# Computational Argumentation Quality Assessment in Natural Language

**Henning Wachsmuth**
Bauhaus-Universität Weimar
Weimar, Germany
henning.wachsmuth@uni-weimar.de

**Nona Naderi**
University of Toronto
Toronto, Canada
nona@cs.toronto.edu

**Yufang Hou**
IBM Research
Dublin, Ireland
yhou@ie.ibm.com

**Yonatan Bilu**
IBM Research
Haifa, Israel
yonatanb@il.ibm.com

**Vinodkumar Prabhakaran**
Stanford University
Stanford, CA, USA
vinod@cs.stanford.edu

**Tim Alberdingk Thijm, Graeme Hirst**
University of Toronto
Toronto, Canada
{thijm, gh}@cs.toronto.edu

**Benno Stein**
Bauhaus-Universität Weimar
Weimar, Germany
benno.stein@uni-weimar.de

## Abstract

Research on computational argumentation faces the problem of how to automatically assess the quality of an argument or argumentation. While different quality dimensions have been approached in natural language processing, a common understanding of argumentation quality is still missing. This paper presents the first holistic work on computational argumentation quality in natural language. We comprehensively survey the diverse existing theories and approaches to assess logical, rhetorical, and dialectical quality dimensions, and we derive a systematic taxonomy from these. In addition, we provide a corpus with 320 arguments, annotated for all 15 dimensions in the taxonomy. Our results establish a common ground for research on computational argumentation quality assessment.

## 1 Introduction

What is a good argument? What premises should it be based on? When is argumentation persuasive? When is it reasonable? We subsume such questions under the term *argumentation quality*; they have driven logicians, rhetoricians, linguists, and argumentation theorists since the Ancient Greeks (Aristotle, 2007). Now that the area of computational argumentation is seeing an influx of research activity, the automatic assessment of argumentation quality is coming into the focus, due to its importance for envisioned applications such as writing support (Stab and Gurevych, 2014) and argument search (Wachsmuth et al., 2017), among others.

Existing research covers the mining of argument units (Al-Khatib et al., 2016), specific types of evidence (Rinott et al., 2015), and argumentative relations (Peldszus and Stede, 2015). Other works classify argumentation schemes (Feng et al., 2014) and frames (Naderi and Hirst, 2015), analyze overall argumentation structures (Wachsmuth et al., 2015), or generate claims (Bilu and Slonim, 2016). Also, theories of argumentation quality exist, and some quality dimensions have been assessed computationally (see Section 2 for details). Until now, however, the assertion of O'Keefe and Jackson (1995) that there is neither a general idea of what constitutes argumentation quality in natural language nor a clear definition of its dimensions still holds.

The reasons for this deficit originate in the varying goals of argumentation: persuading audiences, resolving disputes, achieving agreement, completing inquiries, and recommending actions (Tindale, 2007). As a result, diverse quality dimensions play a role, which relate to the logic of arguments, to the style and rhetorical effect of argumentation, or to its contribution to a discussion. Consider the following argument against the death penalty:[1]

*Everyone has an inalienable human right to life, even those who commit murder; sentencing a person to death and executing them violates that right.*

Although implicit, the conclusion about the death penalty seems sound in terms of (informal) logic, and the argument is clear from a linguistic viewpoint. Some people might not accept the first stated premise, though, especially if emotionally affected by some legal case at hand. Or, they might not be persuaded that the stated argument is the most relevant in the debate on death penalty.

This example reveals three central challenges: (1) Argumentation quality is assessed on different levels of granularity; (2) many quality dimensions are subjective, depending on preconceived opinions; and (3) overall argumentation quality seems hard to measure, as the impact and interaction of the different dimensions remain unclear.

[1] Taken from www.bbc.co.uk/ethics/capitalpunishment.

*Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

This paper does *not* propose a specific approach to assess quality; rather it defines a common ground by providing a so-far-missing holistic view on argumentation quality assessment in natural language. In particular, we first briefly but comprehensively survey all major theories and computational approaches for argumentation quality. Following Blair (2012), we distinguish three main quality aspects, each associated with several quality dimensions:

- *Logical quality* in terms of the cogency or strength of an argument.
- *Rhetorical quality* in terms of the persuasive effect of an argument or argumentation.
- *Dialectical quality* in terms of the reasonableness of argumentation for resolving issues.

We organize the survey along these aspects, discussing quality at four levels of granularity: (1) *argument unit*, i.e., a segment of text that takes the role of a premise or conclusion; (2) *argument*, i.e., a composition of premises and a conclusion, some of which may be implicit; (3) *(monological) argumentation*, i.e., a composition of arguments on a given issue; and (4) *(dialogical) debate*, i.e., a series of interacting argumentation on the same issue.

To unify and to consolidate existing research, we then derive a generally applicable taxonomy of argumentation quality from the survey. The taxonomy systematically decomposes quality assessment based on the interactions of 15 widely accepted quality dimensions (including the overall quality). Moreover, we provide a new annotated corpus with 320 arguments for which three experts assessed all 15 dimensions, resulting in over 14,000 annotations. Our analysis indicates how the dimensions interact and which of them are subjective, making the corpus an adequate benchmark for future research.

In summary, the contributions of this paper are:

1. *A comprehensive survey* of research on argumentation quality assessment (Section 2).
2. *A taxonomy* of all major quality dimensions of natural language argumentation, which clarifies their roles and dependencies (Section 3).
3. *An annotated corpus* for computational argumentation quality assessment (Section 4).[2]

## 2 Survey of Argumentation Quality

This section briefly surveys all major existing theories and the assessment of natural language argumentation quality. While we order the discussions along the three main quality aspects, we point out overlaps and interrelations where relevant.

### 2.1 Theories of Argumentation Quality

We focus on the major fields dealing with argumentation quality in natural language: argumentation theory and rhetoric. Table 1 gives an overview of the quality dimensions that we detail below.

**Logic** Formal argumentation studies the *soundness* of arguments, requiring the truth of an argument's premises and the deductive *validity* of inferring its conclusion. In case of inductive strength, the conclusion becomes probable given the premises. While sound arguments exist in natural language, most are defeasible in nature (Walton, 2006). The desired property of such arguments is *cogency*.

A cogent (or logically good) argument has individually acceptable premises that are relevant to the argument's conclusion and, together, sufficient to draw the conclusion (Johnson and Blair, 2006). Here, *(local) acceptability* means that a premise is rationally worthy of being believed by the target audience of the argument. It replaces truth, which is often unclear (Hamblin, 1970). A premise's *(local) relevance* refers to the level of support it provides for the conclusion, and *(local) sufficiency* captures whether the premises give enough reason to accept the conclusion. In the end, sufficiency thus presupposes relevance (Blair, 2012). While acceptability is more dialectical, overall the three dimensions of cogency are, with slight variations, acknowledged to cover the logical quality of arguments.

Damer (2009) adds that a good argument also depends on the rebuttal it gives to anticipated counterarguments (a dialectical property) as well as on its structural *well-formedness*, i.e., whether it is intrinsically consistent, avoids begging the question, and uses a valid inference rule. These dimensions adopt ideas from the argument model of Toulmin (1958), including rebuttals and warrants, and from the argumentation schemes of Walton et al. (2008), whose critical questions are meant to evaluate inference rules. While not focusing on quality, critical questions particularly help identify fallacies.

Introduced by Aristotle as invalid arguments, fallacies have been brought back to attention by Hamblin (1970). In general, a fallacy has some sort of error in reasoning (Tindale, 2007). Fallacies range from resorting to inapplicable evidence types or irrelevant premises to rhetoric-related errors, such

[2] The corpus is freely available at: http://www.arguana.com

| Aspect | Quality Dimension | Granularity | Sources |
|---|---|---|---|
| Logic | **Cogency** | Argument | Johnson and Blair (2006), Damer (2009), Govier (2010) |
| | Local relevance | Argument (unit) | Johnson and Blair (2006), Damer (2009), Govier (2010) |
| | Local sufficiency | Argument | Johnson and Blair (2006), Damer (2009), Govier (2010) |
| | Well-Formedness | Argument | Walton et al. (2008), Damer (2009) |
| Dialectic | Global sufficiency | Argument | Toulmin (1958), Damer (2009) |
| Dialectic | Local acceptability | Argument (unit) | Johnson and Blair (2006), Damer (2009), Govier (2010) |
| | **Fallaciousness** | Argument (unit) | Hamblin (1970), Tindale (2007), Walton et al. (2008) |
| | Local relevance | Argument (unit) | Hamblin (1970), Tindale (2007) |
| | Local sufficiency | Argument | Hamblin (1970), Tindale (2007) |
| | Validity | Argument | Hamblin (1970), Tindale (2007) |
| | Well-Formedness | Argument | Hamblin (1970), Tindale (2007) |
| | **Strength** | Argument | Perelman et al. (1969), Tindale (2007), Freeman (2011) |
| Rhetoric | **Effectiveness** | Argument(ation) | Perelman et al. (1969), O'Keefe and Jackson (1995) |
| | Arrangement | Argumentation | Aristotle (2007), Damer (2009) |
| | Appropriateness of style | Argumentation | Aristotle (2007) |
| | Clarity of style | Argumentation | Aristotle (2007), Tindale (2007), Govier (2010) |
| | Credibility | Argumentation | Aristotle (2007) |
| | Emotional appeal | Argumentation | Aristotle (2007), Govier (2010) |
| Logic | Soundness | Argument | Aristotle (2007) |
| Dialectic | **Convincingness** | Argumentation | Perelman et al. (1969) |
| | Global acceptability | Argument(ation) | Perelman et al. (1969) |
| | **Reasonableness** | Argumentation, debate | van Eemeren and Grootendorst (2004) |
| | Global acceptability | Argument(ation) | van Eemeren and Grootendorst (2004) |
| | Global relevance | Argument(ation) | van Eemeren and Grootendorst (2004), Walton (2006) |
| | **Global sufficiency** | Argumentation, debate | Cohen (2001) |

Table 1: Theoretical treatment of quality dimensions in the referenced sources for the given granularities of natural language argumentation, grouped by the aspect the bold-faced high-level dimensions refer to.

as unjustified appeals to emotion. They represent an alternative assessment of logical quality. Following Damer (2009), a fallacy can always be seen as a violation of one or more dimensions of good arguments. *Fallaciousness* negatively affects an argument's *strength* (Tindale, 2007).

Argument strength is often referred to, but its meaning remains unclear: "Is a strong argument an effective argument which gains the adherence of the audience, or is it a valid argument, which ought to gain it?" (Perelman et al., 1969). Tindale (2007) sees validity as a possible but not mandatory part of reasoning strength. Freeman (2011) speaks of the strength of support, matching the idea of inductive strength. Blair (2012) roughly equates strength with cogency, and Hoeken (2001) observes correlations between evidence strength and rhetorical persuasiveness. Such dependencies are expected, as the use of true and valid arguments represents one means of persuasion: logos (Aristotle, 2007).

**Rhetoric** Aristotle's work on rhetoric is one of the most systematic to this day. He defines rhetoric as the ability to know how to persuade (Aristotle, 2007). Besides logos, the three means of persuasion he sees include ethos, referring to the arguer's *credibility*, and pathos, the successful *emotional appeal* to the target audience. Govier (2010) outlines how emotions interfere with logic in arguments. Pathos is not necessarily reprehensible; it just aims for an emotional state adequate for persuasion.

In overall terms, rhetorical quality is reflected by the persuasive *effectiveness*, i.e., the success in persuading a target audience of a conclusion (Blair, 2012). It has been suggested that what arguments are considered as effective is subjective (O'Keefe and Jackson, 1995). Unlike persuasiveness, which relates to the actual arguments, effectiveness covers all aspects of an argumentation, including the use of language (van Eemeren, 2015). In particular, the three means of persuasion are meant to be realized by what is said and how (Aristotle, 2007). Several linguistic quality dimensions are connected to argumentation (examples follow in Section 2.2). While many of them are distinguished by Aristotle, he groups them as the *clarity* and the *appropriateness* of style as well as the proper *arrangement*.

Clarity means the use of correct, unambiguous language that avoids unnecessary complexity and deviation from the discussed issue (Aristotle, 2007). Besides ambiguity, vagueness is a major problem impairing clarity (Govier, 2010) and can be a cause of fallacies (Tindale, 2007). So, clarity is a prerequisite of logos. Also, it affects credibility, since it indicates the arguer's skills. An appropriate style in terms of the choice of words supports credibility and emotions. It is tailored to the issue and audience (Aristotle, 2007). Arrangement, finally, addresses the structure of argumentation regarding the presentation of the issue, pros, cons, and conclusions. Damer (2009) outlines that a proper arrangement is governed by the dimensions of a good argument. To be effective, well-arranged argumentation matches the expectations of the target audience and is, thus, related to dialectic (Blair, 2012).

**Dialectic** The dialectical view of argumentation targets the resolution of differences of opinions on the merit (van Eemeren and Grootendorst, 2004). Quality is assessed for well-arranged discussions that seek agreement. In contrast to the subjective nature of effectiveness, people are good in such an assessment (Mercier and Sperber, 2011). In their pragma-dialectical theory, van Eemeren and Grootendorst (2004) develop rules for obtaining *reasonableness* in critical discussions. Reasonableness emerges from two complementary dimensions, intersubjective *(global) acceptability* and problem-solving validity, but effectiveness still remains the underlying goal (van Eemeren, 2015). For argumentation, global acceptability is given when the stated arguments and the way they are stated are acceptable to the whole target audience. Problem-solving validity matches the *(global) relevance* of argumentation that contributes to resolution, helping arrive at an ultimate conclusion (Walton, 2006).

Global relevance implicitly excludes fallacious moves, so reasonable arguments are cogent (van Eemeren, 2015). Van Eemeren sees reasonableness as a precondition for *convincingness*, the rational version of persuasiveness. Following Perelman et al. (1969), persuasive argumentation aims at a particular audience, whereas convincing argumentation aims at the universal audience, i.e., all reasonable beings. This fits the notion that dialectic examines general rather than specific issues (Aristotle, 2007).

Convincingness needs *(global) sufficiency*, i.e., all objections to an argumentation are countered. The dilemma here is that the number of objections could be infinite, but without global sufficiency the required support seems arbitrary (Blair, 2012). A solution is the relaxed view of Damer (2009) that only those counter-arguments that can be anticipated are to be rebutted. For debates, Cohen (2001) speaks of dialectical satisfactoriness, i.e., whether all questions and objections have been sufficiently answered. In case a reasonable debate ends up in either form of global sufficiency, this implies that the discussed difference of opinion is resolved.

**Other** Although closely related, critical thinking (Freeley and Steinberg, 2009) and persuasion research (Zhao et al., 2011) are covered only implicitly here; their views on quality largely match with argumentation theory. We have not discussed deliberation, as it is not concerned with the quality of argumentation primarily but rather with communicative dimensions of group decision-making, e.g., participation and respect (Steenbergen et al., 2003). Also, we have restricted our view to the logic found in natural language. For formal and probabilistic logic, dimensions such as degree of justification (Pollock, 2001), argument strength (Pfeifer, 2013), and premise relevance (Ransom et al., 2015) have been analyzed. As we see below, such logic influenced some practical assessment approaches.

### 2.2 Approaches to Quality Assessment

As for the theories, we survey the automatic quality assessment for natural language argumentation. All discussed approaches are listed in Table 2.

**Logic** Braunstain et al. (2016) deal with logical argument quality in community question answering: Combining relevance-oriented retrieval models and argument-oriented features, they rank sentence-level argument units according to the *level of support* they provide for an answer. Unlike classical essay scoring, Rahimi et al. (2014) score an essay's *evidence*, a quality dimension of argumentation: it captures how sufficiently the given details support the essay's thesis. On the dataset of Correnti et al. (2013) with 1569 student essays and scores from 1 to 4, they find that the concentration and specificity of words related to the essay prompt (i.e., the statement defining the discussed issue) impacts scoring accuracy. Similarly, Stab and Gurevych (2017) introduce an essay corpus with 1029 argument-level annotations of *sufficiency*, following the definition of Johnson and Blair (2006). Their experiments suggest that convolutional neural networks outperform feature-based sufficiency classification.

**Rhetoric** Persing et al. (2010) tackle the proper arrangement of an essay, namely, its *organization* in terms of the logical development of an argument. The authors rely on manual 7-point score annotations for 1003 essays from the ICLE corpus (Granger et al., 2009). In their experiments, sequences of paragraph discourse functions (e.g., introduction or rebuttal) turn out to be most effective. Organization is also analyzed by Rahimi et al. (2015) on the same dataset used for the evidence

| Aspect | Quality Dimension | Granularity | Text Genres | Sources |
|---|---|---|---|---|
| Logic | Evidence | Argumentation | Student essays | Rahimi et al. (2014) |
| | Level of support | Argument unit | Wikipedia articles | Braunstain et al. (2016) |
| | Sufficiency | Argument | Student essays | Stab and Gurevych (2017) |
| Rhetoric | Argument strength | Argumentation | Student essays | Persing and Ng (2015) |
| | Evaluability | Argumentation | Law comments | Park et al. (2015) |
| | Global coherence | Argumentation | Student essays | Feng et al. (2014) |
| | Organization | Argumentation | Student essays | Persing et al. (2010), Rahimi et al. (2015) |
| | Persuasiveness | Argument | Forum discussions | Tan et al. (2016), Wei et al. (2016) |
| | Prompt adherence | Argumentation | Student essays | Persing and Ng (2014) |
| | Thesis clarity | Argumentation | Student essays | Persing and Ng (2013) |
| | Winning side | Debate | Oxford-style debates | Zhang et al. (2016) |
| Dialectic | Acceptability | Argument | Debate portal arguments | Cabrio and Villata (2012) |
| | Convincingness | Argument | Debate portal arguments | Habernal and Gurevych (2016) |
| | Prominence | Argument | Forum discussions | Boltužić and Šnajder (2015) |
| | Relevance | Argument | Diverse genres | Wachsmuth et al. (2017) |

Table 2: Practical assessment of quality dimensions in the referenced sources for the given granularities and text genres of natural language argumentation, grouped by the aspect the quality dimensions refer to.

approach above. Their results indicate a correlation between organization and local coherence. Feng et al. (2014) parse discourse structure to assess *global coherence*, i.e., the continuity of meaning in a text. Lacking ground-truth coherence labels, they evaluate their approach on sentence ordering and organization scoring instead. Coherence affects the clarity of style, as do the *thesis clarity* and *prompt adherence* of essays. Persing and Ng (2013) find the former to suffer from misspellings, while Persing and Ng (2014) use prompt-related keywords and topic models to capture the latter (both for 830 ICLE essays like those mentioned above). For comments in lawmaking, Park et al. (2015) develop an argumentation model that prescribes what information users should give to achieve *evaluability* (e.g., testimony evidence or references to resources).

Not only linguistic quality, but also effectiveness is assessed in recent work: Persing and Ng (2015) score the *argument strength* of essays, which they define rhetorically in terms of how many readers would be persuaded. Although potentially subjective, their manual 7-point score annotations of 1000 ICLE essays differ by at most 1 in 67% of the studied cases. Their best features are heuristic argument unit labels and part-of-speech n-grams. Recently, Wachsmuth et al. (2016) demonstrated that the output of argument mining helps in such argumentation-related essay scoring, obtaining better results for argument strength and organization. Tan et al. (2016) analyze which arguments achieve *persuasiveness* in “change my view” forum discussions, showing that multiple interactions with the view-holder are beneficial as well as an appropriate style and a high number of participants. On similar data, Wei et al. (2016) find that also an author’s reputation impacts persuasiveness. Zhang et al. (2016) discover for Oxford-style debates that attacking the opponents’ arguments tends to be more effective than relying on one’s own arguments. These results indicate the relation of rhetoric and dialectic.

**Dialectic** Dialectical quality has been addressed by Cabrio and Villata (2012). The authors use textual entailment to find ground-truth debate portal arguments that attack others. Based on the formal argumentation framework of Dung (1995), they then assess global argument *acceptability*. Habernal and Gurevych (2016) compare arguments in terms of *convincingness*. However, the subjective nature of their crowdsourced labels actually reflects rhetorical effectiveness. Boltužić and Šnajder (2015) present first steps towards argument *prominence*. Prominence may be a product of popularity, though, making its quality nature questionable, as popularity is often not correlated with merit (Govier, 2010). In contrast, Wachsmuth et al. (2017) adapt the famous PageRank algorithm to objectively derive the *relevance* of an argument at web scale from what other arguments refer to the argument’s premises. On a large ground-truth argument graph, their approach beats several baselines for the benchmark argument rankings that they provide.

**Other** Again, we have left out deliberative quality (Gold et al., 2015). Also, we omit approaches that classify argumentation schemes (Feng and Hirst, 2011), evidence types (Rinott et al., 2015), ethos-related statements (Duthie et al., 2016), and myside bias (Stab and Gurevych, 2016); their output may help assess quality assessment, but they do not actually assess it. The same holds for argument mining,

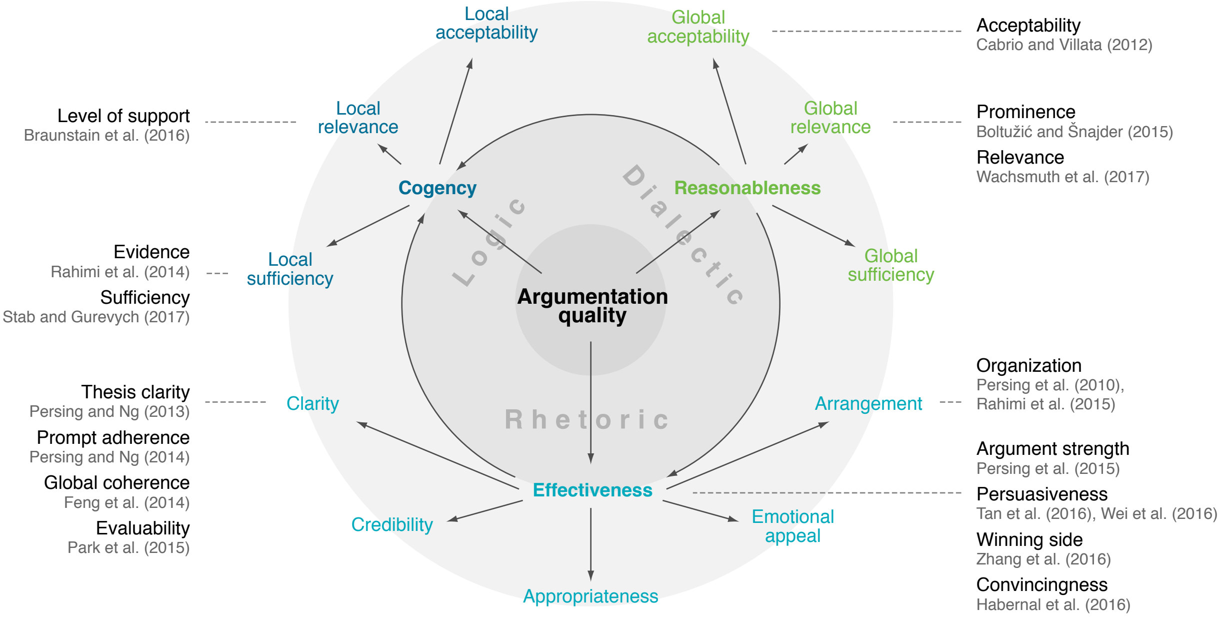

Figure 1: The proposed taxonomy of argumentation quality as well as the mapping of existing assessment approaches to the covered quality dimensions. Arrows show main dependencies between the dimensions.

even if said to aim for argument quality (Swanson et al., 2015). Much work exists for general text quality, most notably in the context of readability (Pitler and Nenkova, 2008) and classical essay scoring. Some scoring approaches derive features from discourse (Burstein et al., 1998), arguments (Ong et al., 2014; Beigman Klebanov et al., 2016; Ghosh et al., 2016), or schemes (Song et al., 2014)—all this may be indicative of quality. However, our focus is approaches that target argumentation quality at heart. Similarly, review helpfulness (Liu et al., 2008) and deception (Ott et al., 2011) are not treated, as arguments only partly play a role there. Also, only few Wikipedia quality flaws relate to arguments, e.g., verifiability (Anderka et al., 2012).

## 3 A Taxonomy of Argumentation Quality

Given all surveyed quality dimensions, we now propose a unifying taxonomy of argumentation quality. The taxonomy decomposes quality assessment systematically, thus organizing and clarifying the roles of practical approaches. It does not require a particular argumentation model, but it rests on the notion of the granularity levels from Section 1.

### 3.1 Overview of the Theory-based Taxonomy

Our objective is not to come up with a new theory, but to provide a unified view of existing theories that is suitable for quality assessment. We aim for a common understanding of the dimensions that affect quality, what interdependencies they have, and how they interact. Figure 1 illustrates the taxonomy that we propose for this purpose. The rationale behind its structure and its layout is as follows.

While Section 2 has outlined overlaps and relations between the three aspects of argumentation, we have identified one dominant high-level quality dimension of *argumentation quality* in theory for each aspect: logical *cogency*, rhetorical *effectiveness*, and dialectical *reasonableness*. The latter two benefit from cogency, and reasonableness depends on effectiveness, as discussed. Often, only one of them will be in the focus of attention in practice, or even only a sub-dimension. In particular, each high-level dimension has a set of sub-dimensions agreed upon. The sub-dimensions are shown on the outer ring in Figure 1, roughly positioned according to the aspects they refer to, e.g., *local acceptability* lies next to the other dialectical dimensions. We ordered the sub-dimensions by their interrelations (left implicit for conciseness), e.g., *appropriateness* supports *credibility* and *emotional appeal*.

Slightly deviating from theory, we match Aristotle's logos dimension with cogency, which better fits real-world argumentation. Similarly, we omit those dimensions from Table 1 in the taxonomy that have unclear definitions, such as strength, or that are covered by others, such as well-formedness, which merely refines the acceptability part of cogency (Govier, 2010). Convincingness is left out, as it is close to effectiveness and as both the feasibility and the need of persuading the universal audience has been questioned (van Eemeren, 2015). Instead, we add *global sufficiency* as part of reasonableness. While global sufficiency may be infeasible, too (Blair, 2012), it forces agreement in critical discussions and, thereby, reasonableness.

### 3.2 Definitions of the Quality Dimensions

Cogency is seen as an argument property, whereas effectiveness and reasonableness are assessed on the argumentation level usually. For generality, we give informal literature-based definitions of these dimensions and all sub-dimensions here for an author who argues about an issue to a target audience:

**Cogency** An argument is cogent if it has acceptable premises that are relevant to its conclusion and that are sufficient to draw the conclusion.

- *Local acceptability:* A premise of an argument is acceptable if it is rationally worthy of being believed to be true.
- *Local relevance:* A premise of an argument is relevant if it contributes to the acceptance or rejection of the argument's conclusion.
- *Local sufficiency:* An argument's premises are sufficient if, together, they give enough support to make it rational to draw its conclusion.

**Effectiveness** Argumentation is effective if it persuades the target audience of (or corroborates agreement with) the author's stance on the issue.

- *Credibility:* Argumentation creates credibility if it conveys arguments and similar in a way that makes the author worthy of credence.
- *Emotional Appeal:* Argumentation makes a successful emotional appeal if it creates emotions in a way that makes the target audience more open to the author's arguments.
- *Clarity:* Argumentation has a clear style if it uses correct and widely unambiguous language as well as if it avoids unnecessary complexity and deviation from the issue.
- *Appropriateness:* Argumentation has an appropriate style if the used language supports the creation of credibility and emotions as well as if it is proportional to the issue.
- *Arrangement:* Argumentation is arranged properly if it presents the issue, the arguments, and its conclusion in the right order.

**Reasonableness** Argumentation is reasonable if it contributes to the issue's resolution in a sufficient way that is acceptable to the target audience.

- *Global acceptability:* Argumentation is acceptable if the target audience accepts both the consideration of the stated arguments for the issue and the way they are stated.
- *Global relevance:* Argumentation is relevant if it contributes to the issue's resolution, i.e., if it states arguments or other information that help to arrive at an ultimate conclusion.
- *Global sufficiency:* Argumentation is sufficient if it adequately rebuts those counter-arguments to it that can be anticipated.

### 3.3 Organization of Assessment Approaches

The taxonomy is meant to define a common ground for assessing argumentation quality, including the organization of practical approaches. The left and right side of Figure 1 show where the approaches surveyed in Section 2.2 are positioned in the taxonomy. Some dimensions have been tackled multiple times (e.g., *clarity*), others not at all (e.g., *credibility*). The taxonomy indicates what sub-dimensions will affect the same high-level dimension.

## 4 The Dagstuhl-15512 ArgQuality Corpus

Finally, we present our new annotated *Dagstuhl-15512 ArgQuality Corpus* for studying argumentation quality based on the developed taxonomy, and we report on a first corpus analysis.[3]

### 4.1 Data and Annotation Process

Our corpus is based on the *UKPConvArgRank* dataset (Habernal and Gurevych, 2016), which contains rankings of 25 to 35 textual debate portal arguments for two stances on 16 issues, such as *evolution vs. creation* and *ban plastic water bottles*. All ranks were derived from crowdsourced convincingness labels. For every issue/stance pair, we took the five top-ranked texts and chose five further via stratified sampling. Thereby, we covered both high-quality arguments and different levels of lower quality. Two example texts follow below in Figure 2.

Before annotating the 320 chosen texts, we carried out a full annotation study with seven authors of this paper on 20 argumentative comments from

[3] The corpus and annotation guidelines are available at http://www.arguana.com. The corpus is named after the Dagstuhl Seminar 15512 “Debating Technologies” that initialized the research in this paper: http://www.dagstuhl.de/15512

| | | (a) Maj. Scores | | | (b) Agreement | | | (c) Pearson Correlation Coefficients | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Quality Dimension** | | **1** | **2** | **3** | $\alpha$ | **full** | **maj.** | **Co** | **LA** | **LR** | **LS** | **Ef** | **Cr** | **Em** | **Cl** | **Ap** | **Ar** | **Re** | **GA** | **GR** | **GS** |
| **Co** | **Cogency** | 150 | 131 | 23 | .44 | 40.1% | 91.8% | | .64 | .61 | **.84** | **.81** | .46 | .27 | .41 | .32 | .55 | .78 | .64 | **.71** | .70 |
| LA | Local acceptability | 84 | 169 | 51 | .46 | 27.0% | 90.8% | .64 | | .51 | .53 | .60 | **.54** | .30 | .40 | .54 | .46 | .68 | .75 | .46 | .45 |
| LR | Local relevance | 25 | 155 | **124** | .47 | 32.6% | 92.4% | .61 | .51 | | .56 | .56 | .39 | .27 | .46 | .35 | .50 | .62 | .58 | .68 | .45 |
| LS | Local sufficiency | 172 | 119 | 13 | .44 | 37.2% | 92.8% | **.84** | .53 | .56 | | .73 | .39 | .25 | .37 | .23 | .51 | .67 | .51 | .68 | **.74** |
| **Ef** | **Effectiveness** | 184 | 111 | 9 | .45 | 42.1% | 94.4% | .81 | .60 | .56 | .73 | | .48 | .31 | .35 | .34 | .54 | .75 | .58 | .66 | .71 |
| Cr | Credibility | 99 | 199 | 6 | .37 | 37.8% | 95.7% | .46 | .54 | .39 | .39 | .48 | | **.37** | .32 | .49 | .37 | .52 | .52 | .36 | .40 |
| Em | Emotional appeal | 48 | **235** | 21 | .26 | 42.8% | 94.4% | .27 | .30 | .27 | .25 | .31 | .37 | | .14 | .30 | .20 | .30 | .26 | .26 | .22 |
| Cl | Clarity | 42 | 191 | 71 | .35 | 29.3% | 89.8% | .41 | .40 | .46 | .37 | .35 | .32 | .14 | | .45 | .56 | .44 | .45 | .38 | .27 |
| Ap | Appropriateness | 43 | 196 | 65 | .36 | 17.4% | 87.5% | .32 | .54 | .35 | .23 | .34 | .49 | .30 | .45 | | .48 | .47 | .59 | .20 | .20 |
| Ar | Arrangement | 91 | 189 | 24 | .39 | 26.6% | 93.4% | .55 | .46 | .50 | .51 | .54 | .37 | .20 | **.56** | .48 | | .55 | .51 | .49 | .48 |
| **Re** | **Reasonableness** | 126 | 159 | 19 | .50 | 41.4% | 95.7% | .78 | .68 | .62 | .67 | .75 | .52 | .30 | .44 | .47 | .55 | | **.78** | .65 | .61 |
| GA | Global acceptability | 88 | 161 | 55 | .44 | 31.6% | 95.4% | .64 | **.75** | .58 | .51 | .58 | .52 | .26 | .45 | **.59** | .51 | .78 | | .46 | .43 |
| GR | Global relevance | 69 | 167 | 68 | .42 | 21.7% | 90.1% | .71 | .46 | **.68** | .68 | .66 | .36 | .26 | .38 | .20 | .49 | .65 | .46 | | .61 |
| GS | Global sufficiency | **231** | 72 | 1 | .27 | **44.7%** | **98.0%** | .70 | .45 | .45 | .74 | .71 | .40 | .22 | .27 | .20 | .48 | .61 | .43 | .61 | |
| **Ov** | **Overall quality** | 152 | 128 | 24 | **.51** | 44.1% | 94.4% | **.84** | .66 | .61 | .74 | **.81** | .52 | .30 | .45 | .42 | **.59** | **.86** | .71 | .70 | .68 |

Table 3: Results for the 304 corpus texts classified as argumentative by all annotators: (a) Distribution of majority scores for each dimension (2 used in case of full disagreement). (b) Krippendorff's $\alpha$ of the most agreeing annotator pair and full/majority agreement of all annotators. (c) Correlation for each dimension pair, averaged over the correlations of all annotators. The highest value in each column is marked bold.

the unshared task dataset of the 3rd Workshop on Argument Mining.[4] The annotators assessed all 15 quality dimensions in the taxonomy for each comment (including its overall quality). Due to simple initial guidelines based on the definitions from Section 3 and the subjectiveness of the task, the agreement of all seven annotators was low for all dimensions, namely, at most .22 in terms of Krippendorff's $\alpha$. The three most agreeing annotators for each dimension achieved much higher $\alpha$-values between .23 (clarity) and .60 (credibility), though.[5]

The study results were discussed by all annotators, leading to a considerably refined version of the guidelines. We then selected three annotators for the corpus annotation based on their availability. They work at two universities and one company in three countries (two females, one male; two PhDs, one PhD student). For each text in the corpus, all annotators first classified whether it was actually argumentative. If so, they assessed all dimensions using ordinal scores from 1 (low) to 3 (high).[6] Additionally, "cannot judge" could be chosen.

## 4.2 Corpus Distribution and Agreement

Table 3(a) lists the majority scores of each dimension for the 304 corpus texts (95%) that are classified as argumentative by all annotators, all covering the whole score range. Five dimensions have the median at score 1, the others at 2. Some seem easier to master, such as *local relevance*, which received the highest majority score 124 times. Others rarely got score 3, above all *global sufficiency*. The latter is explained by the fact that only few texts include any rebuttal of counter-arguments.

Only one of the over 14,000 assessments made by the three annotators was "cannot judge" (for *global relevance*), suggesting that our guidelines were comprehensive. Regarding agreement, we see in Table 3(b) that the $\alpha$-values of all logical and dialectical quality dimensions except for *global sufficiency* lie above 0.4 for the most agreeing annotator pair. As expected, the rhetorical dimensions seem to be more subjective. The lowest $\alpha$ is observed for *emotional appeal* (0.26). The annotators most agreed on the *overall quality* ($\alpha = 0.51$), possibly meaning that the taxonomy adequately guides the assessment. In accordance with the moderate $\alpha$-values, full agreement ranges between 17.4% and 44.7% only. On the contrary, we observe high majority agreement between 87.5% and 98% for all dimensions, even where scores are rather evenly distributed, such as for *global acceptability* (95.4%). In case of full disagreement, it makes sense to use score 2. We hence argue that the corpus is suitable for evaluating argumentation quality assessment.

Figure 2 shows all scores of each annotator for two example arguments from the corpus, referring to the question whether to ban plastic water bottles. Both have majority score 3 for *overall quality (Ov)*,

[4] Unshared task data found at: http://github.com/UKPLab

[5] We use Krippendorff's $\alpha$ as is suitable for small samples, multiple ratings, and ordinal scales (Krippendorff, 2007).

[6] We chose a 3-point scale to foster clear decisions on the quality; in the annotation study, we used a 4-point scale but observed that the annotators only rarely chose score 1 and 4.

| Arguments | **Pro** Water bottles, good or bad? Many people believe plastic water bottles to be good. But the truth is water bottles are polluting land and unnecessary. Plastic water bottles should only be used in emergency purposes only. The water in those plastic are only filtered tap water. In an emergency situation like Katrina no one had access to tap water. In a situation like this water bottles are good because it provides the people in need. Other than that water bottles should not be legal because it pollutes the land and big companies get 1000% of the profit. | **Con** Americans spend billions on bottled water every year. Banning their sale would greatly hurt an already struggling economy. In addition to the actual sale of water bottles, the plastics that they are made out of, and the advertising on both the bottles and packaging are also big business. In addition to this, compostable waters bottle are also coming onto the market, these can be used instead of plastics to eliminate that detriment. Moreover, bottled water not only has a cleaner safety record than municipal water, but it easier to trace when a potential health risk does occur. (http://www.friendsjournal.org/bottled-water) (http://www.cdc.gov/healthywater/drinking/bottled/) |
|---|---|---|

| Scores | Co | LA | LR | LS | Ef | Cr | Em | Cl | Ap | Ar | Re | GA | GR | GS | Ov | Co | LA | LR | LS | Ef | Cr | Em | Cl | Ap | Ar | Re | GA | GR | GS | Ov |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Annotator A | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Annotator B | 2 | 2 | 3 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 3 | 3 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 2 | 3 |
| Annotator C | 2 | 3 | 3 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Majority score | 2 | 3 | 3 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |

Figure 2: The scores of each annotator and the majority score for all considered quality dimensions of one pro and one con argument from our corpus. The arguments refer to the issue *ban plastic water bottles*.

but the pro argument shows more controversy with full disagreement in case of *effectiveness (Ef)*. Especially, *annotator B* seems to be critical, giving one point less for several dimensions. In contrast, the con argument yields majority agreement for all 15 dimensions and full agreement for seven of them. It meets main quality criteria surveyed in Section 2, such as a rebuttal or references to resources. In fact, it constitutes the only corpus text with majority score 3 for *global sufficiency (GS)*.

### 4.3 Correlations between Quality Dimensions

Table 3(c) compares the correlations of all dimension pairs. *Cogency* (.84), *effectiveness* (.81), and *reasonableness* (.86) correlate strongly with *overall quality*, and also much with each other.

Cogency and *local sufficiency* (.84) go hand in hand, whereas *local acceptability* and *local relevance* show the highest correlation with their global counterparts (.75 and .68 respectively). Quite intuitively, *credibility* and *appropriateness* correlate most with the acceptability dimensions. The coefficients of *emotional appeal* seem lower than expected, in particular for effectiveness (.31), indicating the limitation of a correlation analysis: As reflected by the 235 texts with majority score 2 for emotional appeal, many arguments make no use of emotions, thus obliterating effects of those which do. On the other hand, *clarity* was scored 2 in most cases, too, so the very low value there (.14) is more meaningful. Clarity rather correlates with *arrangement* (.56), which in turn shows coefficients above .50 for all high-level dimensions.

Altogether, the correlations largely match the surveyed theory. While an analysis of cause and effect should follow in future work, they provide first evidence for the adequacy of our taxonomy.

## 5 Conclusion

Argumentation quality is of high importance for argument mining, debating technologies, and similar. In computational linguistics, it has been treated only rudimentarily so far. This paper defines a common ground for the automatic assessment of argumentation quality in natural language. Based on a survey of existing theories and approaches, we have developed a taxonomy that unifies all major dimensions of logical, and dialectical argumentation quality. In addition, we freely provide an annotated corpus for studying these dimensions.

The taxonomy is meant to capture *all* aspects of argumentation quality, irrespective of how they can be operationalized. The varying inter-annotator agreement we obtained suggests that some quality dimensions are particularly subjective, raising the need to model the target audience of an argumentation. Still, the observed correlations between the dimensions support the general adequacy of our taxonomy. Moreover, most dimensions have already been approached on a certain abstraction level in previous work, as outlined. While some refinement may be suitable to meet all requirements of the community, we thus propose the taxonomy as the common ground for future research on computational argumentation quality assessment and the corpus as a first benchmark dataset for this purpose.

## Acknowledgments

We thank all attendees of Dagstuhl Seminar 15512, particularly the rest of the quality breakout group: Wolf-Tilo Balke, Ruty Rinott, and Christian Stab. Also, we acknowledge financial support of the Stanford University, the DFG, and the Natural Sciences and Engineering Research Council of Canada.

## References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404. Association for Computational Linguistics.

Maik Anderka, Benno Stein, and Nedim Lipka. 2012. Predicting quality flaws in user-generated content: The case of Wikipedia. In *Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval*, pages 981–990.

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, Translator). Clarendon Aristotle series. Oxford University Press.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75. Association for Computational Linguistics.

Yonatan Bilu and Noam Slonim. 2016. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 525–530. Association for Computational Linguistics.

J. Anthony Blair. 2012. *Groundwork in the Theory of Argumentation*. Springer Netherlands.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115. Association for Computational Linguistics.

Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. 2016. Supporting human answers for advice-seeking questions in CQA sites. In *Proceedings of the 38th European Conference on IR Research*, pages 129–141.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. 1998. Enriching automated essay scoring using discourse marking. In *Discourse Relations and Discourse Markers*.

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

Daniel H. Cohen. 2001. Evaluating arguments and making meta-arguments. *Informal Logic*, 21(2):73–84.

Richard Correnti, Lindsay Clare Matsumura, Laura Hamilton, and Elaine Wang. 2013. Assessing students' skills at writing analytically in response to texts. *The Elementary School Journal*, 114(2):142–177.

T. Edward Damer. 2009. *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments*. Wadsworth, Cengage Learning, Belmont, CA, 6th edition.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Rory Duthie, Katarzyna Budynska, and Chris Reed. 2016. Mining ethos in political debate. In *Proceedings of the Sixth International Conference on Computational Models of Argument*, pages 299–310.

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

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949. Dublin City University and Association for Computational Linguistics.

Austin J. Freeley and David L. Steinberg. 2009. *Argumentation and Debate*. Cengage Learning, Boston, MA, 12th edition.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*. Springer.

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

Valentin Gold, Mennatallah El-Assady, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2015. Visual linguistic analysis of political discussions: Measuring deliberative quality. *Digital Scholarship in the Humanities*.

Trudy Govier. 2010. *A Practical Study of Argument*. Wadsworth, Cengage Learning, Belmont, CA, 7th edition.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. International corpus of learner English (version 2).

Ivan Habernal and Iryna Gurevych. 2016. 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. Association for Computational Linguistics.

Charles L. Hamblin. 1970. *Fallacies*. Methuen, London, UK.

Hans Hoeken. 2001. Anecdotal, statistical, and causal evidence: Their perceived and actual persuasiveness. *Argumentation*, 15(4):425–437.

Ralph H. Johnson and J. Anthony Blair. 2006. *Logical Self-defense*. International Debate Education Association.

Klaus Krippendorff. 2007. Computing Krippendorff's alpha reliability. Technical report, Univ. of Pennsylvania, Annenberg School for Communication.

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2008. Modeling and predicting the helpfulness of online reviews. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 443–452.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34:57–111.

Nona Naderi and Graeme Hirst. 2015. Argumentation mining in parliamentary discourse. In *Principles and Practice of Multi-Agent Systems - International Workshops: IWEC 2014, Gold Coast, QLD, Australia, December 1-5, 2014, and CMNA XV and IWEC 2015, Bertinoro, Italy, October 26, 2015, Revised Selected Papers*, pages 16–25.

Daniel J. O'Keefe and Sally Jackson. 1995. Argument quality and persuasive effects: A review of current approaches. In *Argumentation and Values: Proceedings of the Ninth Alta Conference on Argumentation*, pages 88–92.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and T. Jeffrey Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319. Association for Computational Linguistics.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in eRulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 206–210.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948. Association for Computational Linguistics.

Chaïm Perelman, Lucie Olbrechts-Tyteca, John Wilkinson, and Purcell Weaver. 1969. *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press, Notre Dame, IN.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Niki Pfeifer, 2013. *Bayesian Argumentation: The Practical Side of Probability*, chapter On Argument Strength, pages 185–193. Springer Netherlands, Dordrecht.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics.

John L. Pollock. 2001. Defeasible reasoning with variable degrees of justification. *Artificial Intelligence*, 133(1–2):233–282.

Zahra Rahimi, Diane J. Litman, Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, and Zahid Kisa. 2014. Automatic scoring of an analytical response-to-text assessment. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, pages 601–610.

Zahra Rahimi, Diane Litman, Elaine Wang, and Richard Correnti. 2015. Incorporating coherence of topics as a criterion in automatic response-to-text assessment of the organization of writing. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 20–30. Association for Computational Linguistics.

Keith J. Ransom, Amy Perfors, and Daniel J. Navarro. 2015. Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, pages 1–22.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence — an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450. Association for Computational Linguistics.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2016. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 113–118. Association for Computational Linguistics.

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*.

Marco R. Steenbergen, Andre Bachtiger, Markus Sporndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

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. Association for Computational Linguistics.

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 World Wide Web Conference*, pages 613–624.

Christopher W. Tindale. 2007. *Fallacies and Argument Appraisal. Critical Reasoning and Argumentation*. Cambridge University Press.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Frans H. van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press, Cambridge, UK.

Frans H. van Eemeren. 2015. *Reasonableness and Effectiveness in Argumentative Discourse: Fifty Contributions to the Development of Pragma-Dialectics*. Argumentation Library. Springer International Publishing.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. 2015. Sentiment flow — A general model of web review argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 601–611. Association for Computational Linguistics.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017. “PageRank” for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Douglas Walton. 2006. *Fundamentals of Critical Argumentation*. Cambridge University Press.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? Ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Xiaoquan Zhao, Andrew Strasser, Joseph N. Cappella, Caryn Lerman, and Martin Fishbein. 2011. A measure of perceived argument strength: Reliability and validity. *Communication Methods and Measures*, 5(1):48–75.