The Shift from Logic to Dialectic in Argumentation Theory: Implications for Computational Argument Quality Assessment

Rositsa V Ivanova
Institute of Computer Science
University of St. Gallen, Switzerland
rositsa.ivanova@unisg.ch

Reto Gubelmann
Digital Society Initiative &
Department of Computational Linguistics
University of Zurich, Switzerland
reto.gubelmann@uzh.ch

Abstract

In the field of computational argument quality assessment, logic and dialectic are essential dimensions used to measure the quality of argumentative texts. Both of them have found their way into the field due to their importance to argumentation theory. We trace the development of core logical concepts of validity and soundness from their first use in argumentation theory to their understanding in state-of-the-art research. We show how, in the course of this development, dialectical considerations have taken center stage, at the cost of the logical perspective. Then, we take a closer look at the quality dimensions used in the field of computational argument quality assessment. Based on an analysis of prior empirical work in this field, we show how methodological considerations from argument theory can benefit state-of-the-art methods in computational argument quality assessment. We propose an even clearer separation between the two quality dimensions not only in regards to their definitions, but also in regards to the granularity at which the argumentative text is being annotated and assessed.

1 Introduction

The field of computational argument quality assessment (CA) is dedicated to the goal of accurately assessing the quality of arguments using computational methods. Starting with the assessment of essays (Persing et al., 2010; Persing and Ng, 2013, 2014, 2015) and slowly moving towards the analysis of smaller texts extracted from the internet (Wachsmuth et al., 2014; Braunstain et al., 2016), the field explored texts of various sources and types on various quality dimensions (Tan et al., 2016; Habernal and Gurevych, 2016a; Durmus et al., 2019; Gretz et al., 2020). While many approaches considered individual texts on their own (i.e., absolute quality), a few also compared pairs of arguments (i.e., relative quality, e.g., Habernal and Gurevych 2016b). To systematize the previously explored quality dimensions and to connect them to existing theoretical frameworks of philosophy, Wachsmuth et al. (2017) created a taxonomy for CA, which has ever since been widely referenced and used in the field. We revisit the taxonomy and analyze it from both a philosophical-conceptual as well as an empirical perspective. We thereby particularly focus on the logical and dialectical dimensions including their sub-dimensions. The detailed analysis of rhetoric and its sub-dimensions thus remains for future work.

Making progress in this area of research is important, as we expect that formative feedback to students on their argumentative writing will be increasingly given by systems using artificial intelligence, as opposed to human instructors. Therefore, it is paramount that these systems apply a reasonable, philosophically sound and empirically valid notion of argument quality (AQ). Unfortunately, as we shall see, the topic is a complex one, involving formal as well as informal logic, epistemology, argumentation theory, and of course natural language processing (NLP). The uses of central concepts such as *argument*, *logic*, and *dialectic*[1] often differ already within one single discipline, and they don't travel well across disciplinary boundaries. As a consequence, researchers are systematically talking past each other, and empirical data becomes questionable, as it is framed in ambiguous conceptual terms. Thus, our main contributions consist in (1) describing some of the conceptual discrepancies of this field, (2) empirically determining the tangible empirical consequences of these discrepancies, and (3) making specific recommendations to amend this situation.

To create a common understanding of the terminology used in our work, we first focus on the

[1] In argumentation theory, the term "dialectic" is sometimes also written with an "s". Here, we consistently use the variant without "s" that is standard usage in CA and also found in argumentation theory, e.g., in Johnson (2009, 6).

English term "argument", but we would like to emphasize that, as research in CA becomes increasingly multilingual, it is important to be aware of the different cross-lingual meanings of "argument". Otherwise, one runs the danger of erroneously applying a definition of argument taken from English on a language where this definition is not a good fit. The terminology that we introduce here is widely, but not universally accepted. Importantly, it is followed by Wachsmuth et al. (2017), which will be in the focus of this paper. On this account, it is useful to distinguish three main *levels of granularity*:

Argument Unit. The argument unit is the finest level of granularity, it refers to individual premises or conclusions, that is, individual statements that can be part of an argument (used, e.g., in Trautmann et al. 2020).

Argument. An argument is a set of propositions such that one of it is the conclusion, and one or several others are the premises. The relationship is such that the truth of the premises necessitates, or makes plausible, rational, or likely the truth of the conclusion (Siegel, 2024, 473)[2].

Argumentation. An argumentation typically contains a number of arguments on a particular topic. Wachsmuth et al. (2017) differentiate between a *Monological Argumentation* – presenting multiple arguments from one side of the debate – and a *Dialogical Argumentation* (also called debate), which in addition to a monological argumentation contains counterarguments presented from at least one contrary perspective. The debate is held together by its overarching topic and the history of back and fourth between opposing views on that topic. In contrast, in the literature from informal logic and philosophy, "argumentation" is usually only used in the sense of dialogical argumentation[3], whereas what Wachsmuth et al. (2017) refer to as *monological argumentation* would simply be considered a *long or complex argument*. From a theoretical point of view, it seems that Wachsmuth et al. (2017) under-represent the significance of dialectic in argumentation, yet for reasons of readability, we still follow their terminology in this regard.

[2]Siegel (2024) also points out that argument, as it is currently used, suffers from a process-product ambiguity: It refers both to the process of rationally defending a position and the products of this process; our definition emphasizes the product side, as this is central for CA.

[3]See, e.g., Dutilh Novaes (2022): "Argumentation can be defined as the communicative activity of producing and exchanging reasons in order to support claims or defend/challenge positions [...] It is arguably best conceived as a kind of dialogue [...]".

2 From Logic to Dialectic in Three Steps

In this section, we introduce central concepts of argumentation theory that underpin much of current concepts of CA. We start with introducing the background from classical philosophical logic, and we sketch informal logic's initial reaction to it (2.1). Then, we delineate how this reaction was systematized (2.2), and we conclude by observing a development towards dialectic in the field (at the cost, ironically, of logic), a development that we consider well-founded (2.3).

2.1 From Soundness to The ARS Scheme

The systematic pursuit of logic and argumentation theory has been initiated in western philosophical thought with the work of Aristotle, approximately 400 B.C. He introduced *syllogistic logic*, a systematic way to identify deductively valid inferences in a very restricted search space (Aristotle, 1984). For many centuries, this notion of **deductively valid inference** has dominated the discussion. An inference is deductively valid if it is *necessary* that the conclusion of an argument is true *if* its premises are true; in other words, in a deductively valid argument, if you know that the premises are true, then you also know with certainty that its conclusion is true. Within deductively valid inferences, the species of formally deductively valid inferences (in short: **formally valid inferences**) dominated. Formally valid inferences are such that their validity depends on the form of the propositions involved rather than on the meaning of the concepts. For instance, Example (1) is valid regardless of whether you replace *President Biden*, *cat*, or *fish* with any other concept[4].

(1) President Biden is a cat. All cats like fish. Therefore, President Biden likes fish.

Historically, the works of Frege (1879) and then Russell (1905) have provided arguably the first revolutionary technical advancements in the realm of formal logic since its foundation by introducing and refining the predicate calculus.

The ideal of this tradition is the *sound* argument (or inference): An argument that is valid and whose premises are also *true*. In combination, this guarantees the truth of the conclusion (note that this

[4]Note, however, that there are not only formally deductively valid inferences, but also materially deductively valid inferences, where the validity of the inference depends on the meaning of the concepts involved. See Gubelmann et al. (2024) for a recent overview of this terminology in the field.

also means that a deductively valid inference can have patently false premises, as in (1)). Example (2) presents one such case of a sound argument.

(2) A biological kind belongs to the class of Mammalia ("are mammals") if female individuals of the kind normally produce milk to feed their offspring. Whales are mammals. Therefore, female whales normally produce milk to feed their offspring.

Roughly since the end of WW2, some scholars have become dissatisfied with this general approach towards logical reasoning and argumentation. In brief, they thought that being deductively sound is neither necessary nor sufficient for an argument to be good, reasonable, or reliable. For instance, our Example (3), while clearly being a deductively sound argument (i.e., it is deductively valid and its premises are true), will not be considered a good argument by any standard: Its premise and its conclusion are identical, which means that there is no chance that it will convince any audience of its conclusion that does not already accept it.

(3) All cats are mammals, therefore, all cats are mammals.

In contrast, Example (4) is arguably a good argument, where the truth of the premise gives reason to believe in the truth of the conclusion, even though the truth of the premise does not necessitate the its conclusion, which means that it is not deductively valid, let alone sound. To see why it is not deductively valid, note that it is perfectly possible that the sticker could erroneously have been applied to a non-organic orange, thus creating a situation where the premise is true, while the conclusion is likely (but not necessarily) false.

(4) There is an "Organic"-sticker on that Orange, therefore this orange has been produced organically, therefore, it will contain less traces of pesticides.

These examples prefigure two major ways in which this new approach to arguments and logic, mostly called **informal logic** (but also argumentation theory, or sometimes also simply critical thinking), diverged from the traditional branch: (i) It was interested in the *credibility*, *acceptability*, or *reliability* of the premises as opposed to in their truth, and (ii) it emphasized a kind of validity of inferences where the truth of the premises does not necessitate, but does make reasonable or plausible, or likely the truth of the conclusion.

This kind of validity is called by different names, such as abductively valid (Lipton, 1991), defeasibly valid (Leslie, 2007), or inductively valid (Wilbanks 2010). We here settle for the final and most common option: **inductive validity**. As can be seen in Example (4), there can be arguments where the premises are not certain to be true, and where the truth of the premises does not guarantee the truth of the conclusion, but which are, in normal circumstances, good arguments. For instance, assuming you are in a store that is generally trustworthy, when you see an "organic" sticker on an orange, that makes it reasonable to assume that it is an organically produced orange (but it's not certain to be true), and this, again, makes it reasonable (but not certain!) that it contains less pesticides. The interested reader may refer to Gubelmann et al. (2024) for more details on how these concepts have influenced NLP and natural language inference in particular.

This basic intuition has then been systematized into the so-called *ARS-Scheme* that has explicitly been proposed to provide an umbrella conception for several conceptions of soundness proposed by informal logic (Johnson 2009, xiii and Blair 2015, 36-37): a good argument should be one whose premises are **A**cceptable (but, *pace* soundness, not necessarily known to be true), that are individually **R**elevant and jointly **S**ufficient to support the conclusion (without, *pace* soundness, necessarily necessitating its truth).

2.2 Complementing Logic with Rhetoric and Dialectic

The research programs of argumentation theory and informal logic (compare Groarke 2024 for an overview of the field)[5] have been motivated with this basic conviction that not all (see Example (3)) and not only (see Example (4)) deductively sound arguments are good arguments. Researchers working in these research programs then developed a tripartite systematic of **logic**, **dialectic**, and **rhetoric**. Perelman (1971) is among the first to propose the

[5] In terms of the topics that it addresses, informal logic is largely coextensive with argumentation theory, while it might be distinguished by a particular emphasis on examining and teaching reasonableness of arguments beyond deductively valid and sound ones.

distinction[6]. According to him, logic does not concern itself with acceptability, or with changing one's counterpart's point of view. This is what **dialectic** does (p. 2): it *"begins from theses that are generally accepted with the purpose of gaining acceptance of other theses which could be or are controversial"* (cited after Johnson 2009, 2). The dialectical perspective enjoys a place of prominence in current argumentation theory (Finocchiaro 1987, 13 and Johnson 2009, 2), however, it is also the most difficult to define from among the three concepts (Rescher, 1977, xi). What is common to all conceptions of the dialectical perspective is that it involves two different positions, typically assumed by two different persons, and that it focuses on argumentation (see Section 1), that is, it spans a number of individual arguments.

The **logical** perspective is then to do with the support that premises can lend to conclusions (however, as we show in the following Section 2.3, the significance of the logical aspect has decreased to the benefit of the dialectical aspect).

Finally, **rhetoric** pertains to the persuasive effect that an argument can have on people or audiences – independently of how it was evaluated on the logical or dialectical dimensions. As previously mentioned (see Section 1), due to the scope of this work we will set aside this third perspective in the present contribution, while acknowledging that it would be an essential topic to be addressed by future work.

2.3 Why Dialectic has become more Important to Informal Logic than Logic

Arguably the most complex conceptual terrain surrounds the concept *logic*. As detailed above, in classical argumentation theory, it denotes the logical aspect, as opposed to the dialectical and rhetorical aspect. Outside of argumentation theory, logic primarily refers to the study of the formal-deductive validity or invalidity of different kinds of inferences (see Section 2.1). Furthermore, a research tradition within argumentation theory refers to itself as *informal logic* (see Section 2.1).

In recent years, the field has seen a number of distinguished voices emphasizing the importance, or even the priority of the dialectical aspect. For instance, Van Eemeren (2015, 5) notes that informal logic focuses on the dialectical and rhetorical aspect at the cost of the logical aspect.

This focus on the dialectical aspect is also observable on one of informal logic's central conceptual products, the so-called ARS-scheme (see above, Section 2.1), emphasizing the idea that good arguments should have premises that are *acceptable*, *relevant*, and *sufficient* to support the conclusion. The central importance of the dialectical aspect for informal logic can be seen from the fact that Blair (2012, 93,95,97) emphasizes the central importance of context and target audience for all three criteria: a premise should be acceptable and relevant for a specific target audience in a given situation, and the premises should jointly be sufficient to support the conclusion, again, for the relevant target audience in a given situation. Freeman (1991, 93ff.), in his book-length study of the macrostructure of arguments, also squarely locates the acceptability condition within a dialectical perspective (compare also Goldman 1994 on this topic). In short, the criteria specified in the ARS scheme have come to be considered to belong to the dialectical aspect.

We emphasize that this shift towards the dialectical perspective in the literature seems correct: It always depends on the specific situation, in particular on the *target audience* involved, whether or not a premises is acceptable, whether or not it is individually relevant, and, in conjunction with the other premises stated, sufficient to inductively entail the conclusion. For instance, regardless of its truth, some target audiences might not accept that whales are mammals, thus refusing to accept one of the premises in Example (2). In contrast, another more zoologically educated audience might accept that premise without any further argument. Similarly, to a generally suspicious audience, the premises in Example (4) might not be sufficient to support the conclusion (for instance, they might generally suspect that the big food industry always lies to them). Indeed, it might even be that they consider any "Organic" stickers on an orange as so untrustworthy as to be irrelevant to the conclusion. This illustrates that acceptability, relevance, and sufficiency are always relative to an argument's general context, and to its target audience in particular.

We note that this distinguishes informal logic from traditional formal logic: whether an inference is formal-deductively valid is independent of any target audience; to simplify slightly, whoever does not accept the validity of a deductively valid inference thereby shows either a lack of understanding

[6] Wenzel (1990) also emphasizes these three perspectives, arguing that they are equally valid ways to conceive and evaluate arguments.

of the respective language, or shows that she is speaking her own idiolect.

We now focus on the quality dimensions that have been developed in argumentation theory and that are particularly relevant for the field of CA. At root, the three different perspectives on argumentative discourse (i.e., logic, dialectic, rhetoric) come with three different sets of norms for evaluating the quality of arguments (Blair, 2012; Siegel, 2024). The *logical* perspective evaluates the epistemic validity of a given argument, that is, the extent to which the premise(s) actually support(s) the conclusion. The *dialectical* perspective, in contrast, asks for the rational acceptability of the premises, and hence, in the case of a valid argument, the support that they can lend the conclusion in a given dialectical setting. In the *rhetorical* realm, finally, persuasion is king: Even an invalid argument that comes with untenable conclusion might get high scores if it convinces the crowd.

2.4 Two Theoretical Takeaways

We draw two implications from our theoretical-historical overview:

Isolated Arguments Cannot be Dialectically Assessed. It is generally not possible to evaluate arguments (as opposed to argumentations) such as Example (1) using categories taken from the dialectical perspective. These categories require that one sees how an argumentative text engages the opposite side and tries to win it over using rational means; in other words, it requires an argumentation, or an argument with sufficient context to know its target audience. Thus, it makes little theoretical sense to try to evaluate arguments, considered in isolation without any context, by means of rhetorical or dialectical perspectives and norms. Only when we know the two (or more) sides involved in an *argumentation* with a dialectical structure will we be able to judge its dialectical value. An argument might be excellent in a certain argumentation with a given audience, but entirely useless in another argumentation with another audience. Without any context, the dialectical value of an argument is simply unknown. Therefore, we maintain that, typically, it is *not theoretically sound* to consider an *argument* in the sense specified above (Section 1), such as Example (1), from a *dialectical perspective* when taken in isolation. In this sense, there is an intimate connection between the level of granularity examined and the perspectives that are available for the examination.

The ARS Scheme is at Home in Dialectic. Second, we emphasize that one of the central quality tests for individual arguments in argumentation theory – the ARS-scheme – is designed to be applied in the dialectical aspect. It should be understood to ask whether the premises of an argument are relevant, sufficient, and acceptable *given a specific context and target audience*.

3 Quality Dimensions in Computational Argumentation

Next, we take a closer look at how the quality dimensions – logic and dialectic – have been addressed and annotated by the literature on CA.

Brief overview of the development. Over the last 15 years, the research in CA has been growing and researchers have explored various quality dimensions, typically chosen based on prior work on argumentation. A series of publications by Persing and Ng, for instance, have taken a closer look at essays and debate comments in terms of their organization (2010), thesis clarity (2013), prompt adherence (2014), strength (2015), and persuasiveness (2017). While their work initially focused rather on the structural aspects of student essays, over time this has shifted towards argument-specific aspects (e.g., persuasiveness). With the change in the focus of the analysis, researchers also explored other types of texts such as forum posts, reviews, or online recommendations, which at that time aligned well with the increasing amount of textual data available on various websites. The field took interest in the sentiment of texts (e.g., Walker et al., 2012; Wachsmuth et al., 2014), their persuasiveness (e.g., Tan et al., 2016; Persing and Ng, 2017; Habernal and Gurevych, 2016a,b), and their convincingness (e.g., Gleize et al., 2019; Toledo et al., 2019.

Following, Wachsmuth et al. (2017) defined the first common ground for the work on argumentation quality assessment in the field of natural language processing. By basing their taxonomy on prior research by Blair (2011), the authors included the quality dimensions, which have previously been addressed in the NLP literature, and further extended the range of the taxonomy to a total of three main dimensions (i.e., cogency/logic, reasonableness/dialectic, effectiveness/rhetoric) and 11 subdimensions. Here, the *effectiveness* examines the rhetorical quality of an argumentative text by assessing its clarity, credibility, appropriateness, emo-

tional appeal, and arrangement. Further, *cogency* refers to the logical soundness and validity of an argument, while *reasonableness* concerns the dialectical aspects and thus asks whether an argument is able to "target the resolution of differences of opinions" (Wachsmuth et al., 2017). Both, logic and dialectic, incorporate three sub-dimensions each (i.e., acceptability, relevance, sufficiency), yet the former regards them locally and the latter globally.

Annotated Datasets. Since its publication, Wachsmuth et al.'s taxonomy (2017) has been frequently applied in the field (see e.g., Ng et al., 2020; Gienapp et al., 2020; Marro et al., 2022). At the time of writing the taxonomy is still the most extensive and formally structured representation of quality dimensions in the field of CA and is thus at the core of our analysis. To better understand how it has been used in practice, we explore annotated datasets using the taxonomy as a concrete example. In particular, we take interest in the annotation of the logical and the dialectical dimensions to assess whether the theoretical differentiation between them (or lack thereof, see Section 2.4) also shows in the existing annotations.

Our first example has been created by the authors of the taxonomy themselves. In their own analysis, Wachsmuth et al. (2017) point out that the highest correlation between quality dimensions is between the local and the respective global acceptability and relevance. The local sufficiency correlates the most with logic, yet if the three main quality dimensions (i.e., logic, dialectic, rhetoric) are excluded from the correlation analysis, the local sufficiency also has the highest correlation with its global counterpart. We take a closer look at the quality scores assigned to the pairs of local and global dimensions by the individual annotators. Table 1 depicts the assignment of the same score for local and their global counterpart dimensions per annotator in percent. We notice that one of the annotators (i.e., A1) has assigned the same scores for all of the three pairs in more than 80% of the time. While we do not observe such extremities for the other two annotators, A2 assigned the same sufficiency score for local and global in 82.19% of the annotations and A3 in 83.75% of the annotations for acceptability. In addition, we assess the Pearson's Correlation Coefficient for the individual annotators and discovered similar results (see Appendix A for further details). The observed correlation and assignment of same score may be an indication that the annotators have difficulty differentiating between the local and global dimensions. We hypothesize that this could be because annotators are projecting their own dialectical context in the absence of a predefined one.

Table 1: Assignment of same score for local and global counterpart dimensions per annotator (i.e., A1, A2, A3).

	Acceptability	Relevance	Sufficiency
A1	**85.62%**	**82.50%**	**81.88%**
A2	56.25%	45.94%	**82.19%**
A3	**83.75%**	65.31%	69.69%

Following the taxonomy and annotation guidelines presented by Wachsmuth et al. (2017), Dumani and Schenkel (2020) created another dataset, yet only the three main dimensions - logic, dialectic, and rhetoric - have been annotated. Here, two people annotated premises (and sometimes arguments) for each of the three main dimensions. To better understand how well annotators differentiate between the local and global dimensions, we primarily compare their annotations for logic and dialectic, yet also include an analysis on rhetoric for completeness and comparison. In cases where a text was not argumentative (e.g., "I accept") or the annotators did not know which score to assign for other reasons, the first annotator mostly assigned the lowest possible value (i.e., "1") and the second annotator mostly assigned the "cannot judge" value (i.e., "?"). We exclude the "?" values from the dataset for further analysis. In the case of annotator one we remove only 1 annotation, whereas for annotator two this results in the removal of 278 out of 1376 annotations (i.e., 20%). Table 2 and 3 show the assignment of same score for the three main quality dimensions and all of them for the first and the second annotator respectively. We notice that also in this dataset the second annotator assigned the same values for logic and dialectic for over 84% of the annotations, however they also assigned the same values for all three quality dimensions for 75% of the texts.

Table 2: Assignment of same score for quality dimensions for annotator 1 (excluding '?' values).

	Logic	Dialectic	Rhetoric	All
Logic	-	64.61%	72.31%	-
Dialectic	64.61%	-	68.02%	-
Rhetoric	72.31%	68.02%	-	-
All	-	-	-	55.01%

Table 3: Assignment of same score for quality dimensions for annotator 2 (excluding '?' values).

	Logic	Dialectic	Rhetoric	All
Logic	-	**84.08%**	**88.15%**	-
Dialectic	**84.08%**	-	77.91%	-
Rhetoric	**88.15%**	77.91%	-	-
All	-	-	-	75.07%

Table 4: Assignment of same relative preference (e.g., A>B) for quality dimensions by multiple annotators via crowdsourcing.

	Logic	Dialectic	Rhetoric	All
Logic	-	44.02%	44.68%	-
Dialectic	44.02%	-	44.80%	-
Rhetoric	44.68%	44.80%	-	-
All	-	-	-	23.70%

Table 5: Assignment of relative preference (e.g., A>B) for quality dimensions by multiple annotators (i.e., crowdsourcing) excluding "Ties" (i.e., A=B).

	Logic	Dialectic	Rhetoric	All
Logic	-	57.08%	58.57%	-
Dialectic	57.08%	-	59.26%	-
Rhetoric	58.57%	59.26%	-	-
All	-	-	-	37.45%

In our last example[7] Gienapp et al. (2020) also annotated only the three main quality dimensions - logic, dialectic, and rhetoric - yet chose a relative approach (i.e., A>B, B>A, A=B) and used crowdsourcing for the annotation. This means that a crowd worker (i.e., an annotator, who does not necessarily have background knowledge in argumentation) would view a pair of arguments at a time and make a decision for one particular quality dimension. As such, the three quality dimensions for each pair of arguments were not (necessarily) annotated by the same crowd worker. Table 4 shows the percentage of argument pairs for which the same score was assigned, while the Table 5 shows the same but excludes cases, where one of the scores was a "Tie" (i.e., A=B). The percentages here are noticeably lower than what we have observed for the previous two datasets. We would like to especially point out the low percentage of cases, where the same score was assigned to all three dimensions (i.e., 23.70% and 37.45% respectively). At first sight, this could be taken as a case in point against our worry that, on the argument level, it is generally not possible to apply the quality dimension of dialectic separate from logic. However, upon closer inspection, this is not the case. In this dataset, the three quality dimensions for each pair of arguments were not (necessarily) annotated by the same crowd worker. It is therefore not clear whether the low correlation scores are due to the ability of individual annotators to distinguish between the logical and dialectical dimension, or whether they are due to the different annotation approach (where one annotator assesses only one dimension at a time).

Our analysis of the datasets and their annotation show that the same scores were frequently assigned for *local* and their *global* counterpart dimensions, thus also to logic and dialectic. Gienapp et al. (2020) and Wachsmuth et al. (2017) discuss the correlation of the dimensions. These correlations for both papers slightly contradict each other, which is also what we observe in our analysis. Gienapp et al. (2020) attribute the differences to the choices in sample sizes (i.e., 320 arguments vs. 41 859 pairs) and annotation methodologies (i.e., absolute vs. relative), and point out that *"dialectical quality appears to correlate slightly more with rhetorical quality in our corpus, but with logical quality in their [Wachsmuth et al. (2017)'s] data"* (Gienapp et al., 2020). Considering the generally observed higher agreement scores and thus quality of relative annotation approaches by Gienapp et al. (2020), this lower overlap of the scores may be an indication that the cause for the high correlations between the dimensions may be due to a difficulty in the differentiation between them. However, subjectivity in the assessment of argument quality is a known issue, which may also have an influence on the assigned scores (Lapesa et al., 2023). Therefore, we turn to the analysis of the levels of granularity used for the individual quality dimensions.

Levels of granularity. While the levels of granularity (i.e., argument unit, argument, monological argumentation, dialogical argumentation, see Section 1) are discussed in the context of the surveyed prior work in the field, they are only implicitly included in the taxonomy proposed by Wachsmuth et al. (2017). Based on the choice of wording for the initial (sub-)categories, we deduce that logic is to be evaluated on an argument unit level or an

[7] Marro et al. (2022) also explore the three quality dimensions, yet in the case of *dialectic* they follow the definition of Stapleton and Wu (2015) and thus measure it through the subcategories "rebuttal" and "counterargument". Further, according to the annotation guidelines dialectic *"will only be annotated if there are counterarguments, represented as Claim(s) attacking the Main Claim, or Claim(s) attacking other Claim(s), present in the persuasive essay"*. As only a subset of the dataset includes dialectic and it is defined differently, we did not find this dataset annotation to be suitable for the purpose of our analysis.

argument level, dialectic on an argumentation level, and rhetoric on an argumentation level. Further, in all three datasets analyzed, all quality dimensions were annotated for the same text units.

We provide an even more refined overview of the levels of granularity, which individual datasets have used for the annotation by also including datasets that have not targeted logic, dialectic, and rhetoric simultaneously, yet have annotated at least one (sub-) dimension (see Table 6 for an overview)[8]. For the majority of the dimensions, prior work has annotated arguments, followed by some cases of argument units. We discovered only one annotation considering dialectic on an argumentation level (El Baff et al., 2018). They look at persuasiveness from a dialectical perspective, considering factors such as prior beliefs of the reader and the author, and differentiating themselves from prior work, which they view as relying on subjective assessment.

The definitions of the dimensions and the corresponding annotation guidelines used by prior work do not necessarily imply that one should expect a direct dependency between the scores for the logical/local and the dialectical/global quality dimensions. As a consequence, we suggest that the assignment of the same score may originate from the definition of the dimensions themselves and/or the use of a single level of granularity to assess the quality for all three dimensions.

4 Recommendations for Future Work in Computational Argumentation

Based on the development of logic and dialectic in argumentation theory (Section 2), our theoretical takeaways (Section 2.4), and our analysis of the most popular taxonomy for CA (Section 3), we introduce recommendations for future work.

Quality-Dimensions Should be Tailored to Granularity. The first main difference between logic and dialectic for the computational assessment of argument quality is that the dialectical dimension cannot in general be applied below the granularity level of argumentation. The logical dimension can be used to *analyze a monological text, even a single argument, in isolation*. dialectic requires at least a *pre-defined audience and a context* (in the case of a monological argumentation) and potentially also the *counter-position alongside the argument* (in the case of a dialogical argumentation). As such, *arguments* should be evaluated only in terms of their logic, while *argumentation* can be assessed in both dimensions - in the case of logic one would analyze the individual arguments within the argumentation. Similarly, an *argument unit*, which can be a premise or a conclusion, cannot be assessed regarding logic or dialectic. However, we acknowledge the fact that some sources such as a debate forums follow a structure where a conclusion is given first and then forum users are merely expected to provide premises that support it, thus the premise alongside the conclusion could form an argument and thus also be analyzed regarding logic. This theoretical finding aligns nicely with the empirical data that suggest that annotators struggle with distinguishing the logical from the dialectical dimension when only presented with arguments (as opposed to argumentations).

Acceptability Should be Confined to Dialectic. The differentiation between logic/local and dialectic/global induces confusion due to the complex conceptual landscape surrounding "logic" (see above, Section 2). In the context of *logic*, as conceived by Wachsmuth et al. (2017), relevance and sufficiency resemble the extant necessary-sufficient couple, as classically conceived in formal logic: In a deductively valid argument, the truth of the premise is sufficient for the truth of the conclusion, while the truth of the conclusion is necessary for the truth of the premise. However, for Wachsmuth et al., both relevance and sufficiency pertain to the support that premises might lend to the conclusion: are the premises individually relevant and jointly sufficient to inductively support the conclusion?

More importantly, when the authors suggest that acceptability is also acknowledged to cover the logical quality of arguments, this confuses the logical with the dialectical dimension. More particularly, according to the annotation guidelines provided by Wachsmuth et al. (2017), acceptability evaluates whether a premise of an argument (i.e., an argument unit) "is worthy of being believed". It is problematic to assess belief-worthiness as such, independently of any target audience, as Blair points out that *"depending on the type of premise and the circumstances of the argument, from the recipient's vantage point the norms of acceptability will vary"*. Thus, to improve upon the framework proposed by Wachsmuth et al. (2017), we suggest to entirely refrain from assessing acceptability as a sub-category

[8] As the focus of this work is on logic and dialectic, we provide the levels of granularity for rhetoric in Appendix B (see Table 14).

Table 6: Levels of granularity in annotated datasets for CA for the dimensions of logic and dialectic.

Main Dimension	Sub-Dimension	Granularity	Dataset
Overall AQ		Argument Unit	Swanson et al. (2015)
		Argument	Swanson et al. (2015); Wei et al. (2016); Wachsmuth et al. (2017); Ng et al. (2020)
Logic / Cogency		Argument Unit	Dumani and Schenkel (2020)
		Argument	Wachsmuth et al. (2017); Gienapp et al. (2020); Ng et al. (2020); Marro et al. (2022)
	L.Acceptability	Argument	Wachsmuth et al. (2017)
	L.Relevance	Argument Unit	Braunstain et al. (2016); Dumani and Schenkel (2019)
		Argument	Wachsmuth et al. (2017)
	L.Sufficiency	Argument	Wachsmuth et al. (2017); Stab and Gurevych (2017)
Dialectic/ Reasonableness		Argument Unit	Dumani and Schenkel (2020)
		Argument	Wachsmuth et al. (2017); Gienapp et al. (2020); Ng et al. (2020); Marro et al. (2022)
		Argumentation	El Baff et al. (2018)
	G.Acceptability	Argument	Wachsmuth et al. (2017)
	G.Relevance	Argument	Wachsmuth et al. (2017); Toledo et al. (2019); Gretz et al. (2020); Joshi et al. (2023)
	G.Sufficiency	Argument	Wachsmuth et al. (2017)

of logic (thus confining it to dialectic) and to evaluate logic in terms of relevance and sufficiency only.

5 Conclusion

In this work we revisited Wachsmuth et al.'s taxonomy (2017) for computational argument quality assessment and analyzed it in terms of its philosophical foundations and its practical applications for the creation of annotated datasets for computational assessment. As a result, we suggest for future work to: i) Pay attention to the texts' granularity – while logic can be used to assess individual arguments or potentially even argument units, dialectic can only be assessed in connection to a target audience; and ii) Confine the sub-dimension of acceptability to the dimension of dialectic, as it requires a context and a target audience for proper consideration.

Limitations

In this article we aim to provide a comprehensive and in-depth analysis of the dimensions of logic and dialectic. While the rhetorical dimension is by no means less important, it remains beyond the scope of this work. We believe that due to its high number of sub-dimensions and due to the considerable differences between rhetoric in a monological and in a dialogical context, an in-depth analysis of rhetoric should be addressed in a separate work.

Acknowledgements

We would like to thank Henning Wachsmuth, Lorik Dumani, and Lukas Gienapp for providing us with further details on their work in regards to the annotation process and the definitions for the individual quality dimensions, thus allowing us to build on insight from their work.

This research was funded in part by the Swiss National Science Foundation (SNSF) [Grant number 208220] and by the Digital Society Initiative of the University of Zurich.

References

Alaa Alhamzeh. 2023. Financial argument quality assessment in earnings conference calls. In *International Conference on Database and Expert Systems Applications*, pages 65–81. Springer.

Aristotle. 1984. Prior Analytics. In Jonathan Barnes, editor, *The Complete Works of Aristotle*, pages 39–113. Oxford: Oxford University Press.

J Anthony Blair. 2011. *Groundwork in the theory of argumentation: Selected papers of J. Anthony Blair*, volume 21. Springer Science & Business Media.

J Anthony Blair. 2012. Relevance, acceptability and sufficiency today. *Groundwork in the Theory of Argumentation: Selected Papers of J. Anthony Blair*, pages 87–100.

J Anthony Blair. 2015. What is informal logic? In *Reflections on Theoretical Issues in Argumentation Theory*, pages 27–42. Springer.

Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. 2016. Supporting human

answers for advice-seeking questions in cqa sites. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 129–141. Springer.

Lorik Dumani and Ralf Schenkel. 2019. A systematic comparison of methods for finding good premises for claims. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 957–960.

Lorik Dumani and Ralf Schenkel. 2020. Quality-aware ranking of arguments. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 335–344.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. The role of pragmatic and discourse context in determining argument impact. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5668–5678.

Catarina Dutilh Novaes. 2022. Argument and argumentation. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, fall 2022 edition. Metaphysics Research Lab, Stanford University.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464.

Neele Falk, Eva Maria Vecchi, Iman Jundi, and Gabriella Lapesa. 2024. Moderation in the wild: Investigating user-driven moderation in online discussions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 992–1013.

Maurice A Finocchiaro. 1987. A historical approach in the study of argumentation. *by FH van Eemeren, R. Grootendorst, JA Blair, and Ch. A. Willard. Dordrecht*, pages 81–91.

James B. Freeman. 1991. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. Number 10 in Studies of Argumentation in Pragmatics and Discourse Analysis. Foris Publications, Berlin ; New York.

Gottlob Frege. 1879. *Begriffsschrift, Eine Der Arithmetischen Nachgebildete Formelsprache Des Reinen Denkens*. Verlag von Louis Nebert.

Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient pairwise annotation of argument quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5772–5781.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976.

Alvin I Goldman. 1994. Argumentation and social epistemology. *The Journal of Philosophy*, 91(1):27–49.

Shai Gretz, 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Leo Groarke. 2024. Informal Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, spring 2024 edition. Metaphysics Research Lab, Stanford University.

Reto Gubelmann, Ioannis Katis, Christina Niklaus, and Siegfried Handschuh. 2024. Capturing the varieties of natural language inference: A systematic survey of existing datasets and two novel benchmarks. *Journal of Logic, Language and Information*, 33(1):21–48.

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

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

Ralph H Johnson. 2009. Revisiting the Logical/Dialectical/Rhetorical Triumvirate.

Omkar Joshi, Priya Pitre, and Yashodhara Haribhakta. 2023. Arganalysis35k: A large-scale dataset for argument quality analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13916–13931.

Gabriella Lapesa, Eva Maria Vecchi, Serena Villata, and Henning Wachsmuth. 2023. Mining, assessing, and improving arguments in nlp and the social sciences. In *The 17th Conference of the European Chapter of the Association for Computational Linguistics*, page 1.

Sarah-Jane Leslie. 2007. Generics and the structure of the mind. *Philosophical Perspectives*, 21:375–403.

Peter Lipton. 1991. *Inference to the Best Explanation*. Routledge.

Santiago Marro, Elena Cabrio, and Serena Villata. 2022. Graph embeddings for argumentation quality assessment. In *EMNLP 2022-Conference on Empirical Methods in Natural Language Processing*.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126.

Chaim Perelman. 1971. *The New Rhetoric*. Springer.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *26th International Joint Conference on Artificial Intelligence*, pages 4082–4088.

Nicholas Rescher. 1977. *Dialectics: A Controversy-Oriented Approach to the Theory of Knowledge*. Suny Press.

Bertrand Russell. 1905. On denoting. *Mind*, 14(56):479–493.

Harvey Siegel. 2024. Arguing with Arguments: Argument Quality, Argumentative Norms, and the Strengths of the Epistemic Theory. *Informal Logic*, 43(4):465–526.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990.

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.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 217–226.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pages 613–624.

Assaf Toledo, 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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-Grained Argument Unit Recognition and Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9048–9056.

Frans H Van Eemeren. 2015. Bingo! Promising developments in argumentation theory. *Reasonableness and Effectiveness in Argumentative Discourse: Fifty Contributions to the Development of Pragma-Dialectics*, pages 55–77.

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

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II 15*, pages 115–127. Springer.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, volume 12, pages 812–817. Istanbul, Turkey.

Zhongyu Wei, Yandi Xia, Chen Li, Yang Liu, Zachary Stallbohm, Yi Li, and Yang Jin. 2016. A preliminary study of disputation behavior in online debating forum. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 166–171.

Joseph W Wenzel. 1990. Three perspectives on argument: Rhetoric, dialectic, logic. *Perspectives on argumentation: Essays in honor of Wayne Brockriede*, pages 9–26.

Jan J. Wilbanks. 2010. Defining Deduction, Induction, and Validity. *Argumentation*, 24(1):107–124.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363.

A Pearson's Correlation Coefficient

In this Appendix Section, we present the Pearson's correlation coefficient for each of the annotated datasets discussed in Section 3. Wherever the data is available, we calculate the coefficient for each annotator separately. The highest values per row (diagonals excluded) are represented in bold.

A.1 Wachsmuth et al. (2017)

Table 7: Pearson's correlation coefficient for the sub-dimensions of logic and dialectic for Annotator 1.

Metric	Local Acceptability	Global Acceptability	Local Relevance	Global Relevance	Local Sufficiency	Global Sufficiency
Local Acceptability	1.00	**0.89**	0.70	0.71	0.81	0.75
Global Acceptability	**0.89**	1.00	0.65	0.69	0.78	0.75
Local Relevance	0.70	0.65	1.00	**0.79**	0.62	0.60
Global Relevance	0.71	0.69	**0.79**	1.00	0.63	0.63
Local Sufficiency	**0.81**	0.78	0.62	0.63	1.00	0.80
Global Sufficiency	0.75	0.75	0.60	0.63	**0.80**	1.00

Table 8: Pearson's correlation coefficient for the sub-dimensions of logic and dialectic for Annotator 2.

Metric	Local Acceptability	Global Acceptability	Local Relevance	Global Relevance	Local Sufficiency	Global Sufficiency
Local Acceptability	1.00	0.48	0.46	0.47	**0.62**	0.53
Global Acceptability	0.48	1.00	0.43	0.47	0.57	**0.59**
Local Relevance	0.46	0.43	1.00	**0.75**	0.48	0.46
Global Relevance	0.47	0.47	**0.75**	1.00	0.48	0.51
Local Sufficiency	**0.62**	0.57	0.48	0.48	1.00	0.57
Global Sufficiency	0.53	**0.59**	0.46	0.51	0.57	1.00

Table 9: Pearson's correlation coefficient for the sub-dimensions of logic and dialectic for Annotator 3.

Metric	Local Acceptability	Global Acceptability	Local Relevance	Global Relevance	Local Sufficiency	Global Sufficiency
Local Acceptability	1.00	**0.84**	0.64	0.64	0.74	0.54
Global Acceptability	**0.84**	1.00	0.64	0.67	0.71	0.55
Local Relevance	0.64	0.64	1.00	**0.72**	0.66	0.51
Global Relevance	0.64	0.67	**0.72**	1.00	0.71	0.61
Local Sufficiency	**0.74**	0.71	0.66	0.71	1.00	0.72
Global Sufficiency	0.54	0.55	0.51	0.61	**0.72**	1.00

A.2 Dumani and Schenkel (2020)

Table 10: Pearson's correlation coefficient for all dimensions for Annotator 1.

	Logic	Dialectic	Rhetoric
Logic	1.000	0.718	**0.765**
Dialectic	0.718	1.000	**0.813**
Rhetoric	0.765	**0.813**	1.000

Table 11: Pearson's correlation coefficient for all dimensions for Annotator 2.

	Logic	Dialectic	Rhetoric
Logic	1.000	0.930	**0.947**
Dialectic	**0.930**	1.000	0.906
Rhetoric	**0.947**	0.906	1.000

A.3 Gienapp et al. (2020)

Table 12: Pearson's correlation coefficient for all dimensions and all values (i.e., A, B, Tie).

	Logic	Dialectic	Rhetoric
Logic	1.000	0.083	**0.104**
Dialectic	0.083	1.000	**0.102**
Rhetoric	**0.104**	0.102	1.000

Table 13: Pearson's correlation coefficient for all dimensions for the values A and B, excluding Tie.

	Logic	Dialectic	Rhetoric
Logic	1.000	0.140	**0.170**
Dialectic	0.140	1.000	**0.184**
Rhetoric	0.170	**0.184**	1.000

B Levels of granularity - Rhetoric

Table 14: Levels of granularity in annotated datasets for CA for the dimension of rhetoric.

Main Dimension	**Sub-Dimension**	**Granularity**	**Dataset**
Rhetoric / Rhetoric		Argument Unit	Dumani and Schenkel (2020)
		Argument	Wachsmuth et al. (2017); Gienapp et al. (2020); Ng et al. (2020); Marro et al. (2022)
	Credibility	Argument	Wachsmuth et al. (2017)
	Appropriateness	Argument	Ziegenbein et al. (2023); Wachsmuth et al. (2017)
	Clarity	Argument	Wachsmuth et al. (2017)
		Argumentation	Persing and Ng (2013, 2014)
	Emotional Appeal	Argument Unit	Alhamzeh (2023)
		Argument	Walker et al. (2012); Wachsmuth et al. (2014, 2017); Falk et al. (2024)
		Dialogue	Walker et al. (2012); Falk et al. (2024)
	Arrangement	Argument	Wachsmuth et al. (2017)
	Arrangement	Argumentation	Persing et al. (2010)
	(Strength)	Argument Unit	Alhamzeh (2023)
		Argument	Persing and Ng (2015)
	(Winning Side)	Dialogue	Zhang et al. (2016)
	(Persuasiveness)	Argument Unit	Durmus et al. (2019)
		Argument	Tan et al. (2016); Persing and Ng (2017)
		Argumentation	Tan et al. (2016)
		Dialogue	Tan et al. (2016)
	(Convincingness)	Argument Unit	Gleize et al. (2019)
		Argument	Habernal and Gurevych (2016a,b); Toledo et al. (2019)