annotators. Inconsistencies were resolved by the super-annotator who received anonymized annotations to avoid bias.

All annotations and curations were created with the freely available INCEpTION tool (Klie et al., 2018) and are included in the corpus release as tsv-files.

4.2. Inter-rater reliability

To meaningfully assess inter-rater reliability, we distinguished between annotations at the sentence-level (see sections 3.2.1 through 3.2.3) and token-based annotation of TAP elements (see section 3.2.4). At the sentence-level, annotators only had to code predefined units, i.e., they had to decide for every sentence which categories it belongs to. For token-based annotations, the annotators had to simultaneously unitize the continuous text into segments and code these segments, i.e., identify the boundaries of the TAP elements and assign the corresponding value. We evaluate the individual annotations by using a variety of reliability metrics.

Annotations at the sentence-level. For the sentence-level evaluation, we report **percentage agreement**, i.e., the relative number of sentences for which both annotators agreed on the presence or absence of a certain label, as an intuitively understandable measure of agreement. We further report the chance-corrected measure **Krippendorff's** $_{c}\alpha$ (Krippendorff, 1980) for codings of predefined units. We interpreted sentences without annotations as if the corresponding layer had been annotated with a *None*/*not applicable* label. Regarding the chance-corrected measures, we considered that ratings on the layer major claim, decision-making strategy, accuracy, adequacy and clarity are based on ordinal scales con-

	All sentences				Agreement-only setting				
Layer	**PA**	**F1$_{100\%}$**	**F1$_{50\%}$**	$_c\alpha$	**n**	**PA**	**F1$_{100\%}$**	**F1$_{50\%}$**	$_c\alpha$
Content zone	0.93	0.88	0.93	0.83					
Major claim	0.97	0.98	0.98	0.83					
					Major claims in introduction or conclusion				
Decision-making strategy	0.98	0.70	0.74	0.76	0.04	0.95	0.55	0.69	0.65
					Main part				
Topics	> 0.92	> 0.87	> 0.90	> 0.77	0.70	> 0.92	> 0.89	> 0.92	> 0.80
Position	> 0.76	> 0.70	> 0.69	> 0.68	0.70	> 0.73	> 0.70	> 0.70	> 0.66
Accuracy	0.90	0.61	0.62	0.74	0.70	0.92	0.44	0.49	0.43
Adequacy	0.90	0.62	0.62	0.74	0.70	0.93	0.44	0.49	0.45
Clarity	0.91	0.54	0.57	0.76	0.70	0.96	0.36	0.39	0.19

Table 3: Inter-Annotator Agreement coefficients for sentence-wise annotated layers. As Topics and Position comprise multiple, independently annotated labels, only the lowest inter-annotator agreement values were listed. *Note.* n = Percentage of sentences taken into account, PA = Percentage agreement, F1 = F1 Score, α = Krippendorff's $_c\alpha$.

sisting of ordered categories with unknown distances by applying the according ordinal weights (see Tran et al., 2018, Eq. 19 to 20).

To allow comparability with previous work, especially the PERSUADE corpus, we also report F1 scores. Adopting the evaluation scheme proposed by Crossley et al. (2022), a prediction is considered a true positive if the overlap of tokens between both annotators is greater than 50% in both directions (F1$_{50\%}$). Any unmatched token annotated by annotator 1 is a false negative, and any unmatched token annotated by annotator 2 is a false positive. These three values were used as the basis for computing the F1 score. We calculated the F1 Score for both directions, i.e., switched the roles' of annotator 1 and 2, and afterward averaged the F1 Scores.

We calculated the reliability coefficients based on the complete corpus, i.e., all sentences. Some elements had to be annotated depending on the prior annotation of a different layer: arguments could only occur within the main part of an essay, a decision-making strategy can only be annotated for a specific major claim, and properties of an argument depend on the previous annotation of an argument. To account for these dependencies, we also evaluate these annotations in a so-called *agreement-only* setting, where we only considered cases where both annotators agreed on the superordinate annotation, e.g., we considered only decision-making strategies where both annotators identified the same major claim.

The inter-rater reliability coefficients calculated on the basis of all sentences suggest at least a substantial agreement for all layers (percentage agreement = [.76, .98], F1$_{100\%}$ = [.54, .98], F1$_{50\%}$ = [.57, .98], $_c\alpha$ = [.68, .83], see Table 3), only ratings of the layer clarity (F1$_{50\%}$ = .54) obtained lower reliability values (Artstein and Poesio, 2008). This is not surprising, given the high prevalence of the "None" label for many layers (e.g., only 7% of all sentences were annotated as "Major claim" by at least one annotator and hence 93% of all sentences were handled as having no decision-making strategy and given the "None" label) and the therefore imbalanced datasets (Chicco and Jurman, 2020). However, we consider this evaluation setup useful as it provides a human upper bound for a later automation of the identification of argumentative elements that might operate and be evaluated on the sentence-level.

The inter-rater reliability coefficients calculated in the *agreement-only* setting instead, might show a more realistic picture of the actual agreement of human annotators and indicate a low reliability for the layers accuracy, adequacy, and clarity (see right part of Table 3).

Token-level. We evaluated the reliability values for the token-based annotation of TAP elements differently from the other elements by estimating Krippendorff's four $_u\alpha$ coefficients (Krippendorff et al., 2015) which simultaneously account for the process of freely annotated units of various sizes (unitizing) and categories (categorizing). This is necessary, as in the sentence-level, the size of the segments to be annotated is predetermined, whereas with the TAP elements, the size of the chunks is not defined.

However, the annotators were not able to unitize and code the relevant segments reliably, as the obtained $_u\alpha$ coefficients were below .50. Closer examination of the reliability coefficients for each category of TAP (see Table 5) reveals that the inadequate reliability values could identify only the TAP element *data* satisfactorily. We see that an-

TAP	Miss.	Claim	Data	Warr.	Rebutt.
-	-	59.72	28.06	3.76	3.77
Claim	10.73	62.90	17.03	5.68	3.67
Data	2.94	10.44	82.30	1.92	2.41
Warrant	3.78	34.75	19.16	37.81	4.49
Rebuttal	3.39	20.21	21.68	4.04	50.69

Table 4: Inter-Annotator agreements on TAP elements. Each row adds up to 100%, e.g., Claims were coded by the respective other annotators as 11% missing, 3% unclear, 63% claim, 17% data, 6% warrant, 4% rebuttal. Matches on missing entries were removed.

TAP Element	F1$_{100\%}$	F1$_{50\%}$	$_{(k)u}\alpha$
Claim	0.41	0.61	0.46
Data	0.46	0.71	0.68
Warrant	0.30	0.42	0.08
Rebuttal	0.47	0.58	0.43

Table 5: Inter-Annotator Agreement coefficients of the unitizing textual continua for TAP elements. All sentences of the main part were taken into account (70% of all sentences/ 78% of all tokens).

notators struggled most with discriminating ”Warrants” from ”Claims” indicated by low reliability coefficients ($_{(k)u}\alpha$ = .08). In fact, 34.8% of all ”Warrants” were coded by the annotator 2 as ”Claim” and only 37.8% of all warrants could be reliably annotated (see Table 4).

5. Annotation Analysis

Annotation issues. Table 5 shows the agreements for TAP are comparably low, with only the *data* label getting an agreement of α = .68. Especially, the *warrant* label seems very difficult to annotate with an agreement of only α = .08. We tried to find explanations by analyzing the data, as well as seeking information directly from the annotators. During the annotation process and biweekly discussions, we already learned that annotators struggled with differentiating certain TAP labels, e.g., agreeing if a structure is a *claim* or a *warrant* as well as identifying the *rebuttal* in general.

This seemed to be a problem with texts that were perceived by the annotators as lacking structure and having a weak argumentation or many problems with orthography, whereas they agreed on elements in well-structured texts. Moreover, distinguishing *data* from *claim* was often difficult as some sentences met both definitions. An example is the sentence “The electric car is with a power efficiency of 70% clearly the best”, which arguably refers to data from the material (70%) and gives at the same time an opinion (the best). The issue is further quantified in Table 4: 34.75% of sequences that have been labeled as *warrant* from one annotator, were annotated as *claim* by the second annotator. Conversely, only 5.68% of *claim* labels were labeled as *warrant* by the second annotator, indicating that the problem lies more in agreeing on what a *warrant* is, rather than a *claim*.

Similar problems occurred with the label *rebuttal*, which occurred in only 44% of texts and was often labeled differently by second annotators either as *claim* or *data*, both with around 21%.

What we can conclude from this is that the definitions of TAP, especially the more complex elements, seem not to work on the student’s essays. We have to further investigate if this is a definition problem or additionally a problem with the ability of students to use these elements correctly in an argumentation.

Label Distributions. Analyzing the curated data provides insight into label distributions and areas students struggle with in writing, seen in the absence of elements like introductions or trade-offs. As seen in Table 7 60% of all texts include an introduction and even less (39%) a conclusion. The difference in the latter could be in part due to the time limit under which the students had to finish the text. Also, 23% of texts lack a major claim, and of those who have one, only 12% use the most complex strategy for writing a major claim, a trade-off, whereas 38% of major claims are unfounded. Every text included at least one argument and 81% at least three arguments per text, indicating that most students followed the assignment by discussing at least three topics. As seen in Table 8, only 33% of texts include a *warrant* and 44% a *rebuttal*. This suggests, that many students can profit from getting feedback on their writing on all levels of argumentation, starting from information, which parts of the basic structure are still missing, if enough arguments were written, and where the text can gain from more complex information, etc. On a technical level, it directly shows, that the labels for each task are imbalanced, which has to be considered when using the data for machine learning.

6. Conclusion & Future Work

Our work presents DARIUS, a rigorously annotated corpus of argumentative texts from a diverse sample of secondary school students, including students from multiple grades and school types. The corpus complements the literature, especially in that the texts are written in the German language, the writing tasks are authentic to school classrooms, and we annotated the quality of both content and arguments.

Topic	Energy		Autom.	
Yield/Energy	1427	0.62	823	0.36
Lifetime/Availability	1912	0.83	1545	0.68
Emissions/Noise	1146	0.50	932	0.41
Price/Operation	1890	0.82	1686	0.74
Greenhouse Gas	1210	0.52	1338	0.59
Efficiency	1794	0.78	1302	0.57
Other	818	0.35	864	0.38

Table 6: Topics ocurring within the essay for the two tasks ”Energy” and ”Automotive” (absolute and relative frequencies)

Annotation	Occurance	Proportion
Content Zone		
Introduction	2765	0.60
Main Part	4550	0.99
Conclusion	1807	0.39
Major Claim		
Texts with at least 1 MC	3542	0.77
DMS		
Unsubstantiated	1648	0.38
Cutoff	2195	0.50
Trade-Off	520	0.12

Table 7: Absolute and relative frequency of the annotation for *Content Zone*, *Major Claims (MC)* and *Decision-Making Strategy (DMS)* among all essays in the corpus.

The implications are twofold. First, DARIUS enriches the landscape of available corpora, facilitating the development of more accurate machine learning models for argument mining. Second, it holds the potential for educational impact by serving as the foundation for automated support systems that can assist teachers and students alike in the demanding task of fostering argumentation skills with feedback.

To evaluate this potential, we plan training machine learning models on the annotations to give students automated feedback on their essay drafts. We plan to evaluate how the students perceive such feedback and if it results in a higher quality of revised essays, compared to the original dataset. Furthermore, we are planning to assess the fairness of various algorithms trained on the annotations in regard to the demographic data collected on the students.

TAP	avg. # tokens	text w. TAP elem.	prop.
Claim	19.34	4271	0.93
Data	29.32	4385	0.96
Warrant	16.83	1503	0.33
Rebuttal	13.79	2011	0.44

Table 8: Average length per TAP element in tokens and numbers and percentage of text containing at least one instance of that element

7. Ethical Considerations and Limitations

Our dataset contains essays written by under-age students in the German school system as part of a 90-minute school lecture. We informed both the students and their parents about the collection of their written texts as well as any voluntarily given data such as age, gender, grade and obtained their written consent to use and publish the data for research purposes. They could opt out of having their data stored, which led to the exclusion of 162 texts.

Due to the nature of the writing tasks, we did not expect that the texts contained personal information. Nevertheless, we carefully checked each essay and removed any text from the final corpus that contained any information that would make the student personally identifiable, such as names or addresses.

Another limitation concerns potential biases within the dataset. The data from the German federal state of Schleswig-Holstein might not fully represent Germany, particularly regarding the coverage of minority groups. This situation could lead to models that disadvantage these groups.

8. Acknowledgements

This work was supported by the Deutsche Telekom Stiftung and partially conducted at “CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics” of the FernUniversität in Hagen, Germany. We thank Mira Tanz for her support in annotating and curating the texts.

9. Bibliographical References

2014. *Fachanforderungen Deutsch. Sekundarstufe I/II.* Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein.

Tariq Alhindi and Debanjan Ghosh. 2021. ” sharks are not the threat humans are”: Argument component segmentation in school student essays. *arXiv preprint arXiv:2103.04518*.

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

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1).

Scott A Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (persuade) corpus 1.0. *Assessing Writing*, 54:100667.

Vaille Dawson and Grady Venville. 2009. High-school students' informal reasoning and argumentation about biotechnology: An indicator of scientific literacy? *International Journal of Science Education - INT J SCI EDUC*, 31:1421–1445.

Sabina Eggert and Susanne Bögeholz. 2009. Students' use of decision-making strategies with regard to socioscientific issues: An application of the rasch partial credit model. *Science Education*, 94:230 – 258.

Vanessa Wei Feng and Graeme Hirst. 2011. 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. Association for Computational Linguistics.

Ralph P. Ferretti and Steve Graham. 2019. Argumentative writing: theory, assessment, and instruction. *Reading and Writing*, 32(6):1345–1357.

Patricia Heitmann, Martin Hecht, Julia Schwanewedel, and Stefan Schipolowski. 2014. Students' argumentative writing skills in science and first-language education: Commonalities and differences. *International Journal of Science Education*, 36:3148–3170.

Kurt Heller and Christoph Perleth. 2000. *Kognitiver Fähigkeitstest für 4.-12. Klassen, Revision (KFT 4-12+ R)*.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Andrea Horbach, Dirk Scholten-Akoun, Yuning Ding, and Torsten Zesch. 2017. Fine-grained essay scoring of a complex writing task for native speakers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 357–366, Copenhagen, Denmark. Association for Computational Linguistics.

Shin'ichiro Ishikawa. 2013. The icnale and sophisticated contrastive interlanguage analysis of asian learners of english. *Learner corpus studies in Asia and the world*, 1(1):91–118.

Gregory Kelly, Stephen Druker, and Catherine Chen. 1998. Students' reasoning about electricity: Combining performance assessments with argumentation analysis. *International Journal of Science Education - INT J SCI EDUC*, 20:849–871.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated Essay Scoring*. Springer Nature.

Klaus Krippendorff. 1980. *Content analysis*. Sage Publications.

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2015. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, 50(6):2347–2364.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Yoko Mizuta and Nigel Collier. 2004. Zone identification in biology articles as a basis for information extraction. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 29–35.

NGSS. 2013. *Next Generation Science Standards: For States, By States*. The National Academies Press, Washington, DC.

E Michael Nussbaum, CarolAnne M Kardash, and Steve Ed Graham. 2005. The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of educational psychology*, 97(2):157.

OECD. 2023. *PISA 2025 SCIENCE FRAMEWORK (DRAFT)*. OECD Publishing, Paris.

Jonathan Osborne, Bryan Henderson, Anna Macpherson, Evan Szu, Andrew Wild, and Shi-Ying Yao. 2016. The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, 53:n/a–n/a.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga. 2021. Parsing argumentative structure in English-as-foreign-language essays. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 97–109, Online. Association for Computational Linguistics.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Tanja Riemeier, Claudia Aufschnaiter, Jan Fleischhauer, and Christian Rogge. 2012. Argumentationen von schülern prozessbasiert analysieren: Ansatz, vorgehen, befunde und implikationen. *Zeitschrift für Didaktik der Naturwissenschaften*, 18:141–180.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Paul Stapleton and Yanming Amy Wu. 2015. Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance. *Journal of English for Academic Purposes*, 17:12–23.

Manfred Stede, Sara Mamprin, Andreas Peldszus, André Herzog, David Kaupat, Christian Chiarcos, and Saskia Warzecha. 2015. *Handbuch Textannotation*. Potsdamer Kommentarkorpus 2.0. Universitätsverlag Potsdam.

Manfred Stede and Jodi Schneider. 2019. *Argumentation Mining*, 1 edition. Synthesis Lectures on Human Language Technologies. Springer Cham. EBook Packages: Synthesis Collection of Technology (R0), eBColl Synthesis Collection 8.

Simone Teufel and Marc Moens. 2002. Articles summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Stephen E Toulmin. 1958. *The uses of argument*. Cambridge university press.

Duyet Tran, Anil Dolgun, and Haydar Demirhan. 2018. Weighted inter-rater agreement measures for ordinal outcomes. *Communications in Statistics - Simulation and Computation*, 49(4):989–1003.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. A corpus for argumentative writing support in German. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 856–869, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xinyu Wang, Yohan Lee, and Juneyoung Park. 2022. Automated evaluation for student argumentative writing: A survey. *arXiv preprint arXiv:2205.04083*.

10. Language Resource References

Alhindi, Tariq and Ghosh, Debanjan. 2021. *" Sharks are not the threat humans are": Argument Component Segmentation in School Student Essays*.

Crossley, Scott A and Baffour, Perpetual and Tian, Yu and Picou, Aigner and Benner, Meg and Boser, Ulrich. 2022. *The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0*. Elsevier.

Jan-Christoph Klie and Michael Bugert and Beto Boullosa and Richard Eckart de Castilho and Iryna Gurevych. 2018. *The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation*. Association for Computational Linguistics. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).

Putra, Jan Wira Gotama and Teufel, Simone and Tokunaga, Takenobu. 2021. *Parsing Argumentative Structure in English-as-Foreign-Language Essays*. Association for Computational Linguistics.

Stab, Christian and Gurevych, Iryna. 2014. *Identifying argumentative discourse structures in persuasive essays*.

Stab, Christian and Gurevych, Iryna. 2017. *Parsing Argumentation Structures in Persuasive Essays*. MIT Press.

Wambsganss, Thiemo and Niklaus, Christina and Söllner, Matthias and Handschuh, Siegfried and Leimeister, Jan Marco. 2020. *A Corpus for Argumentative Writing Support in German*. International Committee on Computational Linguistics.