Annotating Multiparty Discourse: Challenges for Agreement Metrics

Nina Wacholder* Smaranda Muresan† Debanjan Ghosh* Mark Aakhus*
*School of Communication and Information, Rutgers University
†Center for Computational Learning Systems, Columbia University
`ninwac|debanjan.ghosh|aakhus@rutgers.edu, smara@ccls.columbia.edu`

Abstract

To computationally model discourse phenomena such as argumentation we need corpora with reliable annotation of the phenomena under study. Annotating complex discourse phenomena poses two challenges: fuzziness of unit boundaries and the need for multiple annotators. We show that current metrics for inter-annotator agreement (IAA) such as P/R/F1 and Krippendorff's α provide inconsistent results for the same text. In addition, IAA metrics do not tell us what parts of a text are easier or harder for human judges to annotate and so do not provide sufficiently specific information for evaluating systems that automatically identify discourse units. We propose a hierarchical clustering approach that aggregates overlapping text segments of text identified by multiple annotators; the more annotators who identify a text segment, the easier we assume that the text segment is to annotate. The clusters make it possible to quantify the extent of agreement judges show about text segments; this information can be used to assess the output of systems that automatically identify discourse units.

1 Introduction

Annotation of discourse typically involves three subtasks: segmentation (identification of discourse units, including their boundaries), segment classification (labeling the role of discourse units) and relation identification (indicating the link between the discourse units) (Peldszus and Stede, 2013a). The difficulty of achieving an Inter-Annotator Agreement (IAA) of .80, which is generally accepted as good agreement, is compounded in studies of discourse annotations since annotators must unitize, i.e. identify the boundaries of discourse units (Artstein and Poesio, 2008). The inconsistent assignment of boundaries in annotation of discourse has been noted at least since Grosz and Sidner (1986) who observed that although annotators tended to identify essentially the same units, the boundaries differed slightly. The need for annotators to identify the boundaries of text segments makes measurement of IAA more difficult because standard coefficients such as κ assume that the units to be coded have been identified before the coding begins (Artstein and Poesio, 2008). A second challenge for measuring IAA for discourse annotation is associated with larger numbers of annotators. Because of the many ways that ideas are expressed in human language, using multiple annotators to study discourse phenomena is important. Such an approach capitalizes on the aggregated intuitions of multiple coders to overcome the potential biases of any one coder and helps identify limitations in the coding scheme, thus adding to the reliability and validity of the annotation study. The more annotators, however, the harder it is to achieve an IAA of .80 (Bayerl and Paul, 2011). What to annotate also depends, among other characteristics, on the phenomenon of interest, the text being annotated, the quality of the annotation scheme and the effectiveness of training. But even if these are excellent, there is natural variability in human judgment for a task that involves subtle distinctions about which competent coders disagree. An accurate computational model should reflect this variability (Aakhus et al., 2013).

LAW VIII - The 8th Linguistic Annotation Workshop, pages 120–128,
Dublin, Ireland, August 23-24 2014.

# Type	Statement
Target	I'm going to quit the iphone and switch to an android phone because I can no long (sic) put up with the AT&T service contract
Callout	I am going to switch too
Callout	There is no point quitting the iphone because of the service package, just jail break it and use the provider you want

Table 1: Examples of Callouts and Targets

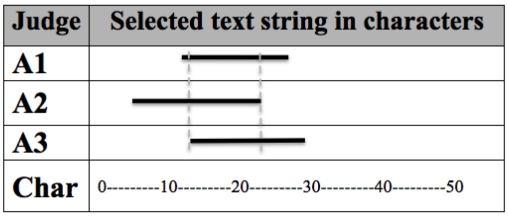

Figure 1: Cluster where 3 judges identify a core

We propose an approach for overcoming these challenges based on evidence from an annotation study of arguments in online interactions. Our scheme for argumentation is based on Pragmatic Argumentation Theory (PAT) (Van Eemeren et al., 1993; Hutchby, 2013; Maynard, 1985). PAT states that argument can arise at any point when two or more actors engage in calling out and making problematic some aspect of another actor's prior contribution for what it (could have) said or meant (Van Eemeren et al., 1993). The argumentative relationships among contributions to a discussion are indicated through links between what is targeted and how it is called out. Table 1 shows two Callouts that refer back to the same Target.

Callouts and Targets are Argumentative Discourse Units (ADUs) in the sense of Peldszus and Stede (2013a), "minimal units of analysis ... inspired ... by a ... relation-based discourse theory" (p.20). In our case the theory is PAT. Callouts are related to Targets by a relationship that we may refer to as Response, though we do not discuss the Response relationship in this paper.

The hierarchical clustering technique that we propose systematically identifies clusters of ADUs; each cluster contains a core of overlapping text that two or more judges have identified. Figure 1 shows a schematic example of a cluster with a core identified by three judges. The variation in boundaries represents the individual judges' differing intuitions; these differences reflect natural variation of human judgments about discourse units. We interpret differences in the number (or percentage) of judges that identify a core as evidence of how hard or easy a discourse unit is to recognize.

The contributions of this paper are two-fold. First, we show that methods for assessing IAA, such as the information retrieval inspired (P/R/F1) approach (Wiebe et al., 2005) and Krippendorff's α (Krippendorff, 1995; Krippendorff, 2004b), which was developed for content analysis in the social sciences, provide inconsistent results when applied to segmentations involving fuzzy boundaries and multiple coders.

In addition, these metrics do not tell us which parts of a text are easier or harder to annotate, or help choose a reliable gold standard. Our second contribution is a new method for assessing IAA using hierarchical clustering to find parts of text that are easier or harder to annotate. These clusters could serve as the basis for assessing the performance of systems that automatically identify ADUs - the system would be rewarded for identifying ADUs that are easier for people to recognize and penalized for identifying ADUs that are relatively hard for people to recognize.

2 Annotation Study of Argumentative Discourse Units: Callouts and Targets

In this section, we describe the annotation study we conducted to determine whether trained human judges can reliably identify Callouts and Targets. The main annotation task was to find Callouts and the Targets to which they are linked and unitize them, i.e., assign boundaries to each ADU. As mentioned above, these are the steps for argument mining delineated in Peldszus and Stede (2013a). The design of the study was consistent with the conditions for generating reliable annotations set forth in Krippendorff (2004a, p. 217).

We selected five blog postings from a corpus crawled from Technorati (technorati.com) between 2008-2010; the comments contain many disputes. We used the first 100 comments on each blog as our corpus, along with the original posting. We refer to each blog and the associated comments as a thread.

The complexity of the phenomenon required the perspective of multiple independent annotators, despite the known difficulty in achieving reliable IAA with more than two annotators. For our initial study, in which our goal was to obtain naturally occurring examples of Callouts and Targets and assess the challenges of reliably identifying them, we engaged five graduate students with a strong humanities background. The coding was performed with the open-source Knowtator software (Ogren, 2006). All five judges annotated all 100 comments in all five threads. While the annotation process was under way, annotators were instructed not to communicate with each other about the study.

The annotators' task was to find each instance of a Callout, determine the boundaries, link the Callout to the most recent Target and determine the boundaries of the Target. We prepared and tested a set of guidelines with definitions and examples of key concepts. The following is an adapted excerpt from the guidelines:

- **Callout:** A Callout is (a part of) a subsequent action that selects (a part of) a prior action and marks and comments on it in some way. In Table 1, Statements 2 and 3 are both Callouts, i.e., they perform the action of calling out on Statement 1. Statement 2 calls out the first part of Statement 1 dealing with switching phones. Statement 3 calls out all of Statement 1 – both what's proposed and the rationale for the disagreement.

- **Target:** A Target is a part of a prior action that has been called out by a subsequent action. Statement 1 is a Target of Statements 2 and 3. But Statements 2 and 3 link to different parts of Statement 1, as described above.

- **Response:** A link between Callout and Target that occurs when a subsequent action refers back to (is a response to) a prior action.

Annotators were instructed to mark any text segment (from words to entire comments) that satisfied the definitions above. A single text segment could be a Target and a Callout. To save effort on a difficult task, judges were asked only to annotate the most recent plausible Target. We plan to study chains of responses in future work.

Prior to the formal study, each annotator spent approximately eight hours in training, spread over about two weeks, under the supervision of a PhD student who had helped to develop the guidelines. Training materials included the guidelines and postings and comments from Technorati that were not used in the formal study. Judges were reminded that our research goal was to find naturally occurring examples of Callouts and Targets and that the research team did not know in advance what were the right answers – the subjects' job was to identify Callouts and Targets that satisfied the definitions in the guidelines. In response to the judges' questions, the guidelines were iteratively updated: definitions were reviewed, additional examples were added, and a list of FAQs was developed[1].

Table 2 shows the wide range of results among the annotators for Callouts that illustrates a problem to be addressed when assessing reliability for multiple annotators.

Averaged over all five threads, A1 identified the fewest Callouts (66.8) while A4 and A5 identified the most (107 and 109, respectively). Furthermore, the number of annotations assigned by A4 and A5 to each corpus is consistently higher than those of the other annotators, while the number of annotations A1 assigned to each thread is consistently lower than that of all of the other annotators. Although these differences could be due to issues with training, we interpret the consistent variation among coders as potential evidence of two distinct types of behavior: some judges are 'lumpers' who consider a text string as a single unit; others are 'splitters' who treat the same text string as two (or more) distinct units. The high degree of variability among coders is consistent with the observations of Peldszus and Stede

[1] The corpus, annotations and guidelines are available at <http://wp.comminfo.rutgers.edu/salts/projects/opposition/>.

Thread	A1	A2	A3	A4	A5
Android	73	99	97	118	110
Ban	46	73	66	86	83
iPad	68	86	85	109	118
Layoffs	71	83	74	109	117
Twitter	76	102	70	113	119
Avg.	66.8	88.6	78.4	107	109.4

Table 2: Callouts per annotator per thread

(2013b). These differences could be due to issues with training and individual differences among coders, but even so, the variability highlights an important challenge for calculating IAA with multiple coders and fuzzy unit boundaries.

3 Some Problems of Unitization Reliability with Existing IAA Metrics

In this section we discuss two state-of-the-art metrics frequently used for measuring IAA for discourse annotation and we show that these methods offer limited informativeness when text boundaries are fuzzy and there are multiple judges. These methods are the information retrieval inspired precision-recall (P/R/F1) metrics used in Wiebe and her collaborators' important work on sentiment analysis (Wiebe et al., 2005; Somasundaran et al., 2008) and Krippendorff's α, a variant of the α family of IAA coefficients specifically designed to handle fuzzy boundaries and multiple annotators (Krippendorff, 1995; Krippendorff, 2004b). Krippendorff's α determines IAA based on observed disagreement relative to expected agreement and calculates differences in annotators' judgments. Although it is possible to use number of words or even clauses to measure IAA, we use length in characters both for consistency with Wiebe's approach and because Krippendorff (2004b, pp.790-791) recommends using "... the smallest distinguishable length, for example the characters in text..." to measure IAA. We next show the results of using P/R/F and Krippendorff's α to measure IAA for our annotation study and provide examples of some challenges that need to be addressed.

3.1 Precision, Recall and F measures

Implementing P/R/F1 requires a gold standard annotation against which the other annotations can be compared. P/R/F1 is calculated here, following (Wiebe et al., 2005), as follows: the units selected by one annotator are taken as the gold standard and the remaining annotators are calculated against the selected gold standard. To determine whether annotators selected the same text span, two different types of matches were considered, as in Somasundaran et al. (2008): exact matches and overlap matches (variation of their lenient match):

- **Exact Matches (EM):** Text spans that vary at the start or end point by five characters or less are considered an exact match. This minor relaxation of exact matching (Somasundaran et al., 2008) compensates for minor inconsistencies such as whether a judge included a sentence ending punctuation mark in the unit.

- **Overlap Matches (OM):** Any overlap between text spans of more than 10% of the total number of characters is considered a match. OM is weaker than EM but still an indicator of shared judgments by annotators.

Tables 3 and 5 and Tables 4 and 6 show the P/R/F1-based IAA using EM and OM respectively. The results are averaged across all five threads. Besides average P/R/F1 we also show Max F1 and Min F1, which represent the maximum and minimum F1 relative to a particular annotator used as gold standard.

These tables show that the results vary greatly. Among the reasons for the variation are the following:

- Results are sensitive to which annotator is selected as the gold standard. In Table 4, pairing A4 with the judge who agrees maximally produces an F measure of 90.2 while pairing A4 with the annotator who agrees minimally produces an F measure of 73.3. In Tables 3 and 4, if we select A4 as the gold standard we get the most variation; selecting A3 produces the least.

Ann	Avg P	Avg R	Avg F1	MaxF1	Min F1
A1	40.7	57.7	47.8	60	36.7
A2	51.7	51.2	51.4	58.3	43
A3	54.2	57.8	55.9	61.4	47.9
A4	59.7	49.1	53.9	61.4	47.3
A5	55	45.6	49.9	58.3	36.7

Table 3: Callouts: EM P/R/F1 over 5 threads

Ann	Avg P	Avg R	Avg F1	MaxF1	Min F1
A1	67.4	95.7	79.1	86.8	73.3
A2	85	83.7	84.3	88.7	76.1
A3	82.7	88	85.2	88.7	80.9
A4	92.7	76.8	84	90.2	73.3
A5	91.4	75.1	82.4	89.6	74

Table 4: Callouts: OM P/R/F1 over 5 threads

Ann	Avg P	Avg R	Avg F1	MaxF1	Min F1
A1	24.1	34.6	28.4	34.5	18.7
A2	26.9	24.7	25.7	37.6	18.7
A3	35.2	35.1	35.1	48.4	19.4
A4	37.3	34.5	35.8	50.4	22.1
A5	36.9	31.4	33.9	50.4	19.9

Table 5: Targets: EM P/R/F1 over 5 threads

Ann	Avg P	Avg R	Avg F1	MaxF1	Min F1
A1	60.1	86.5	70.9	76.1	64.2
A2	74.5	69.4	71.9	79.6	62.9
A3	75.9	74.5	75.1	80.1	67.7
A4	78.1	71.5	74.6	84.2	64
A5	83.8	70.3	76.4	83.8	67.2

Table 6: Targets: OM P/R/F1 over 5 threads

- The type of matching matters. As expected, OM, which is less strict than EM, produces substantially higher F1 scores both for Callouts (Tables 3 and 4) and Targets (Tables 5 and 6).
- Different phenomena are associated with different levels of difficulty of annotation. The F1 scores for Targets are considerably lower than the F1 scores for Callouts. We suspect that Callouts are easier to recognize since they are often introduced with standard expressions that signal agreement or disagreement such as ‘yes’, ‘no’, ‘I agree’, or ‘I disagree’. Targets, on the other hand, generally lack such distinguishing lexical features.

We also observe differences across threads. For example, the Ban thread seems harder to annotate than the other threads. Figure 2 and 3 show IAA results for OM for Callout and Target annotations for annotators A1 and A5 respectively, across the five threads. We chose A1 and A5 because in general A1 annotated the fewest Callouts and A5 annotated the most Callouts in the corpus. These figures show different annotator behavior. For instance, for both Callout and Target annotations, A1 has higher average R than P, while A5 has higher P but lower R. Figures 2 and 3 hint that the Ban thread is harder to annotate than the others.

The examples in this section show two downsides to the P/R/F1 metric. First, the scores do not reflect the extent to which two annotations match. This is crucial information for fuzzy boundary matching, because the agreement between two annotations can be over only a few characters or over the full length of the selected text. Second, the variation across multiple judges demonstrates the disadvantage of arbitrary selection of a gold standard set of annotations against which to measure IAA.

3.2 Krippendorff’s α

Krippendorff’s α calculates IAA based on the observed and expected disagreement between annotators. We use the version of Kripendorff’s α discussed in Krippendorff (2004b) which takes into account multiple annotators and fuzzy boundaries. Detailed proof and an explanation of the calculation can be found in (Krippendorff, 2004b; Krippendorff, 1995).

Thread	F1	Krippendorff’s α
Android	87.8	0.64
Ban	85.3	0.75
iPad	86.0	0.73
Layoffs	87.5	0.87
Twitter	88.5	0.82

Table 7: F1 and α for all 5 threads

Thread Rank by IAA (Descending)	
F1	K’s α
Twitter	Layoffs
Android	Twitter
Layoffs	Ban
iPad	iPad
Ban	Android

Table 8: Threads ranked by IAA in descending order

Comparison of α and P/R/F1 metrics shows that they generate inconsistent results that are difficult to interpret. For example, in Table 7, the F1 measure for Callouts indicates lower agreement on the Ban thread in comparison to Android while α suggests higher agreement on the Ban subcorpus relative to the

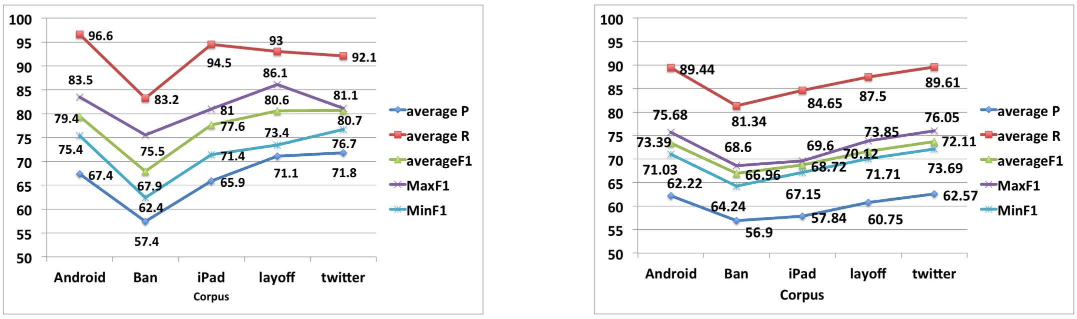

Figure 2: IAA metrics per thread when A1 is gold standard (Left: Callout. Right: Target.)

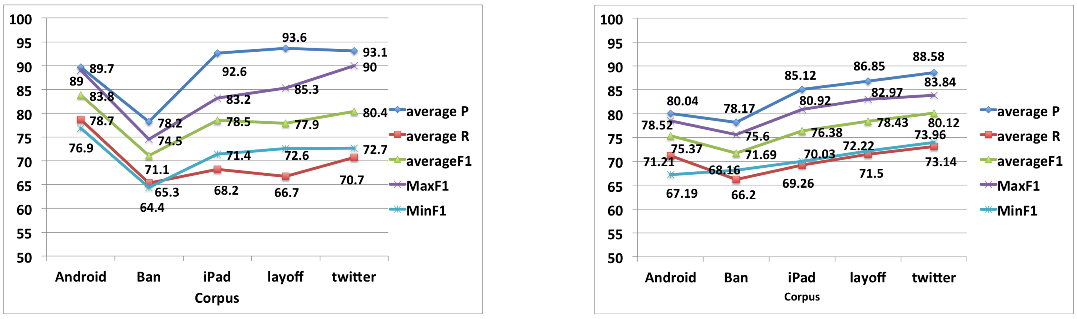

Figure 3: IAA metrics per thread when A5 is gold standard (Left: Callout. Right: Target.)

Android subcorpus. The inconsistencies are also apparent in Table 8, which ranks threads in descending order of IAA. For example, the Android corpus receives the highest IAA using F1 but the lowest using α.

We do not show the results for Krippendorff's α for Targets for the following reason. Relevant units from a continuous text string are assigned to categories by individual annotators. But identification of Targets is dependent on (temporally secondary to) identification of Callouts. In multiple instances we observe that an annotator links multiple Callouts to two or more overlapping Targets. Depending on the Callout, the same unit (i.e., text segment) can represent an annotation (a Target) or a gap between two Targets. Computation of α is based on the overlapping characters of the annotations and the gaps between the annotations. Naturally, if a single text string is assigned different labels (i.e. annotation or a gap between annotations) in different annotations, α does not produce meaningful results. The inapplicability of Krippendorff's α to Targets is a significant limitation for its use in discourse annotation (To save space we only show results for Callouts in subsequent tables.)

The examples in Section 3 show a fundamental limitation of both P/R/F1 and Krippendorff's α: They do not pinpoint the location in a document where the extent of variation can be observed. This limits the usefulness of these measures for studying the discourse phenomenon of interest and for analyzing the impact of factors such as text difficulty, corpus and judges on IAA. The impact of these factors on IAA also makes it hard to pick gold standard examples on a principled basis.

4 Hierarchical Clustering of Discourse Units

In this section we introduce a clustering approach that aggregates overlapping annotations, thereby making it possible to quantify agreement among annotators within a cluster. Then we show examples of clusters from our annotation study in which the extent of annotator support for a core reflects how hard or easy an ADU is for human judges to identify. The hierarchical clustering technique (Hastie et al., 2009) assumes that overlapping annotations by two or more judges constitutes evidence of the approximate location of an instance of the phenomenon of interest. In our case, this is the annotation of ADUs that contain overlapping text. Each ADU starts in its own cluster. The start and end points of each ADU are utilized to identify overlapping characters in pairs of ADUs. Then, using a bottom-up clustering

# Annots	Text selected
A1, A2, A3, A4, A5	I remember Apple telling people give the UI and the keyboard a month and you'll get used to it. Plus all the commercials showing the interface. So, no, you didn't just pick up the iPhone and know how to use it. It was pounded into to you.

Table 9: A cluster in which all five judges agreement on the boundaries of the ADU

# Annots	Text selected
A1	*I'm going to agree that my experience required a bit of getting used to . . .*
A2, A3, A4	*I'm going to agree that my experience required a bit of getting used to . . .* I had arrived to the newly minted 2G Gmail and browsing
A5	*I'm going to agree that my experience required a bit of getting used to . . .* I had arrived to the newly minted 2G Gmail and browsing. Great browser on the iPhone but . . . Opera Mini can work wonders

Table 10: A cluster in which all 5 annotators agree on the core but disagree on the closing boundary of the ADU

technique, pairs of clusters (e.g. pairs of Callout ADUs) with overlapping text strings are merged as they move up in the hierarchy. An ADU that does not overlap with ADUs identified by any other judge will remain in its own cluster.

Aggregating overlapping annotations makes it possible to quantify agreement among the annotators within a cluster. Table 9 shows an example of a cluster that contains five annotations; all five annotators assign identical unit boundaries, which means that there is a single core, with no variation in the extent of the ADU. Table 9 thus shows an optimal case – there is complete agreement among the five annotators. We take this as strong evidence that the text string in Table 9 is an instance of a Callout that is relatively easy to identify.

But of course, natural language does not make optimal annotation easy (even if coders were perfect). Table 10 shows a cluster in which all five annotators agree on the core (shown in italics) but do not agree about the boundaries of the ADU. A1 picked the shortest text segment. A2, A3 and A4 picked the same text segment as A1 but they also included the rest of the sentence, up to the word 'browsing'. In A5's judgment, the ADU is still longer - it also includes the sentence 'Great browser . . . work wonders.' Although not as clear-cut as the examples in Table 9, the fact that in Table 10 all annotators chose overlapping text is evidence that the core has special status in the context of in an annotation task where it is known that even expert annotators disagree about borders. Examples like those in Table 10 can be used to study the reasons for variation in the judges' assignment of boundaries. Besides ease of recognition of an ADU and differing human intuitions, the instructions in the guidelines or characteristics of the Callouts may be also having an effect.

Table 11 shows a more complex annotation pattern in a cluster. Annotators A1 and A2 agree on the boundaries of the ADU, but their annotation does not overlap with A4 at all. A3's boundaries subsume all other annotations. But because A4's boundaries do not overlap with those of A1 and A2, technically this cluster has no core (a text segment included in all ADUs in a cluster). 5% or less of the clusters have this problem. To handle the absence of a core in this type of cluster, we split the clusters that fit this pattern into multiple 'overlapping' clusters, that is, we put A1, A2, and A3 into one cluster and we put A3 and A4 into another cluster. Using this splitting technique, we get two cores, each selected by two judges: i) "actually the only . . . app's developer" from the cluster containing A1, A2, and A3 (shown in italics) and ii) "I think it hilarious . . . another device" from the cluster containing A3 and A4 (shown in bold). The disagreement of the judges in identifying the Callout suggests that judges have quite different judgments about boundaries of the Callouts.

Table 12 and 13 respectively show the number of clusters with overlapping annotations for Callouts for each thread before and after splitting. The splitting process has only a small impact on results. The number of clusters with five and four annotators shows that in each corpus there are Callouts that are evidently easier to identify. On the other hand, clusters selected by only two or three judges are harder to

# Annots	Text selected
A1, A2	*Actually the only one responsible for the YouTube and Twitter multitasking is the app's developer*
A3	*Actually the . . . app's developer.* The Facebook app allows you to watch videos posted by . . . **I think it hilarious that people complain about features that arent even available on another device**
A4	**I think it hilarious that people complain about features that arent even available on another device**

Table 11: A cluster with 2 cores, each selected by 2 judges

identify. The clusters containing a text string picked by only one annotator are hardest to identify. This may be an indication that this text string is not a good example of a Callout, though it also could be an indication that the judge is particularly good at recognizing subtly expressed Callouts. The clustering technique thus scaffolds deeper examination of annotation behavior and annotation/concept refinement. Table 13 also shows that overall, the number of clusters with five or four annotators is well over 50% for each thread except Ban, even when we exclude the clusters with an ADU identified by only one judge. This is another hint that the IAA in this thread should be much lower than in the other threads. (See also Figures 2 and 3).

Thread	# of Clusters	Annots in each cluster				
		5	4	3	2	1
Android	91	52	16	11	7	5
Ban	89	25	18	12	20	14
Ipad	88	41	17	7	13	10
Layoffs	86	41	18	11	6	10
Twitter	84	44	17	14	4	5

Table 12: Callouts: Clusters before splitting process

Thread	# of Clusters	Annots in each cluster				
		5	4	3	2	1
Android	93	51	15	14	8	5
Ban	91	25	19	12	21	14
iPad	89	41	16	9	13	10
Layoffs	89	40	17	14	8	10
Twitter	87	43	15	20	4	5

Table 13: Callouts: Clusters after splitting process

The clusters with cores supported by four or five annotators show strong annotator agreement and are very strong candidates for a gold standard, regardless of the IAA for the entire thread. Clusters with an ADU selected by only one annotator are presumably harder to annotate and are more likely than other clusters not to be actual instances of the ADU. This information can be used to assess the output of systems that automatically identify discourse units. For example a system could be penalized more for missing to identifying ADUs on which all five annotators agree on the boundaries, as in Table 9; the penalty would be decreased for not identifying ADUs on which fewer annotators agree. Qualitative analysis may help discover the reason for the variation in strength of clusters, thereby supporting our ability to interpret IAA and to create accurate computational models of human judgments about discourse units. As a related research, PAT and the clustering technique discussed in this paper allow the development of a finer-grained annotation scheme to analyze the type of links between Target-Callout (e.g., Agree/Disagree/Other), and the nature of Callouts (e.g., Stance/Rationale) (Ghosh et al., 2014).

5 Conclusion and Future Work

Reliability of annotation studies is important both as part of the demonstration of the validity of the phenomena being studied and also to support accurate computational modeling of discourse phenomena. The nature of ADUs, with their fuzzy boundaries, makes it hard to achieve IAA of .80 or higher. Furthermore, the use of a single figure for IAA is a little like relying on an average to convey the range of variation of a set of numbers. The contributions of this paper are i) to provide concrete examples of the difficulties of using state of the art metrics like P/R/F1 and Krippendorff's α to assess IAA for ADUs and ii) to open up a new approach to studying IAA that can help us understand how factors like coder variability and text difficulty affect IAA. Our approach supports reliable identification of discourse units independent of the overall IAA of the document.

References

Mark Aakhus, Smaranda Muresan, and Nina Wacholder. 2013. Integrating natural language processing and pragmatic argumentation theories for argumentation support. pages 1–12.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, Maryland, June. Association for Computational Linguistics.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. 2009. *The elements of statistical learning*, volume 2. Springer.

Ian Hutchby. 2013. *Confrontation talk: Arguments, asymmetries, and power on talk radio*. Routledge.

Klaus Krippendorff. 1995. On the reliability of unitizing continuous data. *Sociological Methodology*, pages 47–76.

Klaus Krippendorff. 2004a. *Content analysis: An introduction to its methodology*. Sage.

Klaus Krippendorff. 2004b. Measuring the reliability of qualitative text analysis data. *Quality & quantity*, 38:787–800.

Douglas W Maynard. 1985. How children start arguments. *Language in society*, 14(01):1–29.

Philip V Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations*, pages 273–275. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2013a. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2013b. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 196–204.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2008. Discourse level opinion relations: An annotation study. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 129–137. Association for Computational Linguistics.

Frans H Van Eemeren, Rob Grootendorst, Sally Jackson, and Scott Jacobs. 1993. *Reconstructing argumentative discourse.* University of Alabama Press.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.