Argument Quality in the Eye of the Annotator: Understanding the Challenges of Annotation Through Eye Tracking

Rositsa V Ivanova[1 [0000−1111−2222−3333]] and Kenan Bektaş[1,2 [0003−2937−0542]]

[1] University of St. Gallen, Switzerland
[2] HOCH Health Ostschweiz, Switzerland
{rositsa.ivanova,kenan.bektas}@unisg.ch

Abstract. In Natural Language Processing, the quality of data annotation directly affects the performance of tools. In the subfield of Computational Argumentation, the annotation of argumentative text is a challenging task due to its subjective nature and high cognitive demand for both experts and novices. In this paper, we provide a systematic overview of these challenges and propose the use of eye tracking as a method that allows us to better understand the experience of annotators. We report on a randomised crossover experiment with 40 participants combining eye movement, self-reported workload using NASA-TLX, and task-specific questionnaires. We demonstrate a rapid decrease in attention throughout the annotation task and during the familiarisation of the annotators with the guidelines. We show that annotators perceive longer guidelines as more mentally demanding, yet more helpful and clear than shorter ones. Our findings show the feasibility of eye tracking for the understanding of the annotation challenges.

Keywords: Computational Argumentation, Natural Language Processing, Eye Tracking, Text Annotation

1 Introduction

In a time when information is available en mass, learning to analyse argumentative texts and to write good arguments is an essential skill that many curricula include. Yet, providing scholars with personalised feedback is very time consuming and still a challenge for most teachers [31]. The field of Computational Argument Quality Assessment (CA) offers support by aiming to automatically assess argumentative texts across various quality dimensions [11, 23, 29]. To achieve this, models need to be created with the help of high quality *gold standard* datasets, which are typically created (i.e., annotated) by experts, via crowdsourcing or using other methods [13, 23].

Despite the already high number of existing datasets in the field, the continuous adaptations and improvements of theoretical foundation of the field (and thus the used quality dimensions) is accompanied by a continuous need for new gold standards. Only recently, Romberg et al. [23] surveyed the datasets in the

field and underlined the importance of various perspectives in the highly subjective task of quality assessment. They suggest that future work shift its focus from a single correct response to multiple correct responses (i.e., the perspectivist turn). Simultaneously, the annotation task is viewed as not only subjective but also demanding for both expert and non-expert annotators [6, 13, 14, 29]. Our work focuses on this issue. Through the use of eye tracking - a method that has not previously been applied to the field of CA - we aim to gain objective insight into aspects of the annotators' behavior, which may have been invisible using other techniques. We formulate our research questions as follows:

RQ1: Does the focus of annotators decrease significantly as they read through annotation guidelines?

RQ2: Do annotators perceive shorter or longer annotation guidelines as more cognitively demanding?

RQ3: At which point throughout the annotation process does the focus of annotators decrease significantly?

The contribution of this work is twofold. First, we highlight prior findings about the annotation process with a focus on CA. Second, we present our experimental setup, which is followed by its results.

2 Related Work

Computational Argumentation. The field of CA finds its roots in the coarse-granular assessment of entire essays [17–20]. From there it has adapted to shorter texts frequently scraped from online sources such as reviews and forum posts (e.g. [2, 30]). Early work focused on rather domain-agnostic quality dimensions such as organization [17], slowly shifting towards more argument-specific quality dimensions such as persuasiveness (e.g., [21, 26]) and convincingness (e.g., [8]). For a recent overview of the field of CA and a detailed description of the existing annotated datasets and quality dimensions refer to Ivanova et al. [11].

Prior work has discussed certain shortcomings of existing annotation approaches with a focus on CA. The use of *modal verbs* (e.g., can) and *non-specific quantities* (e.g., many, some) may lead to disagreements between annotators [1, 24]. In addition, annotators oftentimes interpret arguments containing *irony, sarcasm or rhetorical questions* differently from one another, sometimes even assigning opposing scores [13]. Difference can arise when annotators deem the addressed topics as *"less worthy"* or when annotators disagree about the *text structure* [1], which may also make it difficult to recognise which texts are argumentative [25]. Using crowdsourcing, annotators may have distinct *reference frames*, which dictate how they interpret interval quality scales e.g., "0-2" [6]. To address this, one may add representative examples and a training session [32].

Further, reducing the number and the granularity of dimension in an annotation taxonomy enables the annotators to comprehend the guidelines better [13]. Yet the CA field has previously been dominated by rather practical approaches that assess single dimensions in detail and thus provide fine-grained gold standards, which is necessary when personalised and specific feedback is to be given.

Simultaneously, the use of too many fine-grained quality dimensions is generally too difficult for annotators to work with and can be subjective [16]. This leads to a rather unavoidable issue, namely that (i) detailed taxonomies with fine-grained quality dimensions are necessary for the training of tools which provide detailed feedback on the quality of argumentative writing, yet (ii) the annotation process associated with the use of such fine-grained taxonomies is highly demanding for the annotators. In this work, we aim to clarify how this dilemma can be better addressed with a novel approach.

Eye Tracking. In psycholinguistics' history, researchers have used eye tracking measures (e.g., fixations, saccades, pupil dilation) to study lexical processing, syntactic parsing, sentence comprehension [12, 22]. In NLP, eye tracking has found applications in cognitive modeling as well as in the improvement of computational tasks through the gained insights. In various tasks eye movement data offers a proxy for human behavior and attention in language processing. It has been used to better understand the unique aspects of how humans cognitively process text [10] and in some cases to compare these processing steps with those of automated tools. For example, Tokunaga et al. [27] found that human annotators look at broader context windows and use relational cues more extensively than automatic systems. Further, Mitsuda et al. [15] aim to predict disagreement in the annotation by using linguistic information of the annotation and a comparison of the eye behaviors of annotators.

3 Experiment

Participants The experiment was carried out with 40 participants, who responded to an open call and were BSc, MSc or PhD students and proficient in English. We selected non-experts for the task, as this target group is typically more affordable for researchers and easier to recruit. The participants' genders were reported as follows: 21 male, 18 female, and 1 other. The ages spanned from 19 to 44 years old with the median at 23.5.

Annotation guidelines. We use two sets of guidelines. The first guideline is *minimal* and contains a definition of the quality dimension *clarity* [29]. The second guideline is *extended* and in addition to the minimal guideline contains a brief introduction to *sufficiency* and three examples for argumentative texts - two negative [24] and one positive, which was added by us to account for findings in prior work [32, 6]. Note that examples are not always available within the annotation guidelines used in CA. Further, prior work (e.g., [7, 13]) has pointed out that in order to be able to offer detailed feedback to learners, it is essential to create fine-grained annotation that depicts nuances in the argumentation quality. Therefore we adopt an interval scale from 1 to 3 [5, 29].

Apparatus and Data. Eye-tracking data was captured with a Tobii Pro Fusion 120Hz and subsequently recorded and analyzed using iMotions software. We selected argumentative texts of medium length from the UKP-ConvArg1 Corpus [9] and selected a topic that all participants would be familiar with (i.e.,"TV vs. books") to account for the importance of familiarity with the topic.

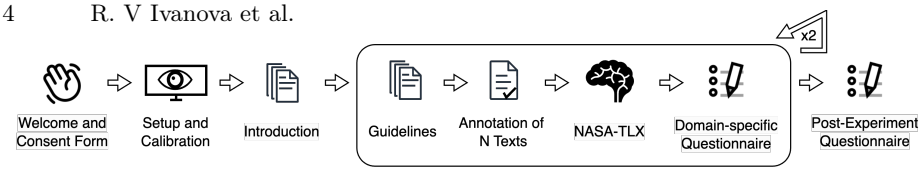


Fig. 1: Overview of the steps of the experiment.

Procedure. Our experiment follows a randomised crossover design, where participants are exposed to all study steps, but in a different order. Through this we aim to avoid a familiarity bias with the quality dimensions - a potential carryover effect, introduced when annotators are affected by prior experiences and knowledge rather than strictly following the only the current guidelines. To account for the familiarity with the annotation process, we rotate the order of appearance of the two guidelines. Through this we create two versions, which are then randomly assigned to and equally distributed across the participants.

The steps of the experiment are depicted in Figure 1. After a welcoming and a calibration, the participants are shown the study instructions. In two iterations, they are presented with one guideline at the time (i.e., minimal or extended), then are asked to annotate 10 arguments, fill out a NASA-TLX questionnaire and answer domain-specific questions. Lastly, after the second iteration, the participants are asked to fill out a post-experiment questionnaire, comparing the two guideline styles and collecting demographic data.

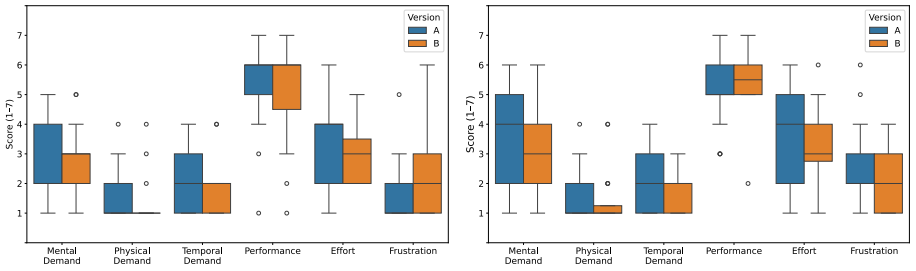
4 Results

We excluded 16 participants from the eye tracking analysis due to the low quality of their gaze recording, yet we considered their responses in the surveys of the experiment. Coincidentally, this left us with 12 participants per guideline.

Reading Behavior. To better understand how annotators read the guidelines we first looked at *the time each participant required to read* them and compared the durations to the average expected reading speed for non-fiction English texts [3] as a baseline. The majority of the participants spent less time than the average reader on the intro, but spent more time on the guidelines pages. Here, eye tracking allows us to determine the time spent on individual lines of the text with the help of *areas of interest*. We calculated the *mean slope in dwell times* for the extended guidelines to be -0.094 ($p<0.001$), which means that participants sped up and spent less time as they progressed through the lines (i.e., dwell time decreased per line). We normalised the durations per line by the word count of each line. The same tendency was found for the intro text ($p<0.001$ and a mean slope -2.834). These results align with prior findings that readers pay more attention to the beginning of a paragraph [4]. In addition, we discover noticeable peaks when annotators reach the beginning of each of the three examples in the guideline text, indicating an increased attention. The three examples display a clear significance in the decrease in attention as the participants pass the middle of a longer paragraph.

Answer to RQ1: The observed results show that the focus of annotators decreases significantly throughout the process of reading a single text / webpage.

However, our findings indicate that the attention may be better maintained when annotators are presented with shorter paragraphs (i.e., more interruptions in the text) rather than lengthy ones. Annotating texts is a cognitively demanding and thus tiring task. We suggest that future work experiments with alternative and more interactive methods such as using gamification or including a training step prior to the actual annotation, which offers feedback to annotators on their accuracy. We acknowledge that the later may be a conflicting suggestion for argument quality dimension, which are inherently subjective (e.g., persuasiveness), yet could be beneficial for the more objective ones (e.g., clarity of writing).



(a) Minimal guideline (b) Extended guideline

Fig. 2: NASA-TLX score distributions for the minimal and the extended guideline (per Version)

Cognitive Demand The complexity of the extended guideline is also reflected by the self-assessment of the participants using *NASA-TLX*. The boxplots in Figures 2a and 2b depict the responses of the participants per question and experiment version (i.e., A or B). The selected scores do not differ significantly between the two used versions, meaning that they were not dictated by the order of familiarisation with the two guidelines (i.e., minimal and extended). We found a significant difference between the scores assigned for the perceived *mental demand* with the extended guideline being significantly more mentally demanding. When asked 22 participants (or 55%) preferred the extended guideline, indicating no significant difference between the two guidelines. 50% preferred the minimal and 22.73% preferred the extended guideline due to the addressed quality dimension. The minimal guideline was preferred, because it is shorter, quicker and easier to understand, yet disliked as it felt too generic, vague, leaving space for individual interpretation, requiring annotators to trust their own judgment and thus feeling more subjective. On the contrary, the extended guideline was perceived as detailed, with clear explanation of the quality dimension, concrete and various examples, and thus offering a sufficient base for the assessment task. Such positive comments about the extended guideline were also made by participants who preferred the minimal guideline due to other factors. Accordingly, downsides of the extended guideline are that it requires more reading and thinking.

Answer to RQ2: The results from the NASA-TLX questionnaire clearly indicate that the extended guideline is perceived by the participants as more mentally

demanding. However, the comments left by the majority of the participants also show that they acknowledge the complexity of the task and consequently feel more prepared for it after familiarising themselves with the detailed guideline and its examples. Considering the low availability of experts in the CA field [6], we need to ensure that the non-experts are familiarised with the essential concepts, which makes the use of long and complex guidelines inevitable. Our experiment demonstrates that the use of multiple examples that cover all labels (i.e., high, medium, and low quality) is perceived as helpful. To reduce the cognitive demand of the task, future work can experiment with different annotation setups, which split long text into smaller chunks that are easier to process.

Focus Maintainability Next, we assess how long annotators can work on a task before their *attention begins to significantly drop*. We use the reading duration per page (normalised by the word count) as a proxy for the attention or effort of participants [28]. We observe that the duration begins to significantly decrease earlier for the group which first assesses using the minimal guideline - at the 5th text. In both guideline versions, the eyetracking shows us that the dwell time on the elements drops significantly after participants move from the first text to the second. It then only increase when the stance changes and then drops again afterwards. Due to this behavior of ignoring parts of webpages that appear to not be changing, we recommend that minor yet essential changes throughout the annotation process are displayed in a manner that cannot be overseen.

Answer to RQ3: In our experiment setting we observe a significant decrease in the time spent per page as annotators progress in the task. This change can be observed on average after 9 texts. Our findings indicate that participants ignore repeating elements (e.g., the definition of the stance) very early on, sometimes beginning already upon the second appearance of the elements. Similarly, a repetitive task such as the assessment of one text after the other, seems to be leading to a decrease in the time spent on a text and thus the maintained focus on the task. To maintain the focus of annotators we recommend more dynamic and interactive annotation settings. However, it is essential to consider that this may also result in a cognitive overload, if applied with a high frequency. Future work should consider how annotators handle various types of external impulses and for how long they would be able to maintain their focus prior to exhaustion.

5 Conclusion

This work provides an insight into the potential benefits of the use of eye tracking as a gateway to the mind and as means to explore subjectivity in the assessment of arguments. Our findings indicate that the attention of annotators rapidly and significantly decreases, if not maintained or demanded by the use of textual features such as new shorter paragraphs or changes in content (e.g., the appearance of an example after a theoretical explanation). Our experiment further provides a positive insight on the use of detailed guidelines, which on the one side are perceived as mentally more demanding, on the other side are acknowledged by annotators to be more helpful for the task at hand.

Acknowledgments. This study was funded by the Behavioral Lab of the University of St. Gallen.
Supplemental Material. The experiment data as well as further details can be found at `rivanova.org/its2026`

References

1. Alhamzeh, A.: Financial argument quality assessment in earnings conference calls. In: International Conference on Database and Expert Systems Applications. pp. 65–81. Springer (2023), https://doi.org/10.1007/978-3-031-39821-6_5
2. Braunstain, L., Kdoiand, O., Carmel, D., Szpektor, I., Shtok, A.: Supporting human answers for advice-seeking questions in cqa sites. In: Advances in Information Retrieval: 38th European Conference on IR Research. pp. 129–141. Springer (2016), https://doi.org/10.1007/978-3-319-30671-1_10
3. Brysbaert, M.: How many words do we read per minute? a review and meta-analysis of reading rate. Journal of memory and language **109**, 104047 (2019), https://doi.org/10.1016/j.jml.2019.104047
4. Duggan, G.B., Payne, S.J.: Skim reading by satisficing: evidence from eye tracking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1141–1150 (2011), https://doi.org/10.1145/1978942.1979114
5. Dumani, L., Schenkel, R.: Quality-aware ranking of arguments. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 335–344 (2020), https://doi.org/10.1145/3340531.3411960
6. Gienapp, L., Stein, B., Hagen, M., Potthast, M.: Efficient pairwise annotation of argument quality. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5772–5781 (2020), https://aclanthology.org/2020.acl-main.511/
7. Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., Slonim, N.: A large-scale dataset for argument quality ranking: Construction and analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7805–7813 (2020), https://doi.org/10.1609/aaai.v34i05.6285
8. Habernal, I., Gurevych, I.: 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. pp. 1214–1223 (2016), https://aclanthology.org/D16-1129/
9. Habernal, I., Gurevych, I.: 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). pp. 1589–1599. Association for Computational Linguistics (2016), https://doi.org/10.18653/v1/P16-1150
10. Iida, R., Mitsuda, K., Tokunaga, T.: Investigation of annotator's behaviour using eye-tracking data. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. pp. 214–222 (2013), https://aclanthology.org/W13-2326/
11. Ivanova, R.V., Huber, T., Niklaus, C.: Let's discuss! quality dimensions and annotated datasets for computational argument quality assessment. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 20749–20779 (2024), https://aclanthology.org/2024.emnlp-main.1155/

12. Just, M.A., Carpenter, P.A.: A theory of reading: From eye fixations to comprehension. Psychological Review **87**(4), 329–354 (1980), https://doi.org/10.1037/0033-295X.87.4.329
13. Lauscher, A., Ng, L., Napoles, C., Tetreault, J.: Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics (2020), https://aclanthology.org/2020.coling-main.402/
14. Mirzakhmedova, N., Gohsen, M., Chang, C.H., Stein, B.: Are Large Language Models Reliable Argument Quality Annotators? In: 1st International Conference on Recent Advances in Robust Argumentation Machines (2024), https://doi.org/10.1007/978-3-031-63536-6_8
15. Mitsuda, K., Iida, R., Tokunaga, T.: Detecting missing annotation disagreement using eye gaze information. In: Proceedings of the 11th Workshop on Asian Language Resources. pp. 19–26 (2013), https://aclanthology.org/W13-4303/
16. Ng, L., Lauscher, A., Tetreault, J., Napoles, C.: Creating a domain-diverse corpus for theory-based argument quality assessment. In: Proceedings of the 7th Workshop on Argument Mining (2020), https://aclanthology.org/2020.argmining-1.13/
17. Persing, I., Davis, A., Ng, V.: Modeling organization in student essays. In: Proceedings of the 2010 Conference on Empirical Methods in Natural lLnguage Processing. pp. 229–239 (2010), https://aclanthology.org/D10-1023/
18. Persing, I., Ng, V.: Modeling thesis clarity in student essays. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 260–269 (2013), https://aclanthology.org/P13-1026/
19. Persing, I., Ng, V.: Modeling prompt adherence in student essays. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1534–1543 (2014), https://aclanthology.org/P14-1144/
20. Persing, I., Ng, V.: 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). pp. 543–552 (2015), https://aclanthology.org/P15-1053/
21. Persing, I., Ng, V.: Why can't you convince me? modeling weaknesses in unpersuasive arguments. In: Proceedings of the International Joint Conferences on Artificial Intelligence. pp. 4082–4088 (2017), https://doi.org/10.24963/ijcai.2017/570
22. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological Bulletin **124**(3), 372–422 (1998), https://doi.org/10.1037/0033-2909.124.3.372
23. Romberg, J., Maurer, M., Wachsmuth, H., Lapesa, G.: Towards a perspectivist turn in argument quality assessment. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (2025), https://aclanthology.org/2025.naacl-long.382/
24. Stab, C., Gurevych, I.: 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. pp. 980–990 (2017), https://aclanthology.org/E17-1092/
25. Swanson, R., Ecker, B., Walker, M.: Argument mining: Extracting arguments from online dialogue. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 217–226 (2015), https://aclanthology.org/W15-4631/

26. Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., Lee, L.: Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In: Proceedings of the 25th International Conference on World Wide Web. pp. 613–624 (2016), https://doi.org/10.1145/2872427.2883081
27. Tokunaga, T., Nishikawa, H., Iwakura, T.: An eye-tracking study of named entity annotation. In: Proceedings of the International Conference Recent Advances in Natural Language Processing. pp. 758–764 (2017), https://doi.org/10.26615/978-954-452-049-6_097
28. Tomanek, K., Hahn, U., Lohmann, S., Ziegler, J.: A cognitive cost model of annotations based on eye-tracking data. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 1158–1167 (2010), https://aclanthology.org/P10-1118/
29. Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: 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. pp. 176–187 (2017), https://aclanthology.org/E17-1017/
30. Wachsmuth, H., Trenkmann, M., Stein, B., Engels, G., Palakarska, T.: A review corpus for argumentation analysis. In: Computational Linguistics and Intelligent Text Processing: 15th International Conference, Proceedings, Part II 15. pp. 115–127. Springer (2014), https://doi.org/10.1007/978-3-642-54903-8_10
31. Wambsganss, T., Janson, A., Käser, T., Leimeister, J.M.: Improving students argumentation learning with adaptive self-evaluation nudging. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2) (2022), https://doi.org/10.1145/3555633
32. Wei, Z., Xia, Y., Li, C., Liu, Y., Stallbohm, Z., Li, Y., Jin, Y.: A preliminary study of disputation behavior in online debating forum. In: Proceedings of the Third Workshop on Argument Mining (2016), https://aclanthology.org/W16-2820/