Creating a Domain-diverse Corpus for Theory-based Argument Quality Assessment

Lily Ng[1]*, Anne Lauscher[2]*, Joel Tetreault[3], Courtney Napoles[1]
[1]Grammarly
[2]Data and Web Science Group, University of Mannheim, Germany
[3]Dataminr, Inc.
[1]first.last@grammarly.com, [2]anne@informatik.uni-mannheim.de,
[3]jtetreault@dataminr.com

Abstract

Computational models of argument quality (AQ) have focused primarily on assessing the overall quality or just one specific characteristic of an argument, such as its *convincingness* or its *clarity*. However, previous work has claimed that assessment based on theoretical dimensions of argumentation could benefit writers, but developing such models has been limited by the lack of annotated data. In this work, we describe GAQCorpus, the first large, domain-diverse annotated corpus of theory-based AQ. We discuss how we designed the annotation task to reliably collect a large number of judgments with crowdsourcing, formulating theory-based guidelines that helped make subjective judgments of AQ more objective. We demonstrate how to identify arguments and adapt the annotation task for three diverse domains. Our work will inform research on theory-based argumentation annotation and enable the creation of more diverse corpora to support computational AQ assessment.

1 Introduction

The notion of *Argumentation Quality (AQ)* plays an important role in many existing argument-related downstream applications, such as argumentative writing support (Stab and Gurevych, 2017), automatic essay grading (Persing and Ng, 2013), and debate systems (Toledo et al., 2019). For some of these applications, the idea is to automatically give feedback to users to help them improve their writing skills or assess their writing capabilities. For others, assessing AQ is an important step in a more complex pipeline for retrieving high-quality arguments.

While grading overall AQ (Toledo et al., 2019) or a specific conceptualization of AQ, such as *prompt adherence* (Persing and Ng, 2014) is relatively well explored, researchers have noted the lack of work in so-called *theory-based AQ*[1] (Wachsmuth et al., 2017b), which can be represented with a taxonomy characterizing overall AQ into several subdimensions and aspects, for instance, as *logic* and *rhetoric*, which therefore provides a more informative and targeted perspective. However, this holistic approach comes with the downside of higher complexity, especially when it comes to annotating textual corpora, which are required for training and developing common computational approaches (see, e.g., Gretz et al. (2020)). In a small study, Wachsmuth et al. (2017a) demonstrate that theory-based AQ annotations can be done both by trained experts and by crowd annotators, though the authors acknowledge the high complexity and subjectivity of the problem and call for the simplification of theory-based AQ annotation in order to reliably create larger-scale corpora. To date, no work has tackled this challenge and accordingly, no larger-scale and no domain-diverse corpus of this kind exists. We aim to close this gap by describing our efforts to create Grammarly Argument Quality Corpus (GAQCorpus) (Lauscher et al., 2020), the largest and the only domain-diverse corpus consisting of 5,285 English arguments annotated with theory-based AQ scores across four dimensions.

*Equal contribution.

[1]In the following, we adopt the term "theory-based AQ," which was proposed by Wachsmuth et al. (2017b) to indicate that the conception of AQ is specifically grounded in argumentation theoretic literature (and not in CL or NLP).

Proceedings of the 7th Workshop on Argument Mining, pages 117–126
Barcelona, Spain (Online), December 13, 2020.

Building on Wachsmuth et al. (2017a), in this work, we modify the annotation task to be suitable for both experts and the crowd while preserving the theoretical basis of the taxonomy. We collect and annotate argumentative texts from web debate forums, as well as community questions and answers forums (CQA), and review forum texts, which are still understudied in computational AQ. The latter domains can consist of rather non-canonical arguments in that they exhibit a lack of explicitness of certain argumentative components; are topic-wise more subjective; or consist of longer, more convoluted text. This makes assessing the quality of such arguments even more challenging, but downstream can result in a more robust model of computational AQ.

Given all these challenges, we work closely with trained linguists to adapt the annotation task, iterating over how best to approach these novel domains and simplify the annotation guidelines for crowdsourcing, allowing us to collect a large number of judgments efficiently. We hope that our work fuels further research on theory-based computational AQ. Our approach to building `GAQCorpus` can inspire and inform AQ annotation in new domains, enriching the domain-diversity of linguistic resources available in this space and consequently expanding computational approaches to AQ.

Structure. We start by surveying previous AQ annotation studies (§2). Next, we describe our efforts to adapt and simplify the annotation task (§3), which is followed by a discussion of the data domains (§4). §5 presents an analysis of the resulting corpus. Finally, we conclude our work and provide directions for future research (§6).

2 Related Work

Most argumentation annotation studies have been conducted on student essays or web debates. Student essays have been annotated for thesis clarity (Persing and Ng, 2013), organization (Persing et al., 2010), and prompt adherence (Persing and Ng, 2014), and Persing and Ng (2015) model argument strength rated on a 4-point Likert scale. Similarly, Stab and Gurevych (2016) annotate the absence of opposing arguments and Stab and Gurevych (2017) predict insufficient premise support in arguments. For web debates, Habernal and Gurevych (2016) conduct an annotation study in which they present debate arguments pairwise to crowd annotators, who then can choose the more convincing argument. Persing and Ng (2017) also annotate the reasons why an argument receives a low persuasive power score.

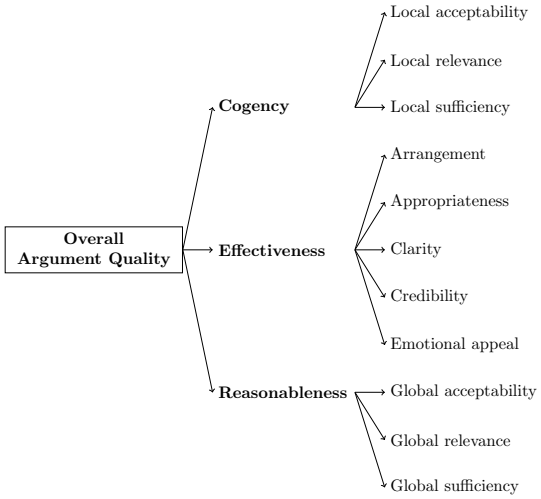


Figure 1: The taxonomy of theory-based argument quality aspects (Wachsmuth et al., 2017b).

Wachsmuth et al. (2017b) developed a taxonomy of AQ synthesized from traditional works in argumentation theory, such as Aristotle (trans 2007). The full taxonomy is depicted in Figure 1, and defines the Overall AQ to consist of the following three subdimensions, each of which is itself defined by several finer-grained AQ aspects:

(1) Cogency relates to the logical aspects of AQ, for instance, whether the an argument's premises are acceptable (local acceptability) or whether they can be seen as relevant for the conclusion (local relevance).

(2) Effectiveness indicates the rhetorical aspects of an argument. Aspects of effectiveness include, for instance, its clarity or its emotional appeal.

(3) Reasonableness reflects the quality of an argument in the overall context of the discussion, as, for instance, its relevance towards arriving at a resolution of the issue (global relevance).

Wachsmuth et al. (2017a) conducted a study in which crowd workers annotated 304 arguments for all 15 quality dimensions (Figure 1), and demonstrated that the theory-based and practical AQ assessments

match to a large extent. Their findings indicate that theory-based annotations can be crowdsourced and that theory-based approaches can inform the practical view, especially. Most importantly, the authors conclude that the annotation task should be simplified to guarantee a reliable crowd-annotation process.

Most recently, Toledo et al. (2019) and Gretz et al. (2020) crowdsourced overall argument quality by presenting pairwise arguments to annotators, who then had to select the argument "they would recommend a friend to use that argument as is in a speech supporting/contesting the topic." This is an extreme simplification of the task, which does not seem to lead to better agreement: the authors (Gretz et al., 2020) report an average inter-annotator agreement of $\kappa = 0.12$ and attribute the low score to the high subjectivity of the task. The authors conducted a theory-based annotation study in the spirit of Wachsmuth et al. (2017b) on a subset of the data (100 arguments) which indicated the highest correlation of the annotations with the effectiveness dimension. Later on, Lauscher et al. (2020) empirically confirmed this observation using computational model predictions across the whole corpus.

Building on this large body of work, we aim to facilitate the annotation of theory-based AQ in diverse domains of real-world argumentative writing and compare expert vs. crowd annotations. Our study results in the largest English corpus annotated with theory-based argumentative quality scores.

3 Annotation Study

In this section, we detail how we developed and designed our annotation task to enable efficient, reliable collection of theory-based AQ judgments with crowdsourcing. We validate Wachsmuth et al. (2017a)'s hypothesis that crowdsourced annotation of theory-based AQ is possible if the task is simplified.

3.1 Simplifying the task

Before collecting any crowdsourced annotations, we conducted 14 pilot experiments with a group of four "expert" annotators, simplifying the TVSP task design through their feedback and observations, as they provided both a deep understanding of the argumentation theory and practical experience annotating the arguments. Each expert annotator was a fluent or native English speaker with an advanced degree in linguistics. Experts underwent training, which included studying guidelines and participating in calibration tasks to analyze debate arguments from three sources: Dagstuhl-ArgQuality-Corpus-V2[2], originally from UKPConvArgRank (Habernal and Gurevych, 2016); the Internet Argument corpus V2[3] (IAC) (Abbott et al., 2016); and ChangeMyView,[4] a Reddit forum. Through the pilots and subsequent debriefs with the experts, we made the following modifications to the annotation task of Wachsmuth et al. (2017a):

(1) Reduce taxonomy complexity. While TVSP defined the task to score all 11 AQ subaspects (Local Acceptability, Local Relevance, etc.), 3 dimensions (Cogency, Effectiveness, Reasonableness), and overall AQ, we reduced the number of qualities scored by only focusing on the 3 higher-level dimensions plus overall AQ. As a result, annotators assessed an argumentative text in terms of 4 scores instead of 15 scores, and instead of 3 different AQ levels, the simplified taxonomy is reduced to 2.

(2) Instruction Modifications. We reworded the TVSP dimension descriptions and added several examples to make the guidelines more understandable. As the annotators were not rating the 11 AQ subaspects, we experimented with different methods to incorporate the subaspects into the guidelines. Instead of explaining the subdimensions in the guidelines and trusting crowd annotators to bear them in mind, we represented each subdimension as a yes/no question in the annotation task itself (Table 1). Our pilot experiments showed that presenting the questions without asking for a response eased the perceived complexity of the task while not affecting agreement.

(3) Five-point scale. While TVSP collected judgments with a three-point rating scale (low, medium, high), we employ a five-point scale (very low, low, medium, high, very high, plus *cannot judge*) to allow for more nuanced judgments, as the expert annotators found too great of a distance between the items on a three-point scale. Scales with 5–9 items have been shown to be optimal, balancing the informational

[2] http://argumentation.bplaced.net/arguana/data

[3] https://nlds.soe.ucsc.edu/iac2

[4] https://www.reddit.com/r/changemyview/

Dimension	Subdimension	Question
Cogency	Local Acceptability	Are the justifications for the argument acceptable/believable?
	Local Relevance	Are the justifications relevant to the author's point?
	Local Sufficiency	Do the justifications provide enough support to draw a conclusion?
Effectiveness	Credibility	Is the author qualified to be making the argument?
	Emotional Appeal	Does the argument evoke emotions that make the audience more likely to agree with the author?
	Clarity	Does the author's language make it easy for you to understand what they are arguing for or against?
	Appropriateness	Is the author's argument and delivery appropriate for an online forum?
	Arrangement	Did the author present their argument in an order that makes sense?
Reasonableness	Global Acceptability	Would the target audience accept the argument and the way it is stated?
	Global Relevance	Does the argument contribute to the resolution of the given issue?
	Global Sufficiency	Does the argument address and adequately rebut counterarguments?

Table 1: Subdimensions represented as questions in the annotation task of debates.

needs of the researcher and the capacity of the raters (Cox III, 1980). We experimented with both three- and five-point scales and found that the larger scale did not negatively affect inter-annotator agreement.

3.2 Validating the Task Design

Our finalized task design is as follows. First, annotators decide whether a text is argumentative. If *yes*, the three high-level dimensions are scored on a five-point scale and subaspect questions are presented to guide the annotator's judgment. The Overall AQ is scored last, also on a five-point scale.

Before collecting annotations from the crowd, we validated our modifications subjectively and objectively. First, we ran a series of pilot tasks with our expert annotators. They initially annotated using the TvsP guidelines and next worked with the simplified taxonomy. In follow-up discussions, the experts confirmed that the new task design reduced the time and cognitive load necessary to rate arguments, and that the guidelines were more understandable. These modifications make the task more approachable, which is vital when presenting it to (untrained) crowd-workers for larger-scale annotation.

	Cogency	Effectiveness	Reasonableness	Overall
Ours	**0.46**	**0.48**	**0.48**	**0.55**
TvsP	0.27	0.38	0.13	0.43

Table 2: Agreement Dagstuhl "gold" annotations and our crowdsourced annotations (Ours) compared to TvsP.

We validated the simplifications quantitatively by reproducing the study of TvsP, which compared their crowd and "expert" annotations. To this end, we randomly sampled 200 arguments from Dagstuhl-ArgQuality-Corpus-V2, which come with author-annotated "gold" ratings. We collected ratings from a crowd (10 ratings per item), following our simplified design[5] (§3.1). All crowd contributors were native or fluent English speakers engaged through Appen (formerly Figure Eight). Crowd contributors did not participate in calibration meetings and all feedback was relayed to contributors through a liaison.

We average the crowd ratings to obtain a single score for each argument and computed the inter-annotator agreement (IAA) with the "gold" annotations using Krippendorff's α (Krippendorff, 2007) (Table 2). Even though the annotation scores are not strong, the IAA between our crowd annotators and the gold annotations generally surpasses the agreement scores reported by TvsP. This is a highly nuanced and subjective task, which is reflected in the agreement levels. Based on these results and annotator observations, we conclude that our task guidelines and design allow for better (or at least comparable) quality crowdsourcing of theory-based AQ annotations.

[5]The only difference is that we used a 3-point scale to more fairly compare to the gold.

4 Data Domains

In this work, we consider three domains: *Debate* forums, *CQA* forums, and *Review* forums. While Debates are generally well-explored in computational AQ, we are unaware of any work involving CQA and Reviews. For each of these domains, we first identified items likely to be argumentative and then adjusted the guidelines in consultation with expert annotators, as described below.

Debate forums. Of these three domains, *Debates* is the most straightforward to annotate. Given a topic or motion, users can define their stance (*pro*/*contra*) and write an argument which supports it. We included data from two online debate forums. ConvinceMe (CM) is a subset of the IAC, where users share their *Stance* on a topic and discuss their point of view, with replies aiming to change the view of the original poster. Change My View (CMV) is a Reddit forum in which participants post their opinion on a topic and ask others to post replies to change their mind. We sampled original posts from CMV, skipping any moderator posts, and the first reply to an original post from CM, in order to limit the context that annotators must consider when evaluating arguments. CMV posts always include the author's perspective in the title, while CM posts may or may not include a stance in the title. In the guidelines, we instruct annotators to judge a post by how successfully it justifies the author's claim.

CQA. In community questions and answers forums, users post questions or ask for advice, which other users can address. We experimented with arguments from Yahoo! Answers[6] (YA). When posting a question, a user can provide background information for their question (*context*) and can later indicate which response is the *best answer* to their question. The forum's looser structure provides for a wide variety of content, which is appealing as a potential source of non-standard arguments, but challenging as many of the posts do not contain any arguments. Through manual analysis, we identified three categories that frequently contained controversial topics, hypothesizing they would have a higher incidence of debates: *Social Science* > *Sociology*, *Society & Culture* > *Other*, and *Politics & Government* > *Law & Ethics*. We empirically selected the category with the highest proportion of arguments in a study on Amazon Mechanical Turk (MTurk). Qualified annotators[7] decided if question and best-answer pairs were argumentative. We collected 10 judgments for 100 pairs from each category and aggregated judgments with a simple majority. *Law & Ethics* had the most argumentative posts (70%, compared to *Sociology* with 40% and *Society & Culture* with 34%), so we sampled posts from this category to annotate.

In the guidelines for this domain, we asked annotators to judge the argumentative strength of an answer with respect to how well it addressed the given question. The guidelines and subdimension questions were altered to encourage this. One obstacle in pilot studies with expert annotators was posts offering, as many users solicited legal advice in the Law & Ethics forum. We decided to consider advice as argumentative as long as the author supported the advice with justification, which mirrors our general approach to the Argumentative dimension.

Reviews. The third domain consists of restaurant reviews from the Yelp-Challenge-Dataset[8]. On Yelp, users write reviews of businesses and rate the quality of their experience from 1 (low) to 5 (high) stars. Unlike the Debate and Q&A forums, the format of Yelp does not support dialogue between users (i.e., users cannot directly reply to other users or posts), and so it is possible to present each post in isolation as a self-contained argument. As most posts do not explicitly state a claim, we pose the star rating as a claim the user is making about the business, and the review as the argument supporting it.

Yelp reviews can be highly subjective in that each review is based on a single user's experience. For instance, a user may rate a restaurant as 5-stars and write only *The food was delicious* in their review. To address this subjectivity, we asked annotators to judge the argumentative quality of each review with respect to how well it supported the rating provided. Another challenge was defining what constituted a counterargument, as these have a very different character than counterarguments in debates (for example, *Everyone says that the pizza crust is too thin here but that's authentic!*). In consultation with our experts,

[6] `https://answers.yahoo.com/`

[7] HIT approval rate >= 97; HITs approved > 500; Location = US

[8] `https://www.yelp.com/dataset`

Domain	Cogency	Effectiveness	Reasonableness	Overall
CQA	0.16	0.31	0.36	0.29
Debates	0.22	0.33	0.20	0.33
Reviews	0.41	0.19	0.21	0.34

Table 4: Agreement (Krippendorff's α) between experts on pilot studies for CQA, Debates, and Reviews (146, 150, and 50 arguments, respectively).

	Cogency	Effectiveness	Reasonableness	Overall
CQA	0.42	0.52	0.52	0.53
Debates	0.14	0.11	0.21	0.19
Reviews	0.32	0.32	0.31	0.33

Table 5: IAA between the mean expert and crowd scores for Cogency, Effectiveness, Reasonableness, and Overall AQ.

we defined counterarguments by the following characteristics: 1) addressing and rebutting the viewpoints of other reviews, 2) addressing and rebutting points that discredit the author's rating, and 3) bringing up favorable points in an unfavorable review and vice versa.

Experts completed a series of pilots before each domain was presented to the crowd, using the task design described in §3.1. Expert agreement on novel domains (YA and Yelp) are shown in Table 4. Feedback on the task and guidelines was gathered during calibration meetings with experts, and they were iteratively altered to be more clear and specific.

5 A Theory-based AQ Corpus

Applying the annotation task design and data selection described above, we created `GAQCorpus`, containing 5,285 arguments across three domains, annotated for theory-based dimensions. All arguments were limited to have a length between 70 and 200 characters. Ratings were provided by the two groups of annotators described above, Experts (§3.1) and the Crowd (§3.2). Each group judged 3,000 arguments, with about 1,000 arguments annotated by both groups for comparison. The size of the corpus is described in Table 3. Annotators worked with the domains in the following order: Debate forums, CQA forums, and Review forums. Before switching to a new domain, annotators completed a small study for calibration. All data and guidelines are available from `https://github.com/grammarly/gaqcorpus`.

	Crowd	Experts			Overlap	
# Annotators	10	1	2	3	11–13	**Total size**
CQA	1,334	626	–	625	500	**2,085**
Debates	1,438	600	–	600	538	**2,100**
Reviews	600	200	400	–	100	**1,100**

Table 3: Number of arguments annotated by experts and the crowd and the number of overlapping instances (annotated by both experts and the crowd) by domain.

5.1 Inter-annotator Agreements (IAA)

We assessed the quality of the crowd annotations by calculating the agreement between the experts and crowd workers on the overlapping portions of `GAQCorpus` using the mean scores (Table 5).

For debate forums, the agreement is weak with $\alpha \leq 0.21$, while for the CQA forums, the agreement is higher: 0.42–0.53. These results suggest that the difficulty of the task is highly dependent on the domain. While our Debates data and the DS data both consist of web debate arguments, the difference in IAA is high, which might be attributed to different complexities of the web debates data. While TVSP only look at single arguments in isolation, often consisting of a single sentence only.

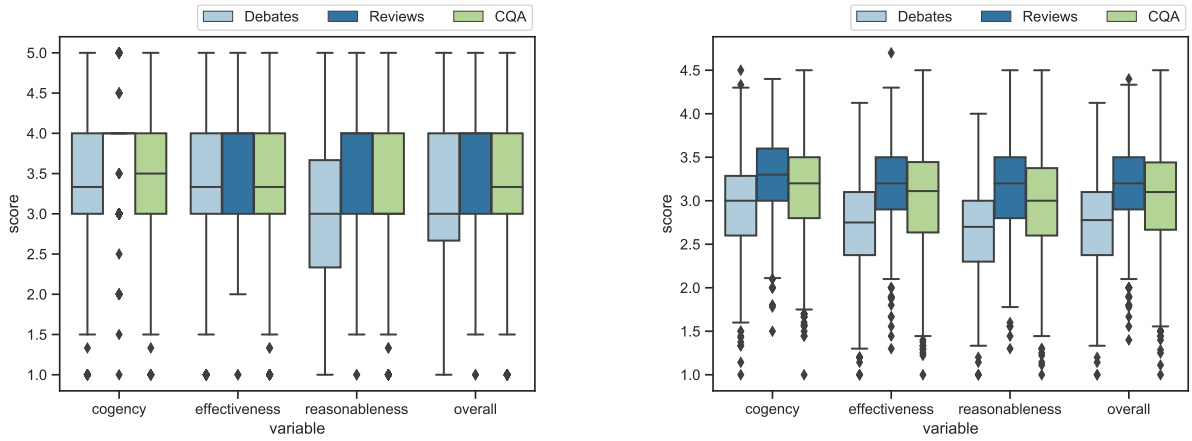
Title: *Should 'blogging' be a capital crime? Iran is considering it...*
Stance: *A government has the right to censor speech (...)*

Text: *My government doesn't give me freedom of speech, so I have to argue for this side. Freedom of speech is bad because ... um ... then Our Leader's beliefs could be challenged. No one wants that. I mean, if everyone would just say and believe what Our Leader says to, we wouldn't need those firing squads altogether! Everyone wins.*

	Cogency	Effectiveness	Reasonableness	Overall
Annotator 1	4	1	1	2
Annotator 2	4	5	3	4
Annotator 3	2	2	2	2

Figure 2: Example argument exhibiting disagreement in the Effectiveness dimension.

One area of disagreement centered on arguments which were sarcastic, ironic, or included rhetorical questions. Consider the argument given in Figure 2, over which the expert annotators expressed



(a) Expert annotations.

(b) Crowd annotations.

Figure 3: Score distributions by domain for expert and crowd annotators.

disagreement. This argument appears to support the stance that a government has the right to censor speech, but several linguistic cues indicate that the argument might be ironic: (a) Punctuation: ellipsis indicates thinking/searching for justifications; similarly, (b) the filler *um*; (c) capitalization: the noun phrase *Our Leader* is capitalized, indicating hyperbolic apotheosis; and finally, (d) the phrase *(...) so I have to argue for this side.* acts like an apologia, which is put in front of the actual argument. Annotators 1 and 2 based their judgments on an interpretation of this text that related to the estimated degree of irony in the post. While Annotator 1 did not perceive irony and judged the argument as *very weak* in *Effectiveness*, Annotator 2 considered it to be highly effective as in their view, the irony positively underlined the perceived stance. Annotator 3 gave medium scores across the board. Such disagreements were regularly discussed and usually revealed that multiple opinions may exist according to how the texts were interpreted, highlighting the high subjectivity of the task.

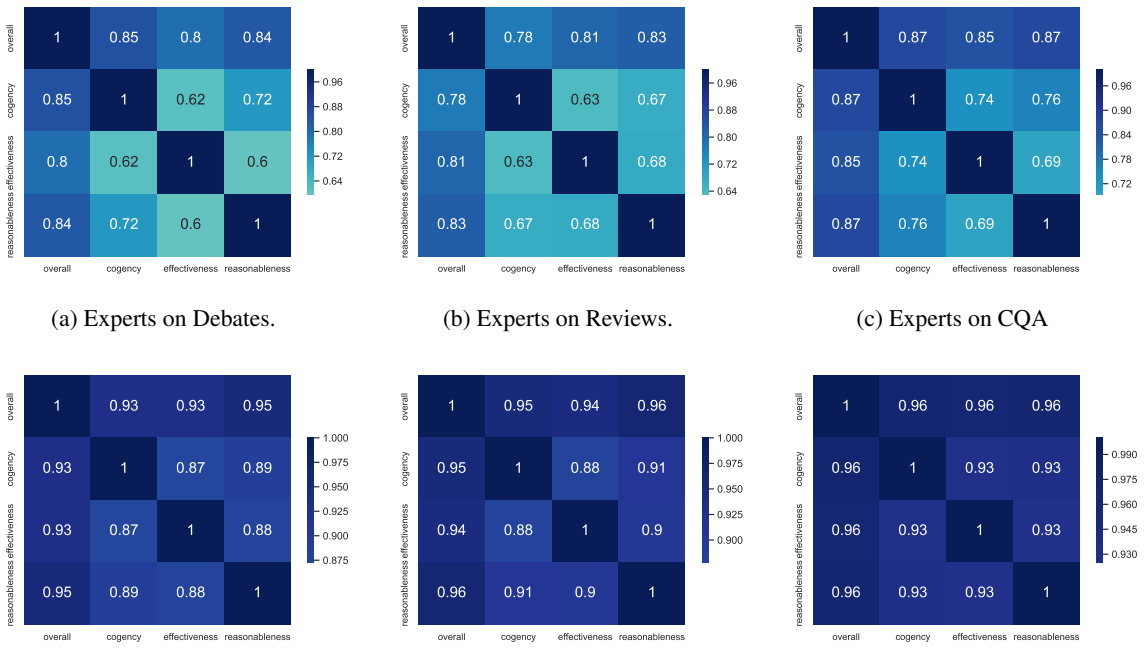
Another area of disagreement was how to judge arguments on topics that were deemed "less worthy" of being discussed, and which usually were humorous in nature or had trivial consequences, such as *Batman vs Superman*, in which users argued for the the superiority of either superhero. In pilots, some experts provided lower ratings of arguments on a topic that they considered less worthy Others thought that writing a strong, serious argument on a less worthy topic was especially difficult, and thus provided higher ratings for such arguments.

5.2 Analysis of Scores

The distributions of mean scores across domains and annotator groups in `GAQCorpus` are depicted in Figures 3a and 3b. In general, the interquartile range of the expert scores was higher than the crowd, suggesting that experts were more specific when scoring items, which is also reflected in the medians: while the crowd exhibits a tendency to score variables equally, expert annotations exhibit more differentiation.

To understand the interrelations between Overall AQ and the dimensions, we compute Pearson correlations between the mean scores (Figure 4). Generally, the trends are similar across all three domains. For instance, for Debates (Figures 4d and 4a), the crowd annotations exhibit stronger correlations between the different dimension scores than the experts, with $0.83 \leq r \leq 0.96$. Interestingly, the variance among the Pearson scores is lower, indicating that the crowd tends to distribute ratings for a single instance more consistently while the experts seem to put more weight on differentiating the dimensions.

Expert ratings of Overall AQ have substantially stronger correlation with the dimensions than any of the dimension scores with each other, further indicating that experts are more discerning in their scores than the crowd. Across both annotator groups and all domains, the correlation between Overall AQ and Reasonableness is highest, which is consistent with earlier observations (Wachsmuth et al., 2017b).



(a) Experts on Debates. (b) Experts on Reviews. (c) Experts on CQA

(d) Crowd on Debates. (e) Crowd on Reviews. (f) Crowd on CQA.

Figure 4: Mean score correlations between the different dimensions for expert and crowd annotators across the three domains (Pearson's r).

5.3 Qualitative Analysis

We next examine low-scoring arguments from all domains to understand how AQ is perceived differently, focusing on the *Reasonableness* dimension. Table 6 shows a low-scoring argument from each domain. The Debate argument raises a counterargument but does not rebut it and additionally neglects to address an obvious counterargument (i.e., the many ethical implications of such a policy). On the other hand, the CQA and Review arguments do not raise or address any counterarguments and are not judged Reasonable for other reasons: the CQA argument jokes about the original poster's question and accuses the poster of malignant behavior, while the Review argument delves into a personal experience that does not contribute to the discussion about the quality of the business.

6 Conclusion

Theory-based AQ assessment provides a holistic and targeted perspective on AQ, but its high complexity makes annotation difficult. In this work, we describe our efforts to create `GAQCorpus`, a multi-domain corpus of 5,295 arguments annotated for quality along theory-based AQ dimensions. We demonstrate that it is possible to collect complex annotations with crowdsourcing in three domains: Debate, CQA, and Review forums. Drawing from the initial study of Wachsmuth et al. (2017a), which suggested the general feasibility, we relied on the intuition of trained linguists to simplify the task and guidelines while preserving the theoretical basis of the task. The agreement between experts and the crowd was higher than the agreement in earlier studies (Wachsmuth et al., 2017a), validating our approach and indicating that it is possible to collect complex ratings using a crowd.

`GAQCorpus` and the findings of our annotation study will serve as a basis for future corpus development and computational model development in theory-based AQ. They are available for download from `https://github.com/grammarly/gaqcorpus`.

Debates		
Cogency	2.0	**Title:** Should you need to pass an IQ test to have kids?– **Stance:** Dumb parents lead to more dumb kids. **Text:** I have a strong opinion that before having children, the prospective parents should have to pass a series of background and IQ tests. Kids being brought into this world need a good foundation to start a successful life with. You may have that limited case where the parents are morons and the kids strive to be different then their failure parents, but in most cases it is an endless line of parasites on our world. We need more smart people.
Effectiveness	1.7	
Reasonableness	1.0	
Overall	1.3	
CQA		
Cogency	2.7	**Question:** Bounced CHECK? **Context:** Does the company holding the bounced check have to send you a certified letter before issuing a warrant for your arrest. I feel almost certain that they do but i am not sure. **Answer:** I always make sure my checks are not printed on rubber. they are just too expensive and not worth it. We all make a mistake from time to time, and usually it is no big deal except for the extreme annoyance and all the bounced check fees. But if you are worried about an arrest warrant then I am sure you are doing this deliberately and trying to defraud the company. You have probably sent them a couple of bad checks already in an attempt to string them along so your guilt is probably pretty well established. You can hope that you do not have to share a jail cell with a gross deviate of some sort.
Effectiveness	2.0	
Reasonableness	1.7	
Overall	2.0	
Reviews		
Cogency	1.0	**Title:** Business review: 2.0 Stars. **Business name:** Cook Out. **City:** Charlotte. **Categories:** Restaurants, Desserts, Food, Fast Food, American (Traditional), Hot Dogs, Burgers **Review:** Burgers are good but I like those other 5 guys burgers instead oh and I guess if your not from around here don't even think about going thru the drive thru it's like the biggest most unreadable confusing hurried crazy thing ever if I ever go again hell with drive thru until I've lived here for at least 5 maybe 10 years and can be a veteran drive thru person I'm walking in it's like if I mix up all the letters in this review and give you 1 minute to read it and figure it out then you gotta move on.
Effectiveness	1.0	
Reasonableness	1.0	
Overall	1.0	

Table 6: Low-scoring arguments from all domains

Acknowledgements

The work of Anne Lauscher is supported by the Eliteprogramm of the Baden-Württemberg Stiftung (AGREE grant). We thank our linguistic expert annotators for providing interesting insights and discussions as well as the anonymous reviewers for their helpful comments. We also thank Henning Wachsmuth for consulting us w.r.t. his previous work and Yahoo! for granting us access to their data.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4445–4452, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Aristotle. trans. 2007. *On Rhetoric: A Theory of Civic Discourse*. Oxford University Press, Oxford, UK. Translated by George A. Kennedy.

Eli P. Cox III. 1980. The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17(4):407–422.

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 AAAI2020*.

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, Berlin, Germany, August. Association for Computational Linguistics.

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

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*.

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, Sofia, Bulgaria, August. 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, Baltimore, Maryland, June. 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.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 4082–4088. AAAI Press.

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*, EMNLP '10, pages 229–239, Stroudsburg, PA, USA. 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.

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, Valencia, Spain, April. Association for Computational Linguistics.

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 5629–5639.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada, July. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. 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, Valencia, Spain, April. Association for Computational Linguistics.