Computational Linguistics in the Netherlands Journal 11 (2021) 257-279 Submitted 10/2021; Published 12/2021

Annotation of a Dutch Essay Corpus with Argument Structures and Quality Indicators

Liqin Zhang[*] LIQIN.ZHANG@OU.NL
Howard Spoelstra[*] HOWARD.SPOELSTRA@OU.NL
Marco Kalz[*][**] KALZ@PH-HEIDELBERG.DE

[*] *Open University of the Netherlands, Valkenburgerweg 177, 6419 AT Heerlen, the Netherlands*

[**] *Heidelberg University of Education, Keplerstraße 87, D-69120 Heidelberg, Germany*

Abstract

Based on the availability of previously annotated text corpora, the technique of argument mining (AM) aims to discover components in texts belonging to an argumentation structure. Due to the lack of such annotated corpus for the Dutch language, this paper presents a Dutch essay corpus with annotations of argumentation structures and quality indicators. We applied the annotation schemes and guidelines derived from previous studies to capture the argument structures of Dutch argumentative essays by identifying and classifying the argument components into major claims, claims, and premises as well as the support/attack relations between the components. Furthermore, we annotated persuasiveness scores and attributes that influence persuasiveness as quality indicators. The annotation task was performed by four native Dutch teachers who annotated 30 student-written Dutch argumentative essays. The inter-rater agreement of the annotations was generally lower compared to similar previous work, due to the less rigid format of the essays in our corpus and more annotators participating in the annotation task. However, the essays in our corpus are more in line with non-worked real-world text examples. To ensure the accuracy, objectivity, and reliability of the corpus a consolidation procedure was applied to the final compilation. This corpus presents a novel and reliable resource for future applications in argument mining tasks in the Dutch context. The corpus is publicly available via GitHub[1].

1. Introduction

Argument mining (AM) is a relatively new area in Natural Language Processing (NLP). Various studies reported on its development and application (Lawrence and Reed 2020, Lippi and Torroni 2016, Palau and Moens 2009). While aiming to automatically find and analyse the argument structures in texts, AM has been applied in several fields, such as education, politics, and social media analysis (Lippi and Torroni 2016). AM has two main goals:

- To identify argument components and their types.
- To identify the relationships between these components.

Several previous studies presented approaches to analyse argumentative essays (Stab and Gurevych 2017, Wambsganss et al. 2020). Corpora with annotations of argumentation components are required to develop NLP models for argumentation analysis. There are a handful of corpus studies that aim to analyse and evaluate the argumentation in English essays (Stab and Gurevych 2014) as well as some other languages such as German (Wambsganss et al. 2020) and Portuguese (Rocha and Cardoso 2017). These corpora contain the annotations of argument component boundaries, argument component types, and the relations between the components. A recent addition is to include the annotation of attributes describing the argumentation quality, such as the persuasiveness value

[1] https://github.com/jayliqinzhang/A-Dutch-essay-corpus-with-argument-structures-and-quality-indicators

of an argument (Carlile et al. 2018) and argumentation-related attributes (Gao et al. 2019, Ke et al. 2019). However, such corpora are not available for the Dutch language.

In earlier work (Zhang et al. 2020), we trained a model with an automatically generated Dutch corpus based on the method proposed by Eger et al. (2018) (See Section 2.4 for the detail of this method). The model is implemented as a deep learning network as in Eger et al. (2017) and is capable of identifying the argument structures in Dutch essays. However, the lack of suitable, more fully annotated Dutch corpora restricts the further development of automatic argument analysis of Dutch essays. Therefore, we report on creating and assessing the annotation quality of a Dutch corpus with annotated argument components, relations, and argument component quality indicators in this article. We recruited four native Dutch speakers to participate in an annotation task to annotate 30 Dutch essays. The main contribution of our study is to provide the first Dutch essay corpus for argumentation analysis.

2. Related work

In this section, we explore related work regarding the annotation of arguments and annotation schemes. We also introduce and explore related work about annotating argument quality and annotated corpora in non-English contexts.

2.1 Annotation of argumentation and annotation schemes

Stab and Gurevych (2014) proposed an annotation scheme to annotate the argument structures of argumentative essays. In their scheme, an argumentative essay exhibited a common structure and the aim of the scheme was to capture the argumentative discourse structure by identifying argument components and classify their types into major claims, claims and premises. Additionally, they defined relations between components: a component "supports" or "attacks" another component. Using this scheme, the authors compiled an annotated corpus using 90 essays from Essayforum[2] and expanded it to 402 essays later on (Stab and Gurevych 2017). Other researchers adopted this method to compile corpora from essays in different languages, such as Portuguese (Rocha and Cardoso 2017), Russian (Fishcheva and Kotelnikov 2019) and German (Wambsganss et al. 2020).

Another scheme was used by Fisas et al. (2016) and Lauscher et al. (2018) to compile a corpus of scientific papers called the Dr Inventor corpus. For this corpus, the scheme consisted of assigning a tag to every sentence in terms of a rhetorical category (such as challenge, background and approach), citation purpose (criticism and comparison), scientific discourse (disadvantage and advantage), and argument components (own claim, background claim, and data component). Visser et al. (2021) annotated the reasoning patterns in the corpus of the televised election in the U.S. from 2016. Their annotation scheme was designed based on the taxonomy from Walton et al. (2008) and the periodic table of arguments from Wagemans (2016). The taxonomy by Walton et al. (2008) was an empirical classification based on conventional argument practice. There were 60 argumentation schemes types in the taxonomy, and these types were classified as reasoning, source-based arguments, and applying rules to cases. In the periodic table from Wagemans (2016), the scheme types were classified as first/second-order, predicate/subject, and propositions of fact, value, and policy.

A relevant work in the Dutch context was from Van Der Vliet et al. (2011) and Redeker et al. (2012). The authors compiled a corpus containing persuasive Dutch texts of fundraising letters and commercial advertisements with the annotation of discourse structures based on the classic rhetorical structure theory (RST) analysis (Mann and Thompson 1988).

[2] https://essayforum.com/

2.2 Annotation of argumentation quality

Recently researchers started to include quality indicators of argument components in their models. Some focused on scoring the quality of the arguments of various topics in large-scale settings by ranking the arguments within a subject (Gretz et al. 2020, Habernal and Gurevych 2016, Toledo et al. 2019, Toledo-Ronen et al. 2020). The arguments stood for themselves without context in the argument-ranking dataset, and they were written to argue for a given topic. Several corpora were compiled for writing-support occasions. On top of the Stab and Gurevych (2014) corpus, Carlile et al. (2018) annotated the quality of the annotated argument components. Specifically, the persuasiveness scores of the identified arguments components were annotated along with the attributes that describe the persuasiveness, such as specificity, eloquence, evidence etc. Apart from persuasiveness, Ke et al. (2019) selected essays from the International Corpus of Learner English (ICLE) corpus (Granger et al. 2009) to evaluate the strengths of the statements after summarising the arguments in an essay. The researchers designed a rubric containing attributes to measure the arguability, specificity, clarity etc. Being inspired by the argument scoring attributes of Gallagher et al. (2015) and Ferretti and Lewis (2019), Gao et al. (2019) recently designed a rubric for annotating the argument quality on student essays. The researchers conducted a pilot annotation task to annotate the essay content into summary content units, elementary discourse units, and their alignments. The argument quality within the above corpora was practical-based, which meant the argumentation qualities were evaluated holistically (the overall quality) or analytically (the attributes relevant to the argumentation). Other studies attempted to assess the quality based on a taxonomy extracted from more fine-grained argumentation theories (Lauscher et al. 2020, Wachsmuth et al. 2017).

2.3 Corpora in other languages

Based on the argumentation theory of Freeman (2011), Peldszus and Stede (2015) annotated the argument structures in a collection of more than 100 short German texts, written as a response to trigger questions, with standpoint and justification. Rocha and Cardoso (2017) annotated the argumentative discourse units identified in Portuguese opinion articles from a news article collection, while Li et al. (2017) annotated the argument components in Chinese hotel reviews based on the annotation scheme of Stab and Gurevych (2014). Iida et al. (2007) annotated the predicate-argument relations in Japanese texts from the Kyoto Text Corpus and the GDA-Tagged Corpus. Regarding quality annotation of corpora, we could not find similar work in languages other than English.

2.4 Alternative corpus generation method

Considering the heavy workload to manually create annotated corpora for various languages, Eger et al. (2018) and Rocha et al. (2018) proposed to apply machine translation and a tag-projection algorithm to transfer the available corpora in a language (e.g. English) to other languages automatically. Recently Toledo-Ronen et al. (2020) used this approach to create multilingual datasets. The datasets were applicable for the argument mining tasks of stance classification, evidence detection and argument quality assessment for multilingual context. We also applied this method to generate a Dutch essay corpus to develop a model for argument component identification, on which we reported in Zhang et al. (2020).

3. Data source

The CSI corpus is one of the few publicly available Dutch essay corpora (Verhoeven and Daelemans 2014). It contains 209 essays with an average length of 1126 words. They are written by native Dutch speakers who took Dutch proficiency courses at the University of Antwerp. The original corpus is compiled for author profiling, containing the meta-data of the authors, including age, gender, region of origin, personality, and sexual orientation. Besides that, grades are given to the essays indicating their quality, although the scoring criterion is not mentioned.

When selecting our source essays, our sampling strategy was to choose the essays whose length were similar to the annotated essay corpus of Stab and Gurevych (2014) and Wambsganss et al. (2020). Moreover, the essays with higher grades were preferred as the holistic quality of an essay was correlated to its argumentation features, such as the number of arguments in an essay (Ghosh et al. 2016). As a result, we decided to select 30 essays shorter than 700 words from the CSI corpus, considering the size of the task. The selected texts covered several subjects as presented in Table 1.

While the essays used in the corpus of Stab and Gurevych (2014) are in a "five-paragraph"[3] format as described in the Purdue Online Writing Lab (Purdue 2021), our Dutch essays are not written in such format. For instance, an essay starts with a story that leads to the topic, or an example in the middle of an essay is described with many words. It is also possible that an author spontaneously writes a new paragraph to describe an example, a background story, or an argument, thereby complicating the annotation task, model development and subsequent accuracy of automatic component identification.

Subjects	**Amount**
Unemployment benefits	7
The quotas for women in the workplace	6
Homosexuality	3
Ecological footprint	2
Civil servants wearing religious symbols	2
Happiness	2
Organ donation	2
Strike	1
Costs of higher education	1
Secondary education reform	1
Lower BAC limit for the young driver	1
European union	1
Longer work and mental health.	1

Table 1: The subjects addressed in the selected essays.

4. Annotation schemes applied

We apply the annotation scheme of Stab and Gurevych (2014) to annotate the argument components and their relations in our Dutch essays since it has been commonly applied for annotating essays. As for annotating quality, our annotations are based on the quality rubric from Carlile et al. (2018) because their quality annotation schemes are based on the annotated argument components in the corpus of Stab and Gurevych (2014).

Figure 1 presents an argumentative essay as in a tree structure, including the quality attributes of each component type. The aim of the annotation task is equivalent to constructing a tree by identifying the argument components, classifying their types, and identifying the relations between the components. In the next sub-sections, the tree structures and the annotation are explained in detail.

[3] A "Five-paragraph" essay starts with a clear and precise statement of the author stance, which is followed by three body paragraphs presenting the evidential supports of the stance. Finally a conclusion or summary is drawn in the last paragraph.

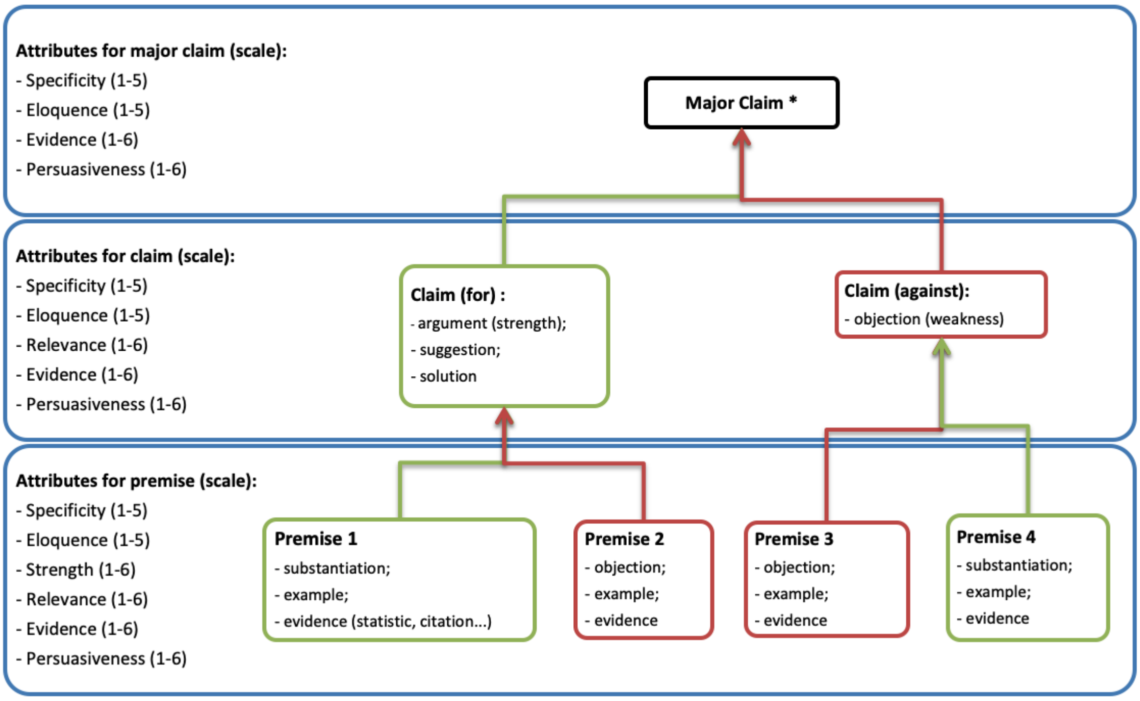

Figure 1: The structure of an argumentative essay with the quality indicators of various types of argument components. *When the "Major Claim" is not clearly present, an annotator will identify a "Topic" component. It will be explained in the later section in detail.

4.1 Annotating the argumentation components and their structure

According to the annotation scheme by Stab and Gurevych (2014), a major claim in an argumentative essay is accompanied by a set of claims and premises. A claim is a statement by the author that supports or objects to the major claim. A premise is a reason to persuade the readers to accept the other argument components.

4.1.1 Annotating argument components and types

An argument component is a node in an argumentation tree. The component is a sentence or part of a sentence containing the elements to construct a complete argument. For instance,

> Een groot nadeel van de kranten voor een euro is dat [**ze bomvol informatie staan waar jij als lezer waarschijnlijk zo goed als niets aan hebt**].
>
> *(A great drawback of the one-euro newspaper is that [**they are packed with information for which you, as a reader, most probably have no use**].)*

The example above is a complete sentence, while the section in bold is an argument component explaining the weakness of one-euro newspapers. The words before the argument component do not belong to argumentative content. Please note that an identified component is not necessary a well-formed sentence due to the particular word order in Dutch (for instance, the verb is perhaps at the end of the component if it is a subordinate clause in a complete sentence). This does, however, not influence the usefulness of the corpus for argument component detection.

To ensure consistency in further analysis, our annotators should follow particular rules when defining argument boundaries. For instance, the punctuation at the end of the component should not be included, and an argument component should not be a loose phrase but should contain all words to potentially create a well-formed Dutch sentence or clause (but disregarding punctuation)[4].

Argument components are classified into one of the component types: major claim, claim, or premise. As illustrated in Figure 1, the major claim in the top layer is the stance of the essay, and it is directly supported by at least one claim. A claim in the second layer is explained by at least one premise, while a premise is explained by at least one other premise or none if it is the leaf node (the end of an argument).

The stance of the author may be not clearly presented in the essay. It means that the "Major Claim" in the figure is not always clearly present in real world examples which leading to problems for annotators to identify it. Instead of not including such essays, we require the annotators to identify the component or sentence describing the essay's main topic and label it as "Topic". This identified topic component is regarded as a "weak" major claim, meaning that this component mentions the information of the subject of the essay but do not clearly state the polarity of the author. The purpose is to ensure the annotators can still annotate the claims and premises related to the topic. In this case, the claims in the essay are the arguments explaining the positive or negative side of the "Topic", allowing us to include essays without a clear major claim in our corpus.

4.1.2 Annotating argument component relations

The argument relations refer to the edges between nodes in the argumentation tree. An edge is directional, representing whether a child supports or attacks its parent. Figure 2 presents a simple complete argument containing only one claim (a supportive argument, objection, etc. to the major claim), one premise (a substantiation, example, or other content that explains the claim), and the directional edge (support or attack) connecting the premise to the claim. In most essays, a more complex complete argument includes more than one premise pointing to a claim and possibly one or more premises supporting or attacking other premises.

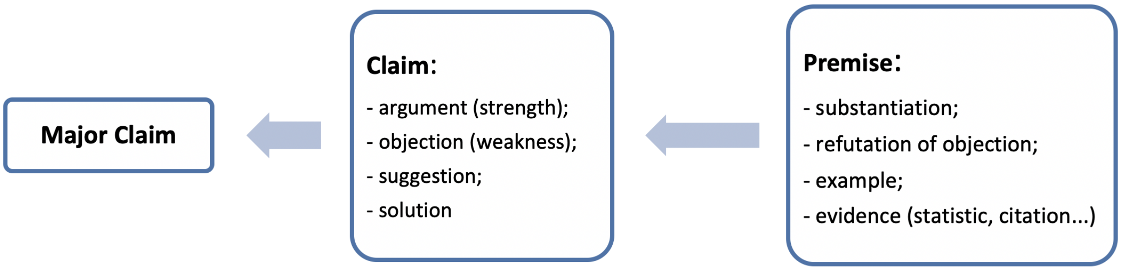

Figure 2: A simple complete argument.

The relation between a claim and a major claim is annotated as "for" and "against", indicating whether a claim supports or attacks the major claim. A relation pair of $< claim, premise >$ or $< premise, premise >$ is annotated as "support" or "attack". Usually, the major claim and claim(s) are not in the same paragraph. Thus, it is possible to create a relation of 'for' and 'against' across paragraphs, which is also the case in the corpora used in earlier research with essays in a structured format. However, in our case, the "support" and "attack" relations were also across paragraphs

[4] See Appendix A for the details of the rules to be applied while annotating.

because of the less well-structured source essays. This is not the case in the corpora with standard formatted essays, as they require that a complete argument is constructed within one paragraph.

4.2 Annotating argument component quality

Carlile et al. (2018) propose an annotation scheme for argument component quality and a rubric for scoring the argument component quality. Argument quality is defined as the persuasiveness of the argument components based on related attributes impacting the persuasiveness. For our annotations, we select the quantitative attributes from the rubric of Carlile et al. (2018). They are specificity, eloquence, strength, relevance, and evidence. As shown in Table 2, persuasiveness is scored on a scale of 1 to 6 (where 1 is low and 6 is high), while the attributes relevant to the persuasiveness are scored 1 to 5 or 1 to 6 (where 1 is low and 5 or 6 is high).

Attribute	**Scale**	**Applicable to**	**Description**
Persuasiveness	1-6	MC, C, P	How persuasive the component is.
Group 1: describe the component			
Specificity	1-5	MC, C, P	How detailed and specific the statement is.
Eloquence	1-5	MC, C, P	How well the idea is presented.
Strength	1-6	P	How well a single statement contributes to persuasiveness.
Group 2: describe the relationship between components			
Relevance	1-6	C, P	The relevance of the statement to the parent statement.
Evidence	1-6	MC, C, P	How well the supporting statements support their parent.

Table 2: Summary of the quality attributes adapted from Carlile et al. (2018) (MC: Major Claim, C: Claim, P: Premise).

We divide the attributes into two groups. The attributes in Group 1 describe the properties of an argument component itself, while the attributes in Group 2 describe the relationships between the components and the connected components. Depending on the characteristics of the component types, the attributes applied to each type vary accordingly.

4.3 Annotation procedure

Figure 3 presents the workflow of the annotation procedure. First, we wrote an annotation guideline, set up an online annotation environment, and then recruited the participants for preparation. The participants in our task were called annotators. The annotators annotated the argument components and types in the essays via BRAT, an online annotation tool (Stenetorp et al. 2012). As it was impossible to start annotating the relations between components without a corpus containing the finalised argument components and type annotations, we consolidated the corpus (resolving differences between annotations) before the annotators annotated relations between components. Similarly, quality could not be annotated without a second consolidation step in which we consolidated the annotated argument components and relations. In the end, a corpus with annotated argument components, types, relations, and qualities was compiled. Figure 3 also explains each stage in detail.

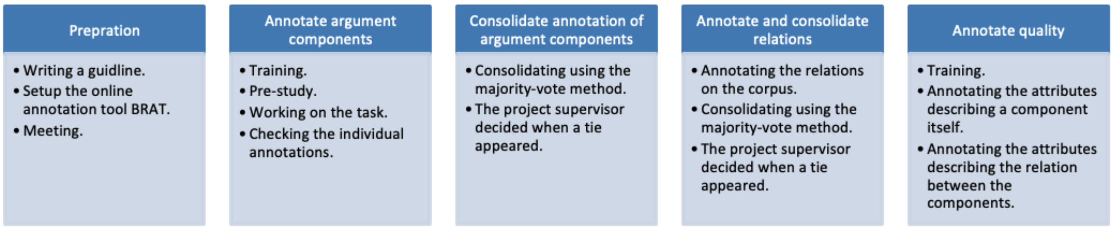
Prepration
• Writing a guidline.
• Setup the online annotation tool BRAT.
• Meeting.
Annotate argument components
• Training.
• Pre-study.
• Working on the task.
• Checking the individual annotations.
Consolidate annotation of argument components
• Consolidating using the majority-vote method.
• The project supervisor decided when a tie appeared.
Annotate and consolidate relations
• Annotating the relations on the corpus.
• Consolidating using the majority-vote method.
• The project supervisor decided when a tie appeared.
Annotate quality
• Training.
• Annotating the attributes describing a component itself.
• Annotating the attributes describing the relation between the components.

Figure 3: The workflow of the annotation task.

Preparation: We recruited four native Dutch speakers who were teachers in secondary or higher education institutions. Each annotator received a compensation of 300 Euro to finish the annotation task. Before the annotation task, an annotation guideline was written based on the guideline from Stab and Gurevych (2014). The guideline included the description of the task, the introduction of the relevant concepts of argumentation and the rules to be applied during annotation. We used BRAT as the annotation environment, a popular tool for various annotation tasks (Stenetorp et al. 2012). The tool was set up online and publicly available with assigned anonymous accounts so that the annotators could work on the task remotely in their own time. To prepare the annotators for their task, we arranged online meetings to explain the task, the tool, the workflow, and other helpful information.

Annotating argument components: A training meeting was arranged for each annotator, introducing the annotation guideline and the instruction to use BRAT in detail. To ensure that the annotators understood the annotation rule and the annotation satisfied the annotation requirement, we conducted a pre-study: every annotator annotated five selected identical essays from the collection of 30 essays, and we discussed the annotation results of these five essays with each annotator. After the pre-study, the annotators annotated the remaining 25 essays. We monitored the annotated essays carefully to fix any technical annotation mistakes. For instance, it was a common mistake that the final letter of an argument component was not selected in an annotation.

Consolidating annotation of argument components: After the four annotators finished annotating the argument components individually, their individual work was consolidated into one annotation following a majority voting principle. In case of a tie, the authors decided on the final annotation. By this approach, we ensured the final annotation was the result of considering the opinions of the four annotators and a fifth person (the author) to guarantee annotation quality and correctness.

Annotating and consolidating relations between argument components: The consolidated corpus resulting from the previous step was presented to the annotators for annotating the relations between the components. As in the previous step, the annotators annotated the relations individually, while their results were consolidated using the majority voting system. Again, the authors decided on annotations in cases of ties. A final consolidated corpus with argument structures was thus generated.

Annotating quality: During a training meeting, the quality annotation procedure was explained to the annotators. The relation between persuasiveness and the other attributes was not mentioned to the annotators to avoid bias, so they regarded persuasiveness as an attribute like the others. The final score of an attribute in a component was the median of the four annotator scores because using the median could eliminate the extreme scores. If the median value was not an integer, the final score was determined by us. As the scoring of the attributes in Group 2 depended on the prior scoring of the attributes in Group 1, the attributes in Group 2 were scored after scoring attributes in Group 1.

5. Analysis of corpus annotations

We employed several inter-annotator agreement measures on the four individual sets of annotations to examine the reliability of the final corpus in order to understand how much the annotators "agreed with" each other. The outcomes were indicators of the quality of the annotation procedure and the design of the annotation guideline in accordance with the previous corpus compilation studies (Stab and Gurevych 2014, Wambsganss et al. 2020). We calculated the Inter-Annotator Agreement (IAA) on annotating the argument components, types, relations, and quality over the collective annotations of all 30 essays. Besides, we applied the confusion probability matrices (CPM) to analyse the disagreement between the four annotators.

5.1 Inter-annotator agreement on argument components

Before determining their types, the participants needed to decide the argument component boundaries (what exactly to select from the essay text). Therefore, we applied Krippendorff's α_U (Krippendorff 2004) to measure the agreement on the component boundaries because Krippendorff's α_U could consider the components boundary difference from more than two annotators. The value ranges from 0 to 1, where 0 indicates "completely disagree" and 1 indicates "completely agree". Krippendorff (2004) suggests that a value greater than 0.667 is acceptable. Table 3 shows the agreement score per argument component type. The result is the average value over all essays. The results indicate that major claim (α_U=0.403) receives moderate agreement. However, both the claim (α_U=0) and premise (α_U=-0.203) scores are relatively low.

Types	**Krippendorff's α_U**
Major claim	0.403
Claim	0
Premise	-0.203

Table 3: Krippendorff's α_U of argument component boundary per type.

.

For the argument component types, we calculated the agreement at the sentence level because only 1.9% of the sentences were annotated with more than one argument type. To calculate the agreement on a type, we transferred the tags on the token level to the sentence level. As a result, every sentence in an essay was assigned with a tag representing the argument type of the sentence, while "none" was assigned to a sentence when there was no argument component annotated. We employed Krippendorff's α (Krippendorff 1980) and Fleiss's κ (Fleiss 1971), which are widely used in measuring agreements in text annotations. The results (see Table 4) show relatively low agreement values for all component types compared to previous corpus studies. The Krippendorff's α values are between 0.28 and 0.37 in our case, while they are between 0.51 and 0.55 in Wambsganss et al. (2020) and between 0.66 and 0.83 in Stab and Gurevych (2014) respectively.

	Krippendorff's α	**Fleiss's κ**
MajorClaim	0.368	0.366
Claim	0.281	0.280
Premise	0.374	0.387
Overall	0.338	0.350

Table 4: The Krippendorff's α and Fleiss's κ of argument component types.

5.2 Inter-annotator agreement on relations between argument components

The calculation of the relation agreement was based on the annotated relation pairs against all possible valid relation pairs generated according to the annotation rules. Suppose a relation between two argument components is denoted as

$$< A, B >= r$$

where A, B are the argument components, r is the relation. When $r \in (support, attack)$, it means that component B supports or attacks A. When $r \in (none)$, it means there is no relation from B to A. A valid pair should satisfy: $< majorclaim, claim >$, $< claim, premise >$, $< premise, premise >$, or $< majorclaim, premise >$.

Among all valid pairs, the proportion of the relation of the type "none", "support", or "attack" was 97.5%, 2.1%, and 0.4%, respectively. The ratio of none relations was much higher compared to previous studies. The main reason is that earlier studies defined the valid pairs inside the same paragraph (unless the pair existed between major claim and claim), while we had to define the valid pairs across paragraphs, leading to more valid pairs in our study due to the less structured nature of our essays. Krippendorff's α was applied as well for IAA measurement. As a result, relations of support (α=0.439) and none (α=0.731) have moderate IAA, while the attack (α=0.282) is relatively low. Considering the low occurrence of the attack relation, we decided to classify the relations into support and non-support only. The agreement for both support and non-support is 0.694, indicating a moderate agreement.

5.3 Inter-annotator agreement on argument component quality

Table 5 presents the distribution of the scores given by each of the annotators. In general, the annotators preferred not to provide extremely low or high scores. There was one notable outlier: when scoring specificity and eloquence (on a scale of 1-5), more than 88% of the scores given by Annotator 1 were over 4, while it was only around 60% for the other annotators. We applied Krippendorff's α again to calculate the agreement of the attributes. As shown in Table 6, the values are considerably lower than the results (between 0.5 to 0.9) obtained by Carlile et al. (2018). We believe this is due to the higher number of annotators involved, as the chance to see different scores gets higher. Also, various annotators may have their own standard and scoring preferences, while resolving the discrepancies is more problematic when dealing with four annotators compared to two annotators, as in Carlile et al. (2018).

Scores	**Annotator1**	**Annotator2**	**Annotator3**	**Annotator4**
The attributes on the scale of 5 (specificity and eloquence)				
1	0.1%	1.5%	2.9%	0.1%
2	1.1%	7.6%	15.3%	1.9%
3	10.2%	24.1%	22.6%	34.1%
4	38.0%	50.7%	18.1%	58.5%
5	50.7%	16.1%	41.0%	5.4%
The attributes on the scale of 6 (persuasiveness, strength, relevance, and evidence)				
1	2.9%	12.2%	5.0%	9.7%
2	5.3%	23.6%	11.0%	2.1%
3	12.0%	17.6%	33.7%	18.3%
4	18.2%	26.7%	29.6%	29.1%
5	29.6%	16.7%	19.9%	32.0%
6	32.0%	3.2%	0.8%	8.8%

Table 5: The distribution of the attribute scores given by each annotator.

Attributes	**Krippendorff's α**
Persuasiveness	0.004
Eloquence	-0.020
Specificity	0.055
Strength	0.206
Evidence	0.041
Relevance	-0.138

Table 6: Krippendorff's α of the argument component quality per attribute.

5.4 Annotator disagreement analysis

To analyse the disagreement between the four annotators, we applied the confusion probability matrices (CPM) proposed by Cinková et al. (2012). CPM is the conditional probability that an annotator assigns a certain tag to an item, given that another annotator assigns the tag to that item. It is applied to analyse the disagreement of an annotation study involving more than two annotators (Stab and Gurevych 2014, Wambsganss et al. 2020). We calculated the CPM for the argument component types, relations types, and quality indicators.

Table 7 presents the CPM for the argument component types. The result shows the primary disagreement between 'none' and the argument types from 0.471 to 0.695. This implies that when an annotator assigns an argument type in a sentence, another annotator probably finds no argument component in this sentence.

	None	**Major claim**	**Claim**	**Premise**
None	0.545	0.031	0.101	0.322
Major claim	0.695	**0.233**	**0.031**	**0.004**
Claim	0.464	**0.007**	**0.215**	**0.314**
Premise	0.471	**0.003**	**0.100**	**0.426**

Table 7: Confusion probability matrix for the argument components types. The annotated "topic" is considered as "none" in this table.

When an annotator did not assign a type in a sentence (assigning "none"), the probability that another annotator also did not assign a type is over 0.545, while the probability to assign 'premise' is more than 0.322. This implies that the annotators have a relatively strong disagreement when deciding whether a sentence contains a premise or no argument. Ignoring the "none" and only focusing on the annotations of the component types, a significant disagreement is found between claim and premise (p=0.314). It is difficult to distinguish between a claim and premise because a claim becomes a premise of another claim if the latter claim is not identified by an annotator. This problem is identified by Stab and Gurevych (2014) and Wambsganss et al. (2020) as well.

We also calculated CPM between the "topic" component and other argument component types mentioned above, when the annotators did not find a clear major claim but instead annotated a "topic" component. When a component is regarded as "topic" by an annotator, the probability of another annotator assigning a major claim to the component is over 0.2, which is higher than the agreement probability (p=0.1). It indicates that the annotators had a high disagreement in determining if a sentence was a clear major claim or not. In other words, although they identified the same component potentially describing the stance of the author, the annotators had different opinions on whether the polarity of the component was clearly recognisable.

Table 8 presents the CPM for the relation annotations, showing that the annotators had a high agreement on determining when a pair showed no relation. However, when an annotator assigned a support/attack relation, there was over 42% probability that it was assigned as "none" by another

	none	**support**	**attack**
None	0.976	0.019	0.005
Support	0.427	0.515	0.058
attack	0.462	0.253	0.286

Table 8: Confusion probability matrix for the annotation of argument component relation.

	1	**2**	**3**	**4**	**5**	**6**
1	0.045	0.120	0.284	0.338	0.172	0.041
2	0.075	0.040	0.264	0.350	0.221	0.049
3	0.091	0.136	0.106	0.344	0.264	0.059
4	0.086	0.143	0.272	0.159	0.276	0.064
5	0.054	0.112	0.259	0.343	0.171	0.061
6	0.053	0.102	0.239	0.327	0.248	0.030

Table 9: Confusion probability matrix for the persuasiveness scores.

annotator. The agreement on "support" relations is over 50%, while the agreement on "attack" relations is only around 28%. The probability of assigning an "attack" relation type is very low when another annotator assigns a "none" or "support" relation. The agreement probability on the "attack" relation is relatively acceptable.

We also calculated the CPM for the persuasiveness scores (see Table 9). The result reveals that there is high confusion between all scoring, showing no significant agreement when scoring the persuasiveness. The CPM for other attributes shows similar results.

6. Corpus example, statistics, and analysis

In this section we first provide an example of the annotations contained in our corpus. Next, the statistic description of the argument components is introduced. Following that, we analyse the annotated quality values in the corpus to understand how the attribute values are relevant to the persuasiveness value. Finally, we compare the annotation results between the essays with and without annotated major claim.

6.1 An example in the corpus

The example below shows a major claim, a claim, and some premises from an essay in the corpus. Moreover, an example of the attribute scores of an argument component is given to provide insight into the attributes representing argumentation quality. Below is a major claim taken from an essay in the corpus.

> Major claim: lijkt het mij dan ook overbodig om op de barricades te gaan staan voor vrouwelijke quota aan de top van het bedrijfsleven (*I think it is unnecessary to stand on the barricades for female quotas at the top of the business world*)

This component is a typical major claim clearly stating the stance of the author that having female quotas in business is not necessary. The author has several arguments. One is that:

> Claim (for): Angela Merkel is één van de mooiste voorbeelden waarom quota niet nodig zijn voor de bedrijfstop (*Angela Merkel is one of the best examples why quotas are not necessary for the corporate top*)

The author proposes Angela Merkel is a good example to support the stance (the major claim of the essay) and he/she continues to provide some reasons:

Premise1: Ze is een self-made vrouw die zich tot de top van het machtigste land van Europa heeft weten op te werken (*She is a self-made woman who has risen to the top of the most powerful country in Europe)*

Premise2: De bondkanselier bepaalt het reilen en zeilen binnen Europa (*The Chancellor determines the ins and outs within Europe*)

Premise3: haar wil is wet (*her will is law*)

These three premises constitute a reasoning chain: Premise 1 is a reason to support the claim, while Premise 2 supports 1 and Premise 3 supports 2. This is one of the reasoning chains in the essay.

Table 10 shows the scores of the attributes representing the quality of the argument component Premise 1. In general, it is a persuasive argument giving its persuasiveness score is 4 out of 6. Both the scores of specificity and eloquence are 4 out of 5. It implies that the argument component is quite specific (using concrete languages "the most powerful country in Europe") and well presented (easy to understand). The Strength score is 4, meaning that the argument is not a very strong argument but still acceptable. On the other hand, the relevance score is only 3 out of 6. It indicates that stating Markel is a self-made woman is not very supportive to the claim of "she is one of the best examples". Nevertheless, its evidence score is 5 out of 6, meaning that the premises that explaining Premise 1 is quite supportive to it.

Attributes	**Scores**
Persuasiveness	4
Eloquence	4
Specificity	4
Strength	4
Evidence	3
Relevance	5

Table 10: The attribute scores of the example premise 1.

6.2 Statistic description of the argument components

Table 11 shows an overview of statistics calculated from the compiled corpus. On average an essay contains 44 sentences and 770 tokens. Twenty-six out of the 30 essays have at least one major claim, while six have no clear major claim. The average number of argument components in the final corpus of the major claim, claim, and premise is 1.03, 4.13, and 18.7, respectively, covering 2.2%, 7.38%, and 36.77% of the words in an essay.

The essays in our corpus contain much more tokens compared to the essays in the similar corpora of Stab and Gurevych (2014) and Wambsganss et al. (2020). In our corpus the proportion between premise and claim is around 4.6. It is higher than the corpora mentioned above (about 2.4 and 1.06, respectively). This indicates that there are more premises explaining the claims in our corpus. In other words, the claims in our corpora are elaborated in more detail, and the authors establish more complex reasoning chains and arguments in the essays.

On average non-argumentative content occupies 53.6% of an essay. This is also much higher compared to 30.2% in the corpus of Stab and Gurevych (2014) and 34.2% in Wambsganss et al.

	All	Ave. per essay	Sd
Sentence	1,326	44.2	7.10
Tokens	23,115	770	130
MajorClaim	31	1.29	0.46
Claim	123	4.13	1.10
Claim (for)	85	2.83	1.26
Claim (against)	38	1.27	0.94
Premise	561	18.7	5.12
Support	482	16.07	4.94
Attack	73	2.7	1.68

Table 11: Statistic of the created corpus.

(2020). In terms of the relation types, there are 2.8 claims with the stance "for", 1.3 with the stance "against", 16.07 "support" relations, and 2.7 "attack" relations on average. Counting the "for" and "against" as "support" and "attack", the ratio of "support" versus "attack" types is 5:1. This ratio is similar to the 7:1 in the corpus of Stab and Gurevych (2014) but much lower compared to the 32:1 found in Wambsganss et al. (2020).

6.3 Analysis of argument component quality annotations

Each argument component in the compiled corpus contains a score on persuasiveness, while the scores on the other five attributes are correlated to this persuasiveness score. To analyse the quality annotations in our corpus, we applied the same method as used in Carlile et al. (2018) to explore how each attribute correlates to the persuasiveness score and which impact these five attributes taken together have on the persuasiveness score.

Our corpus contains 4,061 attribute scores in total representing the argument quality measured by persuasiveness and its related attributes. As shown in Table 12, more than 93% of the given scores of persuasiveness are between 3 and 5, while the extreme scores of 1 and 2 make out only 0.7% of the total. The score distributions of other attributes are similar to the persuasiveness scores distribution that the majority of the scores are in the middle ranges (except for evidence because premises without any descendant are given low evidence scores, leading to more scores at the low level)[5].

Scores	Percentage	Percentage
1	0.004	0.40%
2	0.059	5.90%
3	0.273	27.30%
4	0.511	51.10%
5	0.15	15.00%
6	0.003	0.30%

Table 12: The distribution of the scores for persuasiveness.

The Pearson Correlation Coefficients between each attribute and persuasiveness are shown in Table 13. The attributes of strength, relevance, and evidence are moderately positively correlated to persuasiveness ($r > 0.4$), while specificity and eloquence correlate relatively weak.

[5] See Appendix B for the full distribution in detail.

	Pearson correlation r	**p-value**
Specificity	0.295	<0.001
Eloquence	0.123	0.001
Strength	0.403	<0.001
Relevance	0.436	<0.001
Evidence	0.481	<0.001

Table 13: The Pearson correlation between each attribute to the persuasiveness score.

We also explored how these five attributes together explained the persuasiveness of an argument. Due to the varied attribute combinations for different component types, three logistic regression models were trained for the major claim, claim, and premise, respectively. The models were trained on the arguments as parents. This means that an instance to train a model consisted of the attributes scores of a component itself and the average attributes scores of all its children[6]. For example, for a random premise, p, whose set of children $C_p i$ is not empty, the instance for the premise as a parent is a tuple as follow:

$$(a_{Sp}, a_{El}, a_{St}, a_{Re}, a_{Ev}, \frac{1}{n}\sum c_{Sp}, \frac{1}{n}\sum c_{El}, \frac{1}{n}\sum c_{St}, \frac{1}{n}\sum c_{Re}, \frac{1}{n}\sum c_{Ev},)$$

where a is an attribute score of p, c is an attribute score of a component in C_p, and n is the cardinal number of C_p. The input feature of the model was a tuple, and the predicted result was c_{Per}, the persuasiveness score of p. We conducted a 5-fold cross validation at the document level to evaluate the models. The performances of the models were measured by the Pearson correlation and the mean absolute error[7] (MAE) between the predicted persuasiveness score and the actual persuasiveness score. Notably, the Pearson correlation value represents the correlation between the predicted persuasiveness score and the actual score, while MAE represents the error. Hence, better performance means a higher correlation value and lower MAE value.

Table 14 shows the result of Pearson correlation and MAE per type. The correlation values of the claim and premise indicate that the combination of attributes is positively correlated to persuasiveness, while the value for the major claim indicates “no correlation” . It is probably due to the small amount of major claim data since there are only 29 instances. As for the MAE values, the results are similar in all three types (all below 0.43). It implies that when using the combination of the attributes to explain persuasiveness, the error is within 0.43 to the actual persuasiveness value.

	Claim	**Premise**	**Major Claim**
Pearson correlation	0.42	0.57	NA*
MAE	0.43	0.43	0.42

Table 14: Performances of persuasiveness prediction based on attributes. *The standard deviation of the predicted values is zero. Hence the correlation value is not calculable as “divided by zero”.

[6] We use the average attributes scores of the children of the component because we need to ensure that an instance represents one component. The number of children varies per component and having more children leads to more instances for a component. Using the average scores of the children makes a component represented by one instance.

[7] Mean absolute error is the mean value of the absolute differences between true value and predicted value.

6.4 Comparison between essays with and without clear major claim

	With MC	**Without MC**
# of essays	24	6
Ave. # of sentences	44.8	41.2
Ave. # of claims	3.72	3.83
Ave. # of premises	16.96	15

Table 15: Statistic description of essays with and without major claims.

Some essays in the corpus contain no clear major claim. Hence, we attempted to compare the essays with and without clear major claim to explore whether this has a significant impact on the annotations and quality. There are 24 essays with at least one major claim in the compiled corpus. Table 15 shows the statistic description comparison between essays with and without a clear major claim. It shows that the average number of claims is similar in both cases. The average number of premises for the essays with the major claim (ave=16.96) is slightly higher compared to the essays without a major claim (ave=15), although the difference is not significant (t=0.89, p=0.38).

As for the quality, the distribution of the given scores of the essays with the major claim is similar to the essays without (see Table 16). There is no significant difference between the annotation in an essay with and without a clear major claim. This opens the possibility to include essays without clear major claims in corpora for analysing argumentative essays.

Score	**With MC**	**Without MC**
1	0.017	0.013
2	0.082	0.055
3	0.234	0.18
4	0.453	0.5
5	0.208	0.246
6	0.006	0.005
Total	1	1

Table 16: The distributions of all the collected scores comparing between essays with and without annotated major claims.

7. Conclusions

In this article we report on the creation of the first Dutch essay corpus with annotated argument structures and quality indicators. The corpus contains 30 essays, and each of them contains the annotations of the argument components at the token levels and the relations between the argument components. The quality of the argument components is annotated and measured by persuasiveness scores and the scores of its related attributes. The inter-annotator agreement analysis shows relatively low agreement between the annotators. There are multiple reasons for this: Firstly, the annotations are performed on essays not using a standard format, leading to less organised text structures. Secondly, the essays have a longer length and higher percentage of non-argument contents when compared to essays used in other analyses. These essay characteristics increase the chances of different opinions between annotators. We believe that annotating argument structures in structured essays leads to more accurate annotations because the argument components are easily recognised. Besides, we did not attempt to maximise the agreement between the annotators via in-depth group discussion together with all the annotators. The opinions of the annotators were not unified to reach a similar standard and understanding of the annotation found in other studies. Another factor that leads to low agreement values between annotators is that the agreement calculation was done on the annotation of the entire corpus instead of a small section of the corpus as was done in other studies. One possible solution to increase inter-annotator agreement would be to further improve the guidelines we used and to allow the annotators to annotate the argument components at the sentence level, while omitting the argument types. This method is proposed in a recent work by Putra et al. (2021). Four annotators participated in the annotation tasks, sometimes leading to ambiguous annotations that had to be resolved by the authors. In contrast, previous studies used only one or two annotators to annotate an essay in their final corpus. We believe that although a higher number of annotators may lead to lower inter-annotator agreement values, it can also guarantees the objectivity and correctness of the corpus.

In future studies, more essays from the Dutch CSI corpus will be annotated to expand our annotated corpus. As the essays in our corpus are not well-structured, it might be beneficial to train a model with a hybrid corpus in which well-structed and less well-structured essays are contained. The model resulting from such a corpus is probably more practically applicable, because the training data covers essays in written in different structures.

With the corpus we provide and plan to extend, we aim to create models for argumentation analysis. The created models service to train supervised machine learning tools that can provide automated argument analysis for Dutch essays based on the annotation of the argument components and relations. In the future we hope to develop an automated essay scoring tool on the argumentation aspect of an essay, which is why we included scores of persuasiveness and other relevant attributes in the corpus. Such tools can become beneficial to generate automated feedback so that teachers and students can take advantage of it for improving argument writing and assessment. In future research, we will also explore the role of the non-argument content in argumentative essays and how this content is relevant to connecting the argument components in argumentation trees.

References

Carlile, Winston, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng (2018), Give me more feedback: Annotating argument persuasiveness and related attributes in student essays, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 621–631.

Cinková, Silvie, Martin Holub, and Vincent Kríž (2012), Managing uncertainty in semantic tagging, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 840–850.

Eger, Steffen, Johannes Daxenberger, and Iryna Gurevych (2017), Neural end-to-end learning for computational argumentation mining, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1, pp. 11–22.

Eger, Steffen, Johannes Daxenberger, Christian Stab, and Iryna Gurevych (2018), Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need!, *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 831–844.

Ferretti, Ralph P and William E Lewis (2019), Knowledge of persuasion and writing goals predict the quality of children's persuasive writing, *Reading and Writing* **32** (6), pp. 1411–1430, Springer.

Fisas, Beatríz, Francesco Ronzano, and Horacio Saggion (2016), A multi-layered annotated corpus of scientific papers, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3081–3088.

Fishcheva, Irina and Evgeny V. Kotelnikov (2019), Cross-lingual argumentation mining for Russian texts., *International Conference on Analysis of Images, Social Networks and Texts* pp. 134–144.

Fleiss, Joseph L (1971), Measuring nominal scale agreement among many raters., *Psychological bulletin* **76** (5), pp. 378, American Psychological Association.

Freeman, James B (2011), *Argument Structure: Representation and Theory*, Vol. 18, Springer Science & Business Media.

Gallagher, H Alix, Katrina R Woodworth, and Nicole L Arshan (2015), Impact of the national writing project's college-ready writers program on teachers and students, *Menlo Park, CA: SRI International.*

Gao, Yanjun, Alex Driban, Brennan Xavier McManus, Elena Musi, Patricia Davies, Smaranda Muresan, and Rebecca J Passonneau (2019), Rubric reliability and annotation of content and argument in source-based argument essays, *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 507–518.

Ghosh, Debanjan, Aquila Khanam, Yubo Han, and Smaranda Muresan (2016), Coarse-grained argumentation features for scoring persuasive essays, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 549–554.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, Magali Paquot, et al. (2009), *International corpus of learner English*, Presses universitaires de Louvain Louvain-la-Neuve.

Gretz, Shai, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim (2020), A large-scale dataset for argument quality ranking: Construction and analysis., *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 7805–7813.

Habernal, Ivan and Iryna Gurevych (2016), What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in web argumentation, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1214–1223.

Iida, Ryu, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto (2007), Annotating a Japanese text corpus with predicate-argument and coreference relations, *Proceedings of the linguistic annotation workshop*, pp. 132–139.

Ke, Zixuan, Hrishikesh Inamdar, Hui Lin, and Vincent Ng (2019), Give me more feedback ii: Annotating thesis strength and related attributes in student essays, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3994–4004.

Krippendorff, Klaus (1980), *Content analysis: An introduction to its methodology*, Sage publications.

Krippendorff, Klaus (2004), Measuring the reliability of qualitative text analysis data, *Quality and quantity* **38**, pp. 787–800, Springer.

Lauscher, Anne, Goran Glavaš, and Simone Paolo Ponzetto (2018), An argument-annotated corpus of scientific publications, *Proceedings of the 5th Workshop on Argument Mining*, pp. 40–46.

Lauscher, Anne, Lily Ng, Courtney Napoles, and Joel Tetreault (2020), Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing, *arXiv preprint arXiv:2006.00843.*

Lawrence, John and Chris Reed (2020), Argument Mining: A Survey, *Computational Linguistics* **45** (4), pp. 765–818.

Li, Mengxue, Shiqiang Geng, Yang Gao, Shuhua Peng, Haijing Liu, and Hao Wang (2017), Crowdsourcing argumentation structures in Chinese hotel reviews, *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, pp. 87–92.

Lippi, Marco and Paolo Torroni (2016), Argumentation mining: State of the art and emerging trends, *ACM Transactions on Internet Technology (TOIT)* **16** (2), pp. 1–25, ACM New York, NY, USA.

Mann, William C and Sandra A Thompson (1988), Rhetorical structure theory: Toward a functional theory of text organization, *Text-interdisciplinary Journal for the Study of Discourse* **8** (3), pp. 243–281, Walter de Gruyter, Berlin/New York.

Palau, Raquel Mochales and Marie-Francine Moens (2009), Argumentation mining: the detection, classification and structure of arguments in text, *Proceedings of the 12th international conference on artificial intelligence and law*, pp. 98–107.

Peldszus, Andreas and Manfred Stede (2015), An annotated corpus of argumentative microtexts, *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, Vol. 2, pp. 801–815.

Purdue, Online Writing Lab (2021), Argumentative essays.

Putra, Jan Wira Gotama, Simone Teufel, and Takenobu Tokunaga (2021), Annotating argumentative structure in English-as-a-foreign-language learner essays, *Natural Language Engineering* p. 1–27, Cambridge University Press.

Redeker, Gisela, Ildikó Berzlánovich, Nynke Van Der Vliet, Gosse Bouma, and Markus Egg (2012), Multi-layer discourse annotation of a Dutch text corpus, *age* **1**, pp. 2, Citeseer.

Rocha, Gil and Henrique Lopes Cardoso (2017), Towards a relation-based argument extraction model for argumentation mining, *International Conference on Statistical Language and Speech Processing* pp. 94–105.

Rocha, Gil, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych (2018), Cross-lingual argumentative relation identification: from English to Portuguese, *Proceedings of the 5th Workshop on Argument Mining*, pp. 144–154.

Stab, Christian and Iryna Gurevych (2014), Annotating argument components and relations in persuasive essays, *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pp. 1501–1510.

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

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii (2012), Brat: a web-based tool for NLP-assisted text annotation, *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Avignon, France, pp. 102–107.

Toledo, Assaf, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim (2019), Automatic argument quality assessment - new datasets and methods., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5624–5634.

Toledo-Ronen, Orith, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim (2020), Multilingual argument mining: Datasets and analysis, *arXiv preprint arXiv:2010.06432*.

Van Der Vliet, Nynke, Ildikó Berzlánovich, Gosse Bouma, Markus Egg, and Gisela Redeker (2011), Building a discourse-annotated Dutch text corpus, *Bochumer Linguistische Arbeitsberichte* **3**, pp. 157–171.

Verhoeven, Ben and Walter Daelemans (2014), CLiPS stylometry investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text., *LREC*, pp. 3081–3085.

Visser, Jacky, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton (2021), Annotating argument schemes, *Argumentation* **35** (1), pp. 101–139.

Wachsmuth, Henning, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein (2017), Computational argumentation quality assessment in natural language, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 176–187.

Wagemans, Jean H.M. (2016), Constructing a periodic table of arguments, *Social Science Research Network*.

Walton, Douglas, Christopher Reed, and Fabrizio Macagno (2008), *Argumentation schemes*, Cambridge University Press.

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

Zhang, Liqin, Howard Spoelstra, and Marco Kalz (2020), Argument component identification and its application in feedback on Dutch essays, *Companion Proceedings 10th International Conference on Learning Analytics & Knowledge (LAK20)*, pp. 732–735.

Appendix A. Rules to define boundary of an argument

- Completeness: an argument component should be a complete sentence.

  Example:

  > Door [de exponentiele ontwikkeling van het internet] (1) [barst het van de informatieve websites] (2).
  >
  > *(Due to [the exponential development of the Internet] (1) [it is bursting with informative websites] (2).)*

  Although (1) is the reason of (2), but we cannot consider (1) as an argument component because it is not a complete sentence.

- Relevance: you should include all relevant materials in an argument component.

  Example:

  > Door [de exponentiele ontwikkeling van het internet] (1) [barst het van de informatieve websites] (2).
  >
  > *(Due to [the exponential development of the Internet] (1) [it is bursting with informative websites] (2).)*

  (2) itself is a complete sentence, and it is the core component of this sentence. However, (1) provides extra information in the sentence, and it should be included because it is relevant to (2).

- No shell language: you should not include shell languages such as "mijn standpunt is" (*my standpoint is*), "bijvoorbeeld" (*for example*), because they are context-irrelevant.

  Example:

  > Een groot nadeel van de kranten voor een euro (en een beetje) is dat [ze bomvol informatie staan waar jij als lezer waarschijnlijk zo goed als niets aan hebt].
  >
  > *(A big disadvantage of the newspapers for a euro (and a bit) is that [they are full of information that is probably of no use to you as a reader].)*

  All the words before "dat" are not related to the context in the core part of the sentence. Hence, it should not be included in the argument component.

- Splitting: if a sentence contains more than one argument components, it should be split. Normally these two components are connective by a connective such as "omdat", "want", "en".

  Example:

  > [Reclame wordt ook nu al minder bekeken door een aantal mensen] (1), want [er is waarschijnlijk niemand die voor zijn plezier naar de onderbreking kijkt] (2).
  >
  > *([Advertising is already being watched less by a number of people] (1), because [there is probably no one who watches the interruption for pleasure] (2).)*

It is obvious to identify that (2) is the premise of (1) since these two components are connected by “want”. Hence, this sentence should be split into two argument components.

- Punctuation: the punctuation at the end of an argument component should not be included.

  Example:

> [Vanwege het gemak rijden veel mensen met hun eigen auto].
>
> *([Because of convenience, many people drive their own cars].)*

The full stop is not included in the component.

Appendix B. The distribution of the scores for the persuasiveness and the attributes.

	Scores	**Percentage**
Persuasiveness	1	0.40%
	2	5.90%
	3	27.30%
	4	51.10%
	5	15.00%
	6	0.30%
Specificity	1	0.00%
	2	3.50%
	3	36.80%
	4	54.20%
	5	5.50%
Eloquence	1	0.00%
	2	0.00%
	3	7.5%
	4	69.70%
	5	22.80%
Strength	1	0.00%
	2	2.00%
	3	20.40%
	4	45.00%
	5	32.40%
	6	0.20%
Relevance	1	0.00%
	2	0.70%
	3	2.90%
	4	25.50%
	5	67.60%
	6	3.20%
Evidence	1	4.2%
	2	23.30%
	3	43.30%
	4	27.90%
	5	1.30%
	6	0.00%