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 [44]. The field of Computational Argument Quality Assessment (CA) offers support by aiming to automatically assess argumentative texts across various quality dimensions [41, 19, 32]. 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 [21, 32].

Over the past 16 years dozens of datasets have been created and used in the field of CA [19, 32]. They assess the overall quality of argumentative texts or focus on specific quality dimensions (e.g., persuasiveness). They consist of individual texts (i.e., absolute assessment) [9, 38, 13] or pairs of texts, where one text is

assessed in relation to another (i.e., relative assessment) [11, 38, 10]. Further, the granularity of text entities varies between a single premise or conclusion (i.e., argument unit), an argument containing multiple premises and/or conclusions, a combination of multiple arguments (i.e., argumentation), and a debate [41, 18].

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. [32] surveyed the datasets in the field in regards to the available metadata that describes the annotators who created the dataset and their demographics. The authors underline the importance of various perspectives in the highly subjective task of quality assessment and suggest that future work should shift its focus from a single correct response to multiple correct responses (i.e., the perspectivist turn).

This new research approach addresses one of the main issues in the creation of high-quality gold standard datasets, namely that the task is inherently subjective. Simultaneously, the task is viewed as demanding for both expert and non-expert annotators [41, 21, 10, 22]. Our work focuses on the second part of the annotation issue. Through the use of eye tracking - a method that has not previously been applied to the field of CA - we aim to gain 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 provide an overview of the field of CA (Section 2), highlighting prior findings about the annotation process. In addition, we give a brief overview of Eye tracking (Section 3). Second, we present our experimental setup in Section 4, which is followed by its results and the answers to the research questions in Section 5.

2 Computational Argumentation

The field of CA finds its roots in the coarse-granular assessment of entire essays [26–29]. From there it has adapted to shorter texts frequently scraped from online sources such as reviews and forum posts (e.g. [42, 3, 36]). Early work focused on rather domain-agnostic quality dimensions such as organization [26], slowly shifting towards more argument-specific quality dimensions such as persuasiveness (e.g., [36, 30]) and convincingness (e.g., [14]). Due to its subjective nature, the focus on argumentative texts lead to the exploration of external factors such as the prior belief of the readers [8]. Simultaneously, research groups took interest in the creation of theoretical frameworks to serve as a foundation for the

creation of annotated datasets. As such, Wachsmuth et al. [41] proposed one taxonomy, which at the time of writing is still the most widely used taxonomy in CA. Since then further research has extended the spectrum of some dimensions, by exploring particular quality dimensions in detail[3].

Annotation Approaches. Alongside the creation of new datasets and taxonomies, prior work has also discussed certain shortcomings of existing annotation approaches with a particular focus on CA. In regards to the *linguistic features* of texts modal verbs (e.g., can) and the use of non-specific quantities (e.g., many, some) may lead to disagreements between annotators [34, 1]. The former suggest that extending the guidelines to include information on how such cases are to be handled may increase the agreement between annotators. Further, annotators oftentimes interpret arguments containing irony, sarcasm or rhetorical questions differently from one another, sometimes even assigning opposing quality scores [21]. Shifting towards the *structure* of arguments, various interpretations of the argument structure are also an issue that leads to disagreement [1] and while annotators have difficulties distinguishing between argumentative and non-argumentative texts [35]. The various interpretations and their effect on subjectivity are also address by Ng et al. [25], who note that annotators may judge arguments differently when they deem their topic as "less worthy" or as potentially having "trivial consequences". Especially in the case of crowdsourcing, annotators were found to have different reference frames, which dictate how they interpret interval quality scales such as "0, 1, 2" or "low, medium, high". The authors suggest that annotation guidelines should provide detailed information on the used scales [10]. To address this, one may add time in the annotation process to introduce representative examples and provide a training session as means for the reduction of ambiguity [45].

Lauscher et al. [21] ran the most extensive study on annotation approaches in the field of CA. They found that it was necessary to simplify the 15-dimensional taxonomy [41] down to its three main dimensions in order for the annotators to be able to comprehend the guidelines. Yet the authors also point out that the field has previously been dominated by rather practical approaches that assess single dimensions such as persuasiveness [30] or overall quality [39] and that methods that provide fine-grained feedback are necessary when personalised and specific feedback is to be given. Further, annotation methods that use designs with too many fine-grained quality dimensions are too difficult for annotators to work with and can be rather subjective [25]. 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.

[3] 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. [19] and Romberg et al. [32].

3 Eye Tracking

For more than five decades, eye tracking has been widely used to study the eye movements of individuals (e.g., fixations, saccades, pupil dilation) and specifically to assess their attention, cognitive load, and proficiency in goal-oriented tasks such as reading [31, 5]. Similarly, it has been used for different tasks of the broad field of NLP.

Eye Tracking Foundations and NLP Integration. In psycholinguistics' long history of eye tracking, researchers have used measures such as fixation and saccades to study lexical processing, syntactic parsing, sentence comprehension [31, 20]. In NLP, eye tracking has found applications in cognitive modeling as well as in the improvement of computational tasks through the gained insights. Such applications can be categorized as *sequence labeling* and *sequence classification*, including *syntactic parsing* [2], *text complexity estimation* [12], *named entity recognition* [16], *sentiment and sarcasm detection* [23], and *neural attention regularization* [33]. In these tasks eye movement data offers a proxy for human behavior and attention in language processing.

Annotation with Eye Tracking. In the context of NLP, eye tracking has been used to better understand the unique aspects of how humans cognitively process text [17] and in some cases to compare these processing steps with those of automated tools. For example, Tokunaga et al. [37] found that human annotators look at broader context windows and use relational cues more extensively than automatic systems. Moreover, the creation of datasets for the training of such automated tools has also previously benefited from eye tracking. Here, the eye movements of annotators provide insight into their behavior that can be used to predict annotator disagreement. Mitsuda et al. [24] aim to better understand and predict disagreement in the annotation by using linguistic information of the annotation and a comparison of the eye behaviors of annotators (i.e., sequences of fixations).

4 Experiment

Our experiment setup is built on findings in prior work. With the help of eye tracking, we aim to provide objective insight into the challenges that annotators face when asked to assess the quality of argumentative texts. Our setup was verified by two people, who were asked to complete the experiment and report any issues.

4.1 Participants

The final experiment was then carried out with 40 participants, all of whom responded to an open call and were at the time students of a Bachelor, Master or PhD level, and proficient in English (8 of whom are native speakers). We selected non-experts for the task, as this target group is typically easier to 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. When asked about their reading behavior, 25 participants (i.e., 62.5%) responded that they read a few times a week or (almost) daily. In terms of their previous experience in argumentation, rhetoric or something related, 25 participants reported to have either participated in organised debates or taken courses in a related field.

4.2 Apparatus and Material

In the experimental setup, eye-tracking data was captured with a Tobii Pro Fusion 120Hz eye tracker and subsequently recorded and analyzed using iMotions software.

Annotation guidelines. We use two sets of guidelines for the purpose of this experiment. The first guideline is *minimal* and contains a definition of the quality dimension *clarity* [41]. 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 examples [34], one positive example was added by us to account for findings in prior work [45, 10]. Here, it is noteworthy that examples are not always available within the annotation guidelines used in CA.

Annotation scale. While relative annotation has yielded overall the highest agreement scores in the related literature, prior work (e.g., [13, 21]) has also 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 [41, 7].

Data. There seems to be no consensus in the field in regards to the optimal text length for arguments. In some cases annotators have scored longer arguments as be better [43, 34], in others a negative correlation between the text length and the assigned quality score was discovered [35, 10]. Our selection of texts is based on their classification as argumentative (i.e., at least containing premises on a defined topic) and prior findings that medium length texts are most preferred by annotators. We make use of texts published in the UKP-ConvArg1 Corpus [15]. As the original dataset was extracted from online discussions, the topic and the stance are predefined, thus not explicitly included in the argumentative text itself. Finally, we selected a topic that all participants would be familiar with (i.e.,"TV vs. books") to account for the importance of being familiar with the topic.

4.3 Procedure

Our experiment follows a randomised crossover design, where participants are exposed for all study steps, but in a different order. Through this we aim to avoid a familiarity bias[4] in regards to the definitions of the individual quality

[4] A potential carryover effect, which may be introduced when annotators are affected by their prior experiences and knowledge rather than strictly following the currently relevant guidelines only.

dimensions. To account for the familiarity with the annotation process itself, we rotate the order of appearance of the two guidelines for each participant. In addition, we change the quality dimension to be annotated (i.e. clarity and sufficiency) with each style of guidelines. Through this we create two versions (i.e., A and B), which are then randomly assigned to and equally distributed across the participants.

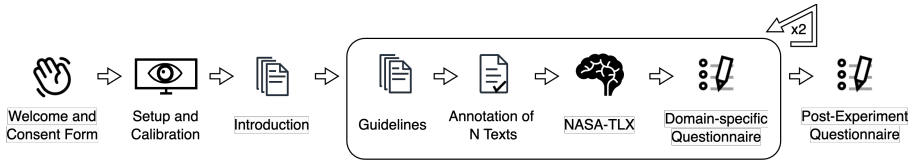


Fig. 1: Overview of the steps of the experiment.

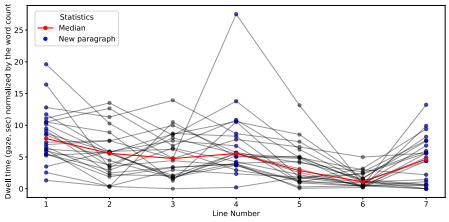
The steps of the experiment are as follows (see Figure 1). First, the principal investigator welcomes one participant at a time and introduces them to the study steps. Next, a calibration for the eye tracking setup is executed. Following, the participant reads an introduction, which briefly introduces them to the study content and their task (i.e., annotation of argumentative texts). In two iterations, the participant is presented with one guideline at the time (i.e., either the minimal or the extended), then is asked to annotated 10 arguments, to fill out a NASA-TLX questionnaire and to answer domain-specific questions, which address (i) whether the participant perceived the guidelines as sufficient for the completion of the task, (ii) whether it was difficult to follow and understand the guidelines, (iii) whether they found any information to be missing in the guidelines, which could have been helpful for the task, and (iv) whether they believe that other people would have assigned the same scores to the texts. Lastly, after the second iteration, the participant is asked to fill out a post-experiment questionnaire, which focuses on the comparison between the two annotation guideline styles and on collecting demographic data about the individual participant.

5 Results

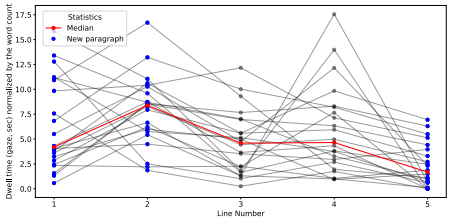
We excluded 16 participants from the eye tracking analysis due to the low quality of their gaze recording, yet we considered their responses in the survey parts of the experiment. We observed an offset in the gaze path of 10 of the remaining participants (e.g., the gaze path was located 30 pixels higher than the text itself). We manually corrected the y-axis of the gaze paths to match the first line on the "Introduction"-step of the experiment and applied the same offset across the entire gaze record of each affected participant. Coincidentally, this left us with 12 participants, who were assigned version A (i.e., minimal guidelines followed by extended guidelines) and 12 participants, who were assigned version B (i.e., extended guidelines followed by minimal guidelines).

5.1 Reading Behavior

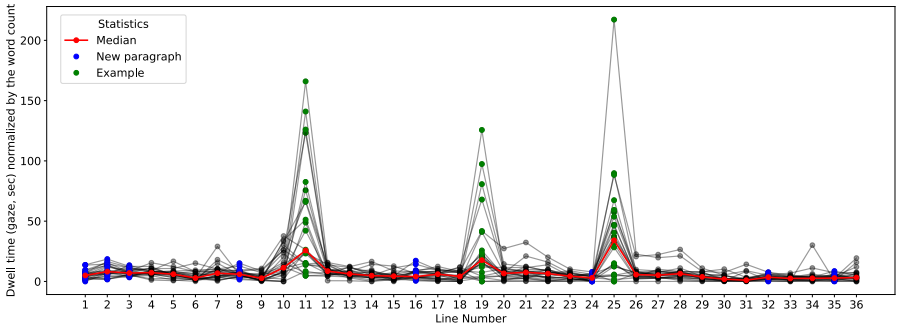
To better understand how annotators read the guidelines we first looked at *the time each participant required to read* them. We compared the durations for the *minimal guideline* (i.e., Minimal) and the *extended guideline* (i.e., Extended) to the Introduction page (i.e., *intro*) as a baseline. In addition, we considered the average expected reading speed for non-fiction English texts of 238 words per minute [4]. The majority of the participants spent less time than the average reader on the intro, but spent more time on the guidelines pages. We hypothesise that this could be due to the various nature of the texts - the guidelines being more complex and essential for the next steps of the experiment than the intro text.



(a) Introduction



(b) Minimal Guidelines



(c) Extended guidelines

Fig. 2: Dwell time spent on reading each line of the respective text per participant normalised by the number of words per line. The red line depicts the median and the blue dots depict the first lines of each new paragraph.

Here, eye tracking allows us to break down the time spent per text down to time spent on individual lines of the text with the help of *areas of interest*. Figures 2a, 2b, and 2c represent the *dwell time* in seconds per participant and line number for the intro, the minimal guideline, and the extended guideline respectively. We calculated the *mean slope in dwell times* using a one-sample t-test for the extended guidelines to be -0.094, which means that participants sped up and spent less time as they progressed through the lines (i.e., dwell

time decreased per line). Note that we normalised the durations per line by the word count of each line. This negative slope is statistically significant ($p<0.001$), therefore the trend is unlikely to be due to chance. The same tendency was found for the intro text ($p<0.001$ and a mean slope -2.834), indicating that participants spent the most time in the beginning of the text. Both of these results align with prior findings in the literature, which conclude that readers pay more attention to the beginning of a paragraph [6]. However, we did not discover any significant difference between the reading speed per line when participants were reading the minimal guidelines ($p=0.269$). The discrepancy between the intro text and the minimal text may once again be due to the difference to how important participants perceive the individual texts to be.

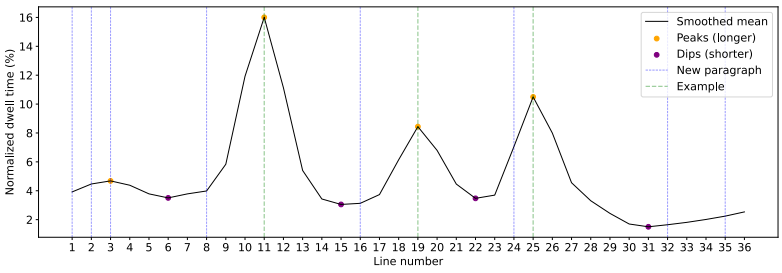


Fig. 3: The smoothed mean pattern of dwell time across text lines in the extended guidelines. The purple vertical lines indicate the first line of a new paragraph and the green ones the beginning of the paragraphs of the examples. The orange and purple dots indicate the peaks (i.e., longer dwell time) and the dips (i.e., shorter dwell time).

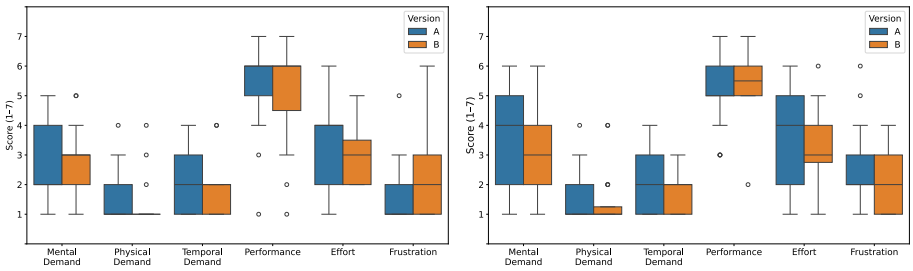
In addition to the tendency of annotators to spend less time per line as they progress through the extended guidelines, we discover noticeable peaks when annotators reach the beginning of each of the three examples in the guideline text (see orange dots in Figure 3), indicating an increased attention. Simultaneously, dips in the dwell time and thus attention can be observed as the participants pass the middle of a longer paragraph (see purple dots). We calculated the significance of the behavioral change per paragraph (longer than 2 lines). The three examples display a clear significance in the decrease in attention. In the beginning of the text (i.e., up until line 8), the participants seem to rather maintain their focus. Interestingly, the attention rather increases in the paragraph, which precede the examples (i.e., lines 8-10). This could be the result of the normalisation of the dwell duration by the word count per line, which was necessary in order to allow for comparability. In the extended guideline text, the lines 10 and 36 consist of 1 and 2 words respectively. Similarly, the first lines of each example text (i.e., lines 11, 19, and 25) contain the words “Example n:”. It is noteworthy that the observed slopes retain their relevance even if one would decide to remove the first lines of the example texts.

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).

5.2 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 4a and 4b depict the responses of the participants in respect to the individual dimensions split based on the experiment version (i.e., A or B) that they were randomly assigned. 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). When inspecting the six individual dimensions, we found a significant difference between the scores assigned for the perceived *mental demand*. Participants found the extended guideline to be significantly more mentally demanding than the minimal guideline.

In regards to the guideline preference, 15 out of the 24 (or 62.5%) participants with high quality eye tracking responded that they prefer the extended guideline. Across all 40 participants 22 (or 55%) preferred the extended guideline, indicating no significant difference between the two guidelines. The added comments in the post-experiment survey allow us to understand the reasoning behind the participants' choices.



(a) Minimal guideline (b) Extended guideline

Fig. 4: NASA-TLX score distributions for the minimal and the extended guideline (per Version)

We discovered that 9 of the 18 participants (50%) who preferred the minimal guideline do so due to the quality dimension addressed in it (i.e., clarity) and their personal familiarity with it. A similar case can be observed for the extended guideline, where 5 out of the 22 (22.73%) participants preferred the sufficiency quality dimension due to its general relevance and deeper dive into the argumentativeness of the texts. The remaining comments reveal that the minimal guideline was preferred by participants, because it is much shorter, quicker and easier to understand. Other participants disliked that it felt too generic, vague, leaving space for individual interpretation, requiring them to trust their own judgment and thus feeling more subjective.

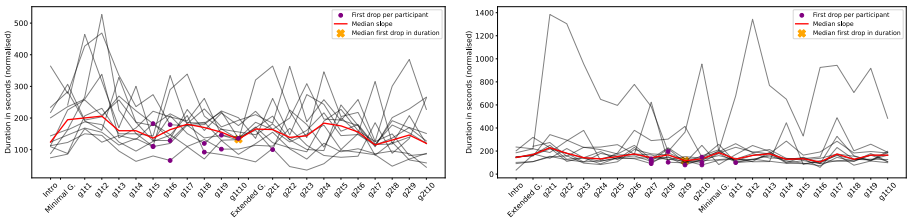
On the contrary, most of the positively perceived aspects of the extended guideline were related to the detailed and clear explanation of the quality dimension, the availability of concrete and various examples, and thus the availability of a sufficient base for the assessment task. It is noteworthy that such positive comments about the extended guideline were also made by participants who ended up selecting the minimal guideline as preferred due to other factors such as its simplicity or the personal familiarity of the quality dimension. 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 taking the time to familiarise themselves with the detailed guideline and its examples. The lack of availability of experts in the field has been reported by researchers in the past (e.g., [10]). Therefore many of the available datasets in the field have been annotated using crwodsourcing. Yet, to achieve a high-quality annotation we need to ensure that the non-experts are familiarised with the essential concepts in the field, which makes the use of long and complex guidelines almost inevitable. Our experiment demonstrates that the use of multiple examples that cover all available labels (i.e., high, medium, and low quality) is perceived as helpful by annotators. To reduce the cognitive demand of the task, future work can experiment with different annotation setups, which break down long text into smaller chunks that are easier to process. Similarly to the suggestions made in Subsection 5.1, we suggest experimenting with interactive approaches.

5.3 Focus Maintainability

Next, we assess how long annotators cam 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 [40]. Figures 5a and 5b depict the durations in seconds per word for each page that the annotators viewed for the participants in version A and B respectively. We observe that the duration begins to significantly decrease earlier for the group which first assesses using the minimal guideline - at the 5th text (i.e., g1t5) -

and slightly later - at the 7th text (i.e., g2t7) - for the other group. In both versions the participant with the latest significant decrease in attention can be found only after the second guideline was introduced. Another similarity is that the median "first drop" for both differs only by one text. Lastly, both figures depict a slight relative rise in the duration for the sixth argumentative text in a row (i.e., g1t6 and g2t6). We hypothesise that this is related to the change in stance, which happens after every 5 tests. Here once again we can make use of eye tracking for further insight.



(a) Version A

(b) Version B

Fig. 5: Time spent per text normalised by the number of words for the participants for each of the experiment versions. The abbreviations e.g. "g1t1" represent guideline (i.e., g) 1 text (i.e., t) 1. The red lines indicate the median duration as a reference. The purple dots depict the first significant reduction in the reading duration for each participant (for $p<0.05$), and the orange x-mark depicts the median first drop.

As there are certain elements that are repeating across the text assessment pages, we controlled for the change in dwell time for the elements that depict the "Stance" and the "Topic" of each text. In both versions A and B, the dwell time on these elements drops significantly after participants move from the first text to the second text. The dwell times only increase when the stance changes (i.e., at the 6th text for each guideline) and then drops again afterwards. Due to this behavior of ignoring parts of webpages that appear to not be changing, we recommend that minor changes throughout the annotation process that are essential should be displayed in a manner that cannot be overseen, such as by using a pop-up window or a separate intermediary page that explicitly states that a change is in place from the given moment onward.

Answer to RQ3: In our experiment setting we observe a significant decrease in the time spent per page / text as annotators progress in the annotation 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 for annotation settings to be created in a more dynamic and

interactive manner, which keeps the focus of annotators high with the help of external impulses. 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.

6 Conclusion

This work provides an insight into the potential benefits of the use of eye tracking for annotation in the field of CA. 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). On a bigger scale, eye tracking can demonstrate changes in attention which is essential throughout the entire annotation process. Thus, changes in the annotation setup can be helpful in re-capturing the focus of annotators. 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. By successfully answering three of our research questions, we demonstrate that the bridge between the disciplines can offer answers to open questions in the field. Further, by utilising eye tracking as a gateway to the mind, we open a new path for the research in CA to explore subjectivity in the assessment of arguments.

We acknowledge that such study can be done with more quality dimensions and different types of texts, yet this was beyond the scope of our funding and equipment availability. Lastly, it is mention-worthy that different quality dimensions may come with their own challenges, making them potentially more difficult to work with and annotate. Therefore, the use of different quality dimensions for the different guidelines may have an effect on the results. We aimed to account for this influence by combining the use of the eye tracking data with the available questionnaire responses.

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. Barrett, M., Hollenstein, N.: Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. Language and Linguistics Compass **14**(11) (2020), `https://doi.org/10.1111/lnc3.12396`

3. 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
4. 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
5. Duchowski, A.T.: Gaze-based interaction: A 30 year retrospective. Computers & Graphics **73**, 59–69 (2018), https://doi.org/10.1016/j.cag.2018.04.002
6. 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
7. 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
8. Durmus, E., Ladhak, F., Cardie, C.: The role of pragmatic and discourse context in determining argument impact. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. pp. 5668–5678 (2019), https://aclanthology.org/D19-1568/
9. El Baff, R., Wachsmuth, H., Al Khatib, K., Stein, B.: Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. pp. 454–464 (2018), https://aclanthology.org/K18-1044/
10. 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/
11. Gleize, M., Shnarch, E., Choshen, L., Dankin, L., Moshkowich, G., Aharonov, R., Slonim, N.: Are you convinced? choosing the more convincing evidence with a siamese network. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 967–976 (2019), https://aclanthology.org/P19-1093/
12. Gonzalez-Garduno, A.V., Søgaard, A.: Using gaze to predict text readability. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 438–443 (2017), https://doi.org/10.18653/v1/W17-5050
13. 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
14. 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/
15. 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

16. Hollenstein, N., Zhang, C.: Entity recognition at first sight: Improving NER with eye movement information. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (2019), https://aclanthology.org/N19-1001/
17. 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/
18. Ivanova, R.V., Gubelmann, R.: The shift from logic to dialectic in argumentation theory: Implications for computational argument quality assessment. In: Proceedings of the 31st International Conference on Computational Linguistics (2025), https://aclanthology.org/2025.coling-main.321/
19. 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/
20. 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
21. 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/
22. 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
23. Mishra, A., Kanojia, D., Nagar, S., Dey, K., Bhattacharyya, P.: Harnessing cognitive features for sarcasm detection. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2016), https://aclanthology.org/P16-1104/
24. 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/
25. 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/
26. 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/
27. 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/
28. 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/

29. 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/
30. 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
31. 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
32. 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/
33. Sood, E., Tannert, S., Müller, P., Bulling, A.: Improving natural language processing tasks with human gaze-guided neural attention. In: 34th Conference on Neural Information Processing Systems (2020)
34. 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/
35. 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/
36. 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
37. 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
38. Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., Slonim, N.: Automatic argument quality assessment-new datasets and methods. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. pp. 5625–5635 (2019), https://aclanthology.org/D19-1564/
39. Toledo-Ronen, O., Orbach, M., Bilu, Y., Spector, A., Slonim, N.: Multilingual argument mining: Datasets and analysis. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 303–317 (2020), https://aclanthology.org/2020.findings-emnlp.29/
40. 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/
41. 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/
42. 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
43. Wachsmuth, H., Werner, T.: Intrinsic quality assessment of arguments. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 6739–6745 (2020), https://aclanthology.org/2020.coling-main.592/
44. 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
45. 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/