A Corpus of Sentence-level Annotations of Local Acceptability with Reasons

Wonsuk Yang **Jung-Ho Kim** **Seungwon Yoon**
ChaeHun Park **Jong C. Park**[†]
School of Computing
Korea Advanced Institute of Science and Technology
{derrick0511, jhkim, swyoon, ddehun, park}@nlp.kaist.ac.kr

Abstract

News editorials are presented with arguments of different quality, which readers may or may not accept as presented. In this work, we present a corpus of news editorials with sentence-level annotations of local acceptability and a set of related attributes, where the annotators also provided detailed reasons in natural language for each attribute. The annotations were performed in both in-house and crowdsourcing environments. In total, 105 news editorials were annotated for 3,591 sentences, with an average of four annotations from different annotators per sentence, resulting in about 14K sentence-level annotations with detailed reasons in natural language (in total 121K tokens written in 1K aggregated hours). We analyze the reasons to see why given information is accepted or rejected, examine the correlation among the attributes, and compare our annotation result against argumentation strategies. This is the first corpus to provide sentence-level annotations with attributes such as local acceptability, which we argue is critical for a fine-grained and advanced analysis of argumentation quality.

1 Introduction

Argumentation is an activity aimed at promoting the acceptability of a controversial standpoint (van Eemeren et al., 1996; Stab and Gurevych, 2017). The research on argumentation quality is of much relevance to computer-assisted writing. *Local acceptability* is considered as one of the important factors that influence the quality of argumentation in writing, which is defined for the premises of an argument as to how much they are "rationally worthy of being believed to be true" (Wachsmuth et al., 2017). It also reveals the process where a particular reader accepts or rejects information delivered by each premise presented in an argument while reading it. A deeper understanding into this process is thus vital for the development of computer-assisted writing systems that can identify weak spots of an author's argument.

Development of such a system has also been considered as an important goal of many studies on computational argumentation such as parsing argumentation structures (Stab and Gurevych, 2017). Recently, there have been studies on the annotation and analysis of eloquence, evidence, and more attributes that give multiple types of feedback on persuasiveness for effective writing support systems (Carlile et al., 2018). However, persuasiveness is an indicator of how "persuasive" a particular sentence is, and takes a different dimension from local acceptability (Wachsmuth et al., 2017). For instance, it is reported that the inter-annotator agreement for persuasiveness (0.5-0.7 alpha, (Carlile et al., 2018)) is quite higher than that for local acceptability (0.22-0.45 alpha, (Wachsmuth et al., 2017)), indicating that the latter is arguably more influenced by individual subjectivity. Wachsmuth et al. (2017) proposed local acceptability as one of the 15 factors affecting argumentation quality, with which they also annotated textual debate portal arguments for two stances on several issues, such as evolution vs. creation. However, as all the attributes were scored only with a 1-3 scale and with argument-level (paragraph-

[†]Corresponding author

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 288-297, Hakodate, Japan, September 13-15, 2019

	Score	Description
Local Acceptability	7	I **strongly accept** the information given by the sentence to be true. I have sound and cogent arguments to justify my acceptance.
	6	I **accept** the information given by the sentence to be true. I have some arguments to justify my acceptance.
	5	I **weakly accept** the information given by the sentence to be true. I do not have arguments justifying my acceptance. Still, I will accept it rather than reject it.
	4	It is **hard to judge** whether I should accept or reject the information given by the sentence to be true.
	3	I **weakly reject** the information given by the sentence to be true. I do not have arguments for the rejection. Still, I will reject it rather than accept it.
	2	I **reject** the information given by the sentence to be true, and I have arguments for the rejection.
	1	I **strongly reject** the information given by the sentence to be true. I have sound and cogent arguments for the rejection.
Knowledge Awareness	3	I **already knew** the information before I read this document.
	2	I **did not know** the information before I read this document, **but came to know** it by reading the previous sentences in this document.
	1	I **did not know** the information.
Verifiability	5	I can verify it **using my knowledge**. It is a common sense. I do not need to google it to verify.
	4	I can verify it by **short-time googling**.
	3	I can verify it by **long-time googling**. I could verify it using deduction if I google it for some time for deeper understanding.
	2	I might find an **off-line way** to verify it, but it will be very hard. It needs specific witness or testimony to verify, and there may not be any evidence in written form.
	1	There is **no way to verify** it.
Disputability	4	Whether or not it is reasonable to accept the information given by the sentence as true, it is **not disputable**.
	3	Whether or not it is reasonable to accept the information given by the sentence as true, it is **weakly disputable**.
	2	Whether or not it is reasonable to accept the information given by the sentence as true, it is **disputable**.
	1	Whether or not it is reasonable to accept the information given by the sentence as true, it is **highly disputable**.

Table 1: Description of local acceptability, knowledge awareness, verifiability, and disputability.

level) granularity, it is not possible to track a specific process where the reader accepts or rejects each premise presented in the argument. This is important because knowing at which sentence a reader starts to reject the given information enables the writing support system to determine from where it needs to start editing.

In this work, we present a corpus of sentence-level annotations of local acceptability and a set of predefined, possibly related attributes for each sentence in news editorials. For a deeper understanding into the individual judgments, we also collected the reasons for the particular attribute values by each annotator on each sentence. The corpus can thus be utilized to experimentally verify to what extent it is possible to predict the very subjective reaction of the target readers accepting or rejecting the information given by each sentence.

The contributions of this paper are as follows. (1) We introduce a large corpus of 105 news editorials with 14K sentential annotations for local acceptability and three possibly related attributes, with reasons for each attribute written in natural language, in both in-house and crowdsourced settings. The gathered text for the reasons amounts to 121K tokens written in 1K aggregated hours[1]. (2) We show experimentally that the three attributes are meaningfully correlated with local acceptability. (3) We provide a detailed analysis of the reasons provided by the annotators for each of the attributes. (4) We provide key insights through the comparison against argumentation strategy, and suggest that our corpus can be utilized for future research that leads to computational argumentation.

2 Related Work

Traditionally, research on computational argumentation has focused on parsing argumentation structures, whose goal is to identify argument components of a given document such as major claims, other claims, and premises, and to analyze how they are related to one another (Stab and Gurevych, 2014a; Stab and Gurevych, 2014b; Stab and Gurevych, 2017).

Recently, research on the quality of argumentation has received much attention. Persing and Ng (2015) defined argument strength based on the expected number of readers who would be persuaded, and annotated student essays with it. For convincingness, which concerns the universal audience (Perelman et al., 1969), there has been an in-depth study to see which of two given arguments is more convincing and why (Habernal and Gurevych, 2016a; Habernal and Gurevych, 2016b). For persuasiveness, which concerns a particular audience

[1]Our corpus is available at http://credon.kaist.ac.kr.

Local Acceptability	strong accept	statement factual nature, author state facts, personally agree statement, evidence support claim, commonly know fact
	accept	author quote another, author sufficient credibility, would grind author, author focus part, reveal subjective interpretation
	weak accept	author reveal subjective, reveal subjective interpretation, author sufficient credibility, likely thing happen, not know enough
	hard to judge	not enough background, not know enough, enough background knowledge, make judgements statement, not enough knowledge
	weak reject	personally not agree, author lack credibility, not agree statement, author reveal subjective, opinion not fact
	reject	controversial author opinion, claim base controversial, claim controversial difficult, one side opinion, long search would
	strong reject	provide counter examples, evidence easily dismiss, not logical reason, highly inaccurate projections, find speculative highly

Table 2: The top five most frequent trigrams (lemmatized verbs, cf. Section 4.3.1) used for describing the reason for local acceptability. (*grind* is a lemmatization error.)

(Perelman et al., 1969), Carlile et al. (2018) annotated argument components, such as major claims, other claims, and premises, with component-specific sub-attributes such as eloquence and evidence. Ke et al. (2018) developed neural network systems to predict persuasiveness and sub-attributes. Tan et al. (2016) utilized the Reddit forum ChangeMyView, used the record of user interactions as the proxy to measure the persuasiveness of arguments in the forum, and studied multiple factors that infludence persuasiveness. El Baff et al. (2018) modeled the argumentation quality of news editorials based on whether they challenge or reinforce the stance of the reader, and gathered document-level annotations for 1,000 news editorials on the model. In addition, as part of the SemEval 2018 shared task, the Argument Reasoning Comprehension task (Habernal et al., 2018) has been offered to select the appropriate warrant for a given argument consisting of a claim and a reason. While we focus on local acceptability in this work, there is another type of acceptability of an argument or argumentation, called global acceptability. It is investigated by Cabrio and Villata (2012) who identified ground-truth debate portal arguments using textual entailments based on the formal argumentation framework of Dung (1995), and assessed the global acceptability of the arguments.

3 Data

The source data we choose to annotate is composed of 105 news editorials randomly chosen from the Webis-Editorials-16 corpus provided by Al-Khatib et al. (2016), which includes news editorials published by Al-Jazeera, FoxNews, and the Guardian. This corpus classifies argumentative discourse units (ADUs) into the following six types for the analysis of argumentation strategies: (1) *Common Ground*, (2) *Assumption*, (3) *Testimony*, (4) *Statistics*, (5) *Anecdote*, and (6) *Other*.

We used their corpus as the basis for our annotation for the following three reasons. (1) The Webis-Editorial-16 corpus is annotated with reliable quality. (2) The corpus has been used as the basis for an analytical study of topic-dependent argumentation strategies (Al-Khatib et al., 2017) and has inspired subsequent research on argumentation synthesis through rhetoric strategies (Wachsmuth et al., 2018). (3) We anticipate that the six types used for the analysis of argumentation strategies would be closely related to local acceptability. For example, the local acceptability for *Statistics* is expected to be higher than that for *Anecdote*. Therefore, we anticipate that the annotations of argumentation strategy will help the quality assessment of local acceptability annotation (see Section 4.3.5).

4 Annotation

4.1 Annotation Scheme

In this study, we annotate each of the sentences in news editorials with local acceptability and a set of predefined attributes related to it.

Our definition of local acceptability follows Wachsmuth et al. (2017), who defined local accept-

Figure 1: The word cloud for the trigram frequencies for *accept* (left) and *reject* (right) for local acceptability.

ability of a premise as: A premise is locally acceptable if it is "rationally worthy of being believed to be true." It is noted that an argument is composed of a claim and one or more premises, where a premise is a reason for justifying (or refuting) a claim and the claim is a possibly controversial statement and the central argument component (Stab and Gurevych, 2017). We define the local acceptability of a sentence based on the truth-value of the sentence following the *truth-conditional theory* (Lewis, 1970): A sentence is locally acceptable if the truth-value of the sentence is rationally worthy of being believed to be true. For complex cases where the truth-values of the phrases in a sentence are combined to generate the truth-value of a sentence, we also follow Lewis (1970).

We also annotated three possibly related attributes as follows. *Knowledge Awareness* asks whether or not an annotator already knew the information. *Verifiability* indicates how easy it is to verify the information. *Disputability* is about how controversial the information is. We chose the attributes for a deeper understanding of local acceptability, focusing on fact-checking (Wintersieck et al., 2018) and journalistic aspects (Cheruiyot and Ferrer-Conill, 2018; Aharoni and Tenenboim-Weinblatt, 2019). Table 1 shows detailed rubrics for the local acceptability and the three attributes that we used for our annotation.

4.2 Annotation Procedure

4.2.1 In-house Annotations

An in-house annotation was conducted on 105 news editorials by eight undergraduate students with native competence in English, where four of them were student journalists responsible for the school newspaper. We introduced the rubrics to the students through one seminar so that they familiarized themselves with the rubrics over one week. Then, for the following two weeks, each student had two separate meetings with the authors for further instruction on the rubrics. The whole annotation process took about 6 months.

4.2.2 Crowdsourcing Annotations

We also conducted the annotation through the Amazon Mechanical Turk (AMT) crowdsourcing platform. Each AMT annotator received specific annotation guidelines, including a detailed description of such cases that could cause confusion, identified as such during the training period of the in-house annotation. In each assignment, the workers were presented with a URL for a news editorial. At the URL, they were asked to annotate following the guidelines, and to write the reasons for choosing each attribute value. The crowdsourced annotation took

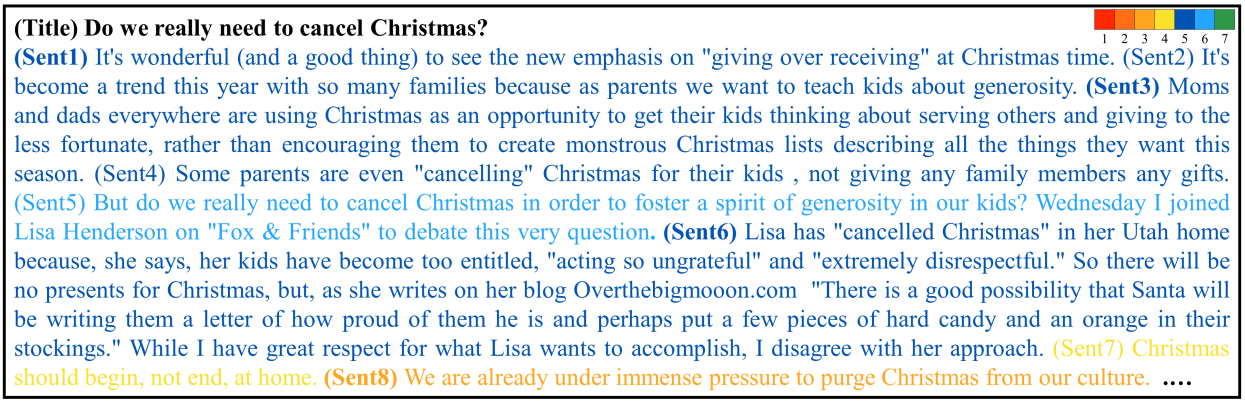

Figure 2: Example news editorial from the presented corpus. The color coding indicates the mean value of the Local Acceptability of each sentence across the three annotators. We rounded each mean value to the first decimal place. (article link: `https://www.foxnews.com/opinion/do-we-really-need-to-cancel-christmas`)

about two weeks, where it started after the in-house annotation was over. During the whole annotation process, 183 workers (with $\geq$ 95% acceptance rate for previous tasks in AMT) participated, where one annotator was allowed to annotate multiple news editorials.

4.2.3 Corpus Statistics

As a result, per news editorial, we obtained annotations from an average of 1.3 different students and from an average of 2.7 different workers (resulting in an average of 4.0 annotators per news editorial), amounting to 14,225 sentence-level annotations for 3,591 sentences in 105 news editorials in total. The students and workers spent about 4 and 5 minutes on average, respectively, to annotate a single sentence including the time to read and select the attribute values and to write down specific reasons, in total 1,108 aggregated hours for the 14,225 anntotations.

4.3 Analysis of Annotations

4.3.1 Reason

The students participating in the in-house annotation described the reasons using 4.2 words per attribute value on average, and the workers used 10.6 words on average. In calculating the average length, we filtered out all non-English words using Natural Language Toolkit (NLTK) [2] and did not count them.

Table 2 shows the top five most frequent trigrams for local acceptability, and Figure 1 shows word clouds of trigram frequencies of the reasons specified for the attribute values of *accept* and *reject*. For the trigram frequency, we preprocessed the reasons written in natural language as follows. We first removed all non-English words, stop words except for *not*, and punctuation marks from the reasons specified by the annotators. We also removed the words that are used to describe the corresponding attribute value itself in the rubrics, but did not remove those that are used to explain other attribute values. For example, in the case of *accept* of local acceptability, words such as *justify* that appear in the explanation of *accept* were removed, but words such as *cogent* that appear in the explanation of *strong accept* were not removed. We lemmatized verbs and measured trigram frequencies across all the (preprocessed) reasons. When one annotator repeatedly used a certain trigram for a specific attribute, it was counted only once.

From the trigram frequency, we find that the major reasons to accept a piece of information are factuality, related evidence, third party information source, and prominence of the information. The reasons to reject are controversy, bias, counter evidence, logical inconsistency, and non-factual nature of the information.

4.3.2 Example

In order to illustrate the overall process of how the (three) annotators accepted or rejected the sentences while reading a given news editorial, we present eight example sentences and the average of the three

[2] http://www.nltk.org

Sentences		**LA**	KA	V	D
(Sent3) Moms and dads everywhere are using Christmas as an opportunity to get their kids thinking about serving others and giving to the less fortunate , rather than encouraging them to create monstrous Christmas lists describing all the things they want this season .	Student1	(5) weak accept, because no counter evidence to reject	(1) did not know	(1) no way to verify, because general statements cannot be verified	(2) weakly disputable, because (…), as some parents may not agree
	Worker1	(3) weak reject, because it seems to be a rhetorical device to make a point.	(1) did not know, becaus e I am not aware that this is a widespread behavior.	(3) long-time googling, because it might be possi ble to find many individu al cases online (…)	(3) disputable, because this is an unsourced opinion.
	Worker2	(6) accept, because I can verify and provide supporting examples	(3) already knew, because (…) my family is trying to make [Christmas] less focused on material things	(4) short-time googling, because you can find blogs and stories of families doing this	(3) disputable, because y ou can argue that there is only a small group of pe ople doing this and that it isn't a Movement
(Sent6) Lisa has "cancelled Christmas" in her Utah home because, she says, her kids have become too entitled, "acting so ungrateful" and "extremely disrespectful." So there will be no presents for (…) While I have great respect for what Lisa wants to accomplish, I disagree with her approach.	Student1	(5) weak accept, because no reason to reject author 's personal opinion	(1) did not know	(1) no way to verify, because author's personal opinion on Lisa's approach to Christmas	(2) weakly disputable, because personal opinions on Christmas may vary
	Worker1	(6) accept, because it is what Lisa claims it is, and what the other speaker stated.	(1) did not know, because Lisa, her plan (…) are all introduced for the first time here.	(4) short-time googling, because it is from her blogoverthebigmoon.co m. The video (…)	(2) weakly disputable, because this is second hand information, and a personal opinion.
	Worker2	(5) weak accept, because I accept their points of view but it is hard to accept this statement as an absolute fact	(1) did not know, because (…) there is no way I would know this without knowing her or being one of her blog readers	(4) short-time googling, because you could read her blog article or talk with her and the interviewer to confirm her opinion	(2) weakly disputable, because (…) on the topic but you [could] if you found scientific or other evidence that proved them wrong
(Sent8) We are already under immense pressure to purge Christmas from our culture.	Student1	(4) hard to judge, because do not fully agree with the statement, but no reason to reject	(1) did not know	(1) no way to verify, because general statements cannot be verified	(3) disputable, because general statements can be controversial, especially involving the audience ("we")
	Worker1	(1) strong reject, because there is no immense pressure to purge Christmas in "our culture"	(1) did not know, because I know this to be untrue.	(4) short-time googling, because this is a common [right-wing] Christian view, but it does not represent reality.	(4) highly disputable, because this is simply not happening.
	Worker2	(3) weak reject, because I need more evidence to prove this is true without it I'm not sure if I believe the claims in this Story	(3) already knew, because I don't agree with this statement but [i] am aware there are people who believe this	(1) no way to verify, because this is an opinion and you can find articles (…)	(4) highly disputable, because some people believe there is a war on [christmas] (…)

Figure 3: The annotation results for the three sentences that we find the most interesting in the paragraph of Figure 2. Each color bar above the column name indicates the color coding for the values of the corresponding attribute in the column.

attribute values for local acceptability for each sentence (Figure 2). It is noted that the annotators start to reject the given sentence at Sent 8.

Moreover, we present the annotation results of the three sentences (Sent3, 6, 8) that we find the most interesting among the annotation results of the eight sentences (Figure 3). In the case of Sent3, we see that each annotator focused on different aspects: (1) counter-evidence (Student1), (2) rhetoric device (Worker1), and (3) verifiability and supporting examples (Worker2). Thus, they made different, subjective judgments on local acceptability for the same sentence. Sent6 was relatively more complex, and we found that each annotator focused on a different phrase in the sentence in accepting or rejecting it. In the case of Sent8, all three annotators responded negatively to local acceptability or gave the *hard to judge* response.

	Pearson Correlation Coefficient			
	LA	KA	V	D
LA		.235 (.154)	.466 (.402)	-.658 (-.481)
KA	.235 (.154)		.407 (.285)	-.118 (.040)
V	.466 (.402)	.407 (.285)		-.425 (-.397)
D	-.658 (-.481)	-.118 (.040)	-.425 (-.397)	

Table 3: Correlation between local acceptability and the related attributes. All the p-values were less than 0.001. The numbers in parentheses are for the in-house annotation only, whereas the numbers outside the parentheses are for the entire (i.e., both in-house and crowdsourced) annotation.

Overall, the example annotation results show that both the in-house students and the AMT workers produced high-quality annotations, as indicated by the overall length and specific content of the reasons. We also note that they tend to make a highly subjective, yet reasonable judgment, and that each of their judgments cannot be considered irrational simply because of such subjectivity and uniqueness.

4.3.3 Inter-Annotator Agreement

As can be seen in the examples in Section 4.3.2, the annotators responded to a sentence differently based on their (different) viewpoints. Even for the case where different annotators chose the same attribute value, the reason was not necessarily the same. Even when the annotators chose different attribute values, their reasons for the choices were apparently valid. From the fact that we can argue for the quality of the annotations in Figure 3 based on the validity of the reasons suggested by the annotators, we see that (the validity of) the reason itself is a good indicator for the quality of annotation. We do not use inter-annotator agreement (IAA) as the quality measure for our corpus in this paper. In Yang et al. (2019), we present an in-depth analysis on the IAA for local acceptability with a new method of quality control that uses the validity of the reasons.

4.3.4 Correlation Analysis

To understand the significance of how the attributes are related to local acceptability, we computed the Pearson's Correlation Coefficient (PC) between local acceptability and each of the attributes along with the corresponding p-values. Results are as shown in Table 3. For all of the calculated PC, the p-value was less than 0.001.

Overall, local acceptability and disputability show a high negative correlation. This is not surprising because it is hard to (strongly) accept highly disputable information. Local acceptability and verifiability show a positive correlation. We speculate that this is because it is less likely that an author delivers a piece of misinformation about easily verifiable information and it is more likely to be acceptable information. Local acceptability and knowledge awareness have a positive correlation. We speculate that this is because the information already known to a reader would be easier to be accepted by that reader. The results show that the three attributes that we speculated as possibly related are actually related to local acceptability.

4.3.5 Local Acceptability and Argumentation Strategy

We look into the relationship between our annotation results and previously annotated argumentation strategies for further insights. Note that during the entire annotation process, the annotators were not provided with information about the argumentation strategy pre-annotated in the corpus, and conducted annotation only on plain text.

Of a total of 3,591 sentences in the 105 news editorials, 3,150 (87.7%) sentences are found to contain only a single type of argumentation strategy (other than the type *Other*), and 441 (12.3%) sentences contain more than one type or do not contain any type. We counted the annotations for the same sentence by different annotators separately (as different annotations). As we note that the labels for argumentation strategy type were imbalanced, we compared the relative frequency (the ratio in the table) of the attribute values for each type. For example, we found that 1,802 annotations on 458 sentences were mapped to the type *Anecdote*, and that 657 (36.5%) of them had the local acceptability of *strong accept*. In this case, the relative frequency of *strong accept* for *Anecdote* is 36.5%. For local acceptability and each of the other attributes, we compared the relative frequencies across different argumentation strategy types in the same way, as shown in Table 4.

Strong accept of local acceptability occurred most frequently with *Statistics*, *Anecdote*, and *Common Ground*, whereas the relatively lower value *accept* occurred most frequently with *Assumption* and *Testimony* (not counting *N/A*). The local acceptability for

		Count (Ratio)						Total
		AS	AN	ST	TE	CO	N/A	
LA	strong accept	1461 (15.8)	657 (**36.5**)	200 (**40.9**)	234 (30.1)	70 (**36.3**)	383 (22.0)	3005 (21.1)
	accept	2341 (**25.4**)	553 (30.7)	175 (35.8)	305 (**39.2**)	66 (34.2)	498 (**28.6**)	3938 (**27.7**)
	weak accept	2232 (24.2)	373 (20.7)	73 (14.9)	127 (16.3)	28 (14.5)	343 (19.7)	3176 (22.3)
	hard to judge	1595 (17.3)	132 (7.3)	13 (2.7)	52 (6.7)	15 (7.8)	334 (19.2)	2141 (15.1)
	weak reject	719 (7.8)	51 (2.8)	12 (2.5)	38 (4.9)	7 (3.6)	87 (5.0)	914 (6.4)
	reject	573 (6.2)	23 (1.3)	11 (2.2)	12 (1.5)	4 (2.1)	64 (3.7)	687 (4.8)
	strong reject	300 (3.3)	13 (0.7)	5 (1.0)	10 (1.3)	3 (1.6)	33 (1.9)	364 (2.6)
KA	already knew	2137 (23.2)	258 (14.3)	44 (9.0)	47 (6.0)	125 (**64.8**)	255 (14.6)	2866 (20.1)
	came to know	1233 (13.4)	135 (7.5)	65 (13.3)	148 (19.0)	15 (7.8)	194 (11.1)	1790 (12.6)
	did not know	5851 (**63.5**)	1409 (**78.2**)	380 (**77.7**)	583 (**74.9**)	53 (27.5)	1293 (**74.2**)	9569 (**67.3**)
V	using my knowledge	1284 (13.9)	141 (7.8)	26 (5.3)	26 (3.3)	82 (**42.5**)	142 (8.2)	1701 (12.0)
	short-time googling	1712 (18.6)	552 (30.6)	232 (**47.4**)	296 (**38.0**)	42 (21.8)	477 (**27.4**)	3311 (23.3)
	long-time googling	1971 (21.4)	557 (**30.9**)	211 (43.1)	264 (33.9)	39 (20.2)	445 (25.5)	3487 (**24.5**)
	off-line way	763 (8.3)	275 (15.3)	12 (2.5)	125 (16.1)	5 (2.6)	170 (9.8)	1350 (9.5)
	no way to verify	2610 (**28.3**)	229 (12.7)	4 (0.8)	53 (6.8)	19 (9.8)	281 (16.1)	3196 (22.5)
	none of the above	881 (9.6)	48 (2.7)	4 (0.8)	14 (1.8)	6 (3.1)	227 (13.0)	1180 (8.3)
D	not disputable	2562 (27.8)	1318 (**73.1**)	358 (**73.2**)	496 (**63.8**)	101 (**52.3**)	882 (**50.6**)	5717 (**40.2**)
	weakly disputable	2875 (**31.2**)	298 (16.5)	82 (16.8)	139 (17.9)	56 (29.0)	425 (24.4)	3875 (27.2)
	disputable	2765 (30.0)	141 (7.8)	34 (7.0)	115 (14.8)	25 (13.0)	326 (18.7)	3406 (23.9)
	highly disputable	1019 (11.1)	45 (2.5)	15 (3.1)	28 (3.6)	11 (5.7)	109 (6.3)	1227 (8.6)
	Total	9221 (100)	1802 (100)	489 (100)	778 (100)	193 (100)	1742 (100)	14225 (100)

Table 4: Occurrence frequencies of the attribute values for each argumentation strategy type. The number in parenthesis is the relative occurrence frequency of an attribute value (row index) for an argumentation strategy type (column index). For each strategy type, the highest value of the relative frequency for each attribute is marked bold. AS, AN, ST, TE, and CO indicate *Assumption*, *Anecdote*, *Statistics*, *Testimony*, and *Common Ground*, respectively. N/A indicates the annotations for the sentences that contain more than one type or do not contain any type (other than the type *Other*).

Assumption was higher than we initially expected. We speculate that this is due to the accumulated credibility of the publishers (Al-Jazeera, FoxNews, and the Guardian) and/or their authors, which may indicate the importance of the author credibility or authority perceived by the readers.

For knowledge awareness, *did not know* occurred most frequently with all the types except for *Common Ground*, and *already knew* occurred most frequently with *Common Ground*. For verifiability, *using my knowledge* occurred most frequently with *Common Ground*, whereas *no way to verify* occurred most frequently with *Assumption*. *Short-time googling* occurred most frequently with *Statistics* and *Testimony*, and *long-time googling* occurred most frequently with *Anecdote*. For disputability, *not disputable* occurred most frequently with all the types except for *Assumption*, whereas *weakly disputable* occurred most frequently with *Assumption*. As such, we find that the relationship between the argumentation strategies and our attributes indicated by the relative frequencies precisely matches good linguistic intuition, suggesting further that our corpus is overall of remarkable quality.

5 Conclusion

In this study, we presented a corpus of 105 news editorials, in which each sentence is annotated with local acceptability and a predefined set of related attributes where reasons for each attribute are also provided in natural language by the annotators. We collected the reasons accepting or rejecting the information given by each sentence, in total 121K tokens written in 1K aggregated hours. A detailed analysis demonstrates that our corpus is overall of remarkable quality. We anticipate that our corpus, the first of its kind at sentence-level annotation with notes, may be utilized meaningfully for computer-assisted writing, as well as for a deeper understanding into the argumentation strategy. Our corpus is available at http://credon.kaist.ac.kr.

Acknowledgements

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00582-002, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection).

References

Tali Aharoni and Keren Tenenboim-Weinblatt. 2019. Unpacking journalists(dis) trust: Expressions of suspicion in the narratives of journalists covering the israeli-palestinian conflict. *The International Journal of Press/Politics*.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1351–1357.

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 (ACL, Short Papers)*, volume 2, pages 208–212.

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

David Cheruiyot and Raul Ferrer-Conill. 2018. "Fact-Checking Africa" epistemologies, data and the expansion of journalistic discourse. *Digital Journalism*, 6(8):964–975.

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.

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

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

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

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. SemEval-2018 Task 12: The Argument Reasoning Comprehension Task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 763–772.

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to Give Feedback: Modeling Attributes Affecting Argument Persuasiveness in Student Essays. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4130–4136.

David Lewis. 1970. General semantics. *Synthese*, 22(1):18–67.

Chaim Perelman, Lucie Olbrechts-Tyteca, John Wilkinson, and Purcell Weaver. 1969. *The new rhetoric: a treatise on argumentation*. University of Notre Dame Press.

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 (ACL-IJCNLP, Long Papers)*, volume 1, pages 543–552.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 1501–1510.

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

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

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion

strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pages 613–624.

Frans H. van Eemeren, Rob Grootendorst, Francisca Snoeck Henkemans, J. Anthony Blair, Ralph H. Johnson, Erik C. W. Krabbe, Christian Plantin, Douglas N. Walton, Charles A. Willard, John Woods, and David Zarefsky. 1996. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments.* Lawrence Erlbaum Associates, Inc.

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

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation Synthesis following Rhetorical Strategies. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3753–3765.

Amanda Wintersieck, Kim Fridkin, and Patrick Kenney. 2018. The message matters: The influence of fact-checking on evaluations of political messages. *Journal of Political Marketing*, pages 1–28.

Wonsuk Yang, Seungwon Yoon, Ada Carpenter, and Jong C. Park. 2019. Nonsense!: Quality control via two-step reason selection for annotating local acceptability and related attributes in news editorials. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). (*to appear).