Amendment to Eichel and Schulte im Walde (2023): A PAP Case Study on AMT vs. Non-AMT Annotation

Annerose Eichel, Sabine Schulte im Walde
Institute for Natural Language Processing, University of Stuttgart
{annerose.eichel,schulte}@ims.uni-stuttgart.de

1 Motivation

In Eichel and Schulte im Walde (2023), we presented a novel dataset for physical and abstract plausibility of events in English. Based on naturally occurring sentences extracted from Wikipedia, we incorporated degrees of abstractness, and automatically generate perturbed pseudo-implausible events. We annotated a filtered and balanced subset for plausibility using crowd-sourcing, and performed extensive cleansing to ensure annotation quality. In-depth quantitative analyses indicated that annotators favor plausibility over implausibility and disagree more on implausible events. Furthermore, our plausibility dataset was the first to capture abstractness in events to the same extent as concreteness, and we found that event abstractness has an impact on plausibility ratings: more concrete event participants triggered a perception of implausibility.

As we conducted a relatively large annotation experiment via AMT (Amazon Mechanical Turk) crowd-sourcing, we aimed to apply post-processing methods minimising the impact of unreliable annotations on our analyses. With more than 500 different final annotators and a very subjective annotation task, we however noted the possibility of potentially wrong annotations due to errors, limitations of task instructions, or the interface (Pradhan et al., 2012; Poesio et al., 2019; Uma et al., 2022). This was especially true for the implausible portion of the dataset where no comparison with an attested triple label was possible. In our work, we considered approaches of mitigation which could be concentrating on triples with high (im)plausibility ratings or use e.g., probabilistic methods to aggregate labels. We thus provided a dataset version with labels aggregated using MACE (Hovy et al., 2013).

In this amendment to Eichel and Schulte im Walde (2023), we present an additional small-scale human annotation study to further understand the influence of AMT annotation of plausibility. Specifically, we zoom in on pseudo-implausible events which have been rated plausible by AMT annotators. In the context of this study, we aim to see whether our findings regarding plausibility are reproducible given a much smaller set of annotators (10 non-AMT vs. 500 AMT annotators) from a more diverse set of countries (non-AMT: 4 continents and 8 countries and vs. AMT: 2 continents and 2 countries). Our research questions are as follows:

- RQ1. Do non-AMT annotations of originally pseudo-implausible events rated plausible differ from AMT annotations (Eichel and Schulte im Walde, 2023)? If so, in which dimensions and how does this impact PAP annotation (aggregations) and findings?
- RQ2. Given originally pseudo-implausible events, are annotators more likely to rate an events as plausible if event participants are more abstract? Conversely, is event participant concreteness connected with annotators rating events are more plausible?
- RQ3. Do annotators agree more on what is plausible as opposed to what might not be within the realm of plausibility based on their subjective opinion? Furthermore, do annotators disagree more when rating events which encompass more concrete, abstract, or mid-range event participants?

The remainder of this amendment is organized as follows. First, we present the design of the non-AMT human annotation task and setup. Second, we analyze the collected ratings w.r.t annotator agreement and present collection statistics. We then compare non-AMT vs. AMT annotation results and discuss our findings as well as recommendations for using the PAP dataset.

2 Non-AMT Human Annotation

2.1 Annotation Task

Task Design In this study, we focus on a sample from the PAP dataset (Eichel and Schulte im Walde, 2023) encompassing 81 pseudo-implausible events and 27 plausible events that have been rated plausible by AMT workers. Specifically, each of these events received a strict plausible majority, i.e. more than 70% of annotations in an aggregated binary setup (plausible vs. implausible)[1].

Each event participant p can be either *highly abstract* (a), *mid-range* (m), or *highly concrete* (c). Taking the Cartesian product, we define 27 possible event combinations, e.g., events consisting of words with very high concrete ratings only, e.g., (c, c, c) or fully mixed events, e.g., (c, m, a). For each abstractness combination, we annotate 3 pseudo-implausible and 1 originally plausible events, amounting to 81 pseudo-implausible and 27 plausible triples.

Setup Within the study, we collect human-produced annotations of event triples with respect to subjective assessments of plausibility on a degree scale (1–5) ranging from implausible to plausible. In our task description, we replicate the general tone of the annotation task in Eichel and Schulte im Walde (2023) and make sure that annotators are informed about the necessity to provide their personal opinion regarding the plausibility of an event. Each triple is annotated by 10 annotators. In particular, we ask annotators to indicate whether a given sentence is implausible or plausible using a dropdown menu (corresponding to a scale from 1 to 4)[2]. To avoid bias, the dropdown is by default set to no value. Annotators are required to choose an option from the menu, thereby deciding for either plausible or implausible.

Corresponding to our objective of comparing non-AMT with AMT annotations, we recruit English native speakers[3] through an offline process. We recruit 1 participant from Africa (Nigeria), 3 participants from Europe (UK), 2 participants from North America (US, Canada), 5 participants from Asia (India, Nepal, Iran, Korea). Plausibility judgements are collected in a remote setup via Google Forms and Google Tables with declarations of consent and details for payment collected via Google Forms and plausibility ratings stored separately in Google Tables. We run a pilot for time frame assessment and find 30 minutes a very fair assumption of a time frame needed to read through all instructions, declare consent, and complete the study without any need for rush. We pay 7€ upon completion of the study, thus exceeding hourly minimum wage (12€/h, Germany, July 2023).

2.2 Analysis of Non-AMT Human Judgements

We assess Inter-Annotator Agreement (IAA) by calculating both pairwise joint probabilities as well as Cohen's Kappa considering each plausibility assessment in a multi-class as well an aggregated binary setup. Due to poor agreement, we exclude one participant (UK) from further analyses. Pairwise joint probabilities are shown in Fig. 1. Cohen's Kappa is presented in Fig. 2.

	# instances
No Disagreement	7
Bi-Disagreement	20
Tri-Disagreement	45
Quadruple-Disagreement	36

Table 1: Number of disagreeing values chosen by annotators for a given target. Bi-disagreement refers e.g., to annotators choosing either two values from one binary category, e.g. (*implausible, rather implausible*) or two disagreeing values (*implausible, plausible*). Tri-/Quadruple-Disagreement illustrates that annotators disagree regarding plausibility / implausibility.

We further analyze disagreements in non-AMT annotations. For this, we calculate the number of targets for which annotators disagree on 0 to 4 values of which they can choose. Results are presented in Table 1. Bi-Disagreement, for example, refers to cases where annotators chose from two different values, while annotators chose from all 4 values in case of Quadruple-Disagreement (33.33%of all cases).

2.3 Comparison to AMT Annotation

When comparing non-AMT to AMT annotations (cf. RQ1), we note the following observations. First, as show in Table 2, the non-AMT annota-

[1] The original annotations included a scale with five possible values between 1 and 5 corresponding to plausible and implausible, respectively. The value 3 is set as a default and cannot be chosen as a valid value which means that participants have to decide for either plausible or implausible.

[2] As we use a dropdown menu with the default set to no value at all, we have a 4-point scale with 1 corresponding to implausible and 4 considered plausible.

[3] 6 participants have a second native language.

	Original Label	AMT	Non-AMT
Plausible	27	108	45
Pseudo-Implausible	81	0	32
Disagree	-	0	31

Table 2: Absolute label distributions for original labels and strict majorities based on the original label as well as AMT and non-AMT annotations.

	Original Label	AMT	Non-AMT
Plausible	-	81	31
Pseudo-Implausible	81	-	27
Disagree	-	-	22

Table 3: Absolute label distributions for non-AMT annotation based on strict majorities for 81 originally pseudo-implausible events rated plausible by AMT annotators.

tions confirm the strict plausible majority label by the AMT annotations in 45 cases. For 32 events, non-AMT annotators agree on the event being implausible. Further, we find disagreeing annotations in 31 cases.

As shown in Table 3, non-AMT annotators find the pseudo-implausible event actually being plausible in 22 cases. For AMT annotations (at least as cleaned/aggregated by now), this number is a lot higher (by approx. 4 times). This indicates, that (i) to some extent (non-AMT annotations: 27%), pseudo-implausible events are in fact not plausible, and (ii) our AMT annotations tend to over-estimate the plausibility of events. We thus conclude, **generating fully implausible events automatically given our conditions is not trivial** and cannot be considered to be 100% clean. This is in line with our previous work.

Considering the hypothesis in RQ2 that implausibility might be easier to catch given more concrete event participants, we explore strict majority labels for both originally plausible and implausible events. As illustrated in Fig. 3, **implausibility seems to be easier to catch given a concrete verb** (peaks at: a-c-c, a-a-c), while subject and object abstract-/concreteness is quite well distributed among abstractness ranges. Moreover, **plausibility, seems to be connected with abstractness** and is annotated given mostly mid-range and abstract participants. **Disagreement can be mostly observed for mid-range events with mid-range/abstract verbs**, while more concrete verbs seem to trigger less uncertainty. These findings underline the results presented in Eichel and Schulte im Walde (2023).

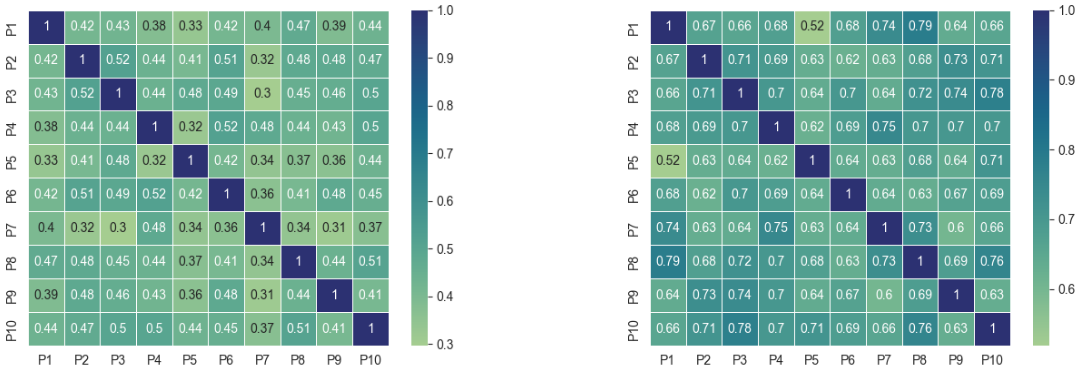
Next, we look at only originally pseudo-implausible events (excluding the originally plausible control events). As shown in Fig. 4a, it can be observed that implausibility labels increase with higher abstract-/concreteness. Peak values are annotated for events with concrete verb/objects. Disagreement is mostly seen for mid-range events where, in particular, the verb is never concrete. While plausible majority labels can be observed, they are not as dominant with peak values for mostly mid-range events with mid-range or abstract objects and mixed verbs. Regarding RQ3, we hold that **events with clearly concrete or abstract event participants seem to trigger less disagreement**, while mid-range events seem to be less clear. This observations further refines our previous findings from concreteness in general to the importance of concrete verbs. We further note that implausibility is observed to trigger slightly more disagreement, however, this is true for events with more abstract event participants only.

3 Discussion

Comparing AMT vs. Non-AMT Annotation When comparing between AMT vs. non AMT annotations for plausibility assessment with a focus on pseudo-implausible events, we do encounter differences. A key finding is that AMT annotation might over-estimate plausibility as compared to non-AMT annotations. The share of plausible events is, however, still bigger than the implausible share. Additionally, a quite substantial number of events are disagreed on (no 70% majority could be reached). We note that our non-AMT study is comparably small and focuses on a specific set on PAP samples. Hence, results could be very different for other portions of the dataset, i.e. events originally labelled plausible.

As presented in the previous section, meta-level findings (RQ2 and RQ3) are consistent across both studies with even more refined findings for RQ3.

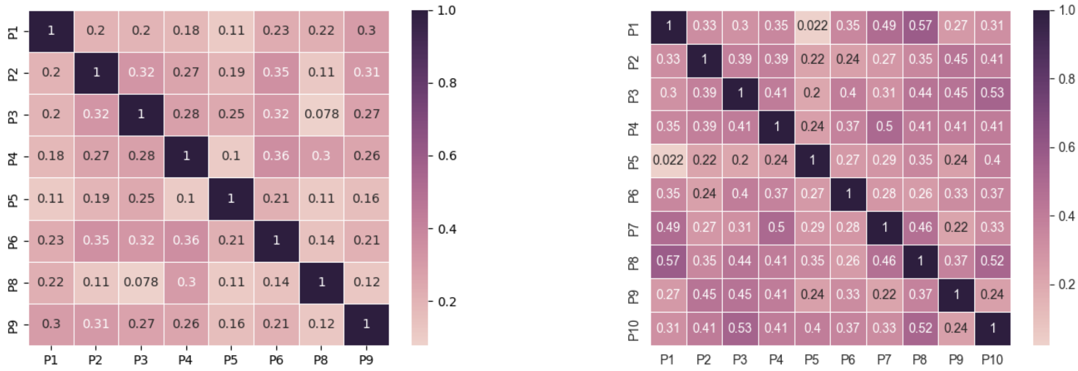
Recommendations Regarding Annotation Aggregation As part of the original dataset, we released probabilistic aggregations based on MACE (Hovy et al., 2013). While these might over-estimate plausibility in comparison to the non-AMT annotations, we observe an overlap of aggregations with non-AMT annotations where MACE labels correspond to implausible or disagree in 90% of cases. In addition, the PAP dataset still repre-



(a) Pairwise Joint Probability using a multiclass setup.

(b) Pairwise Joint Probability using a binary setup.

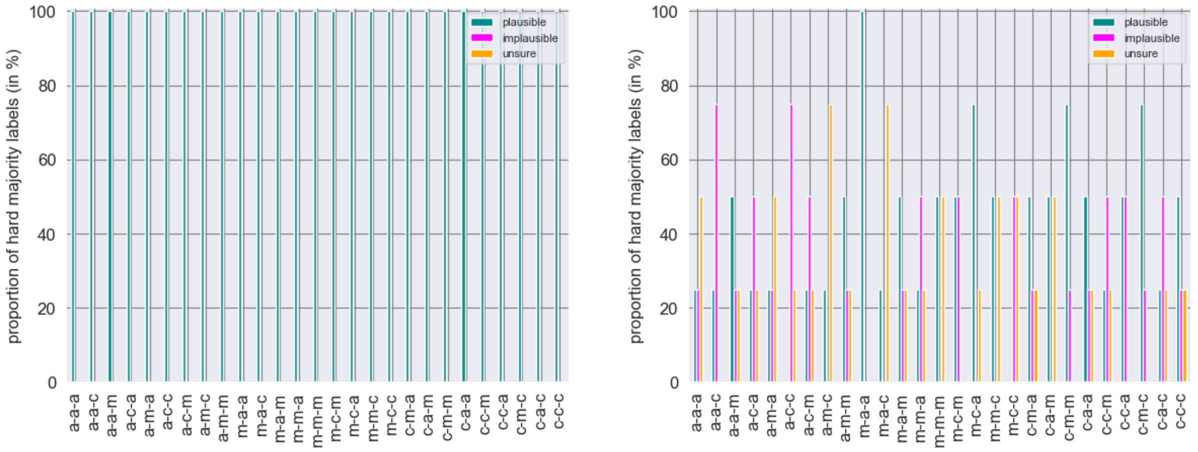
Figure 1: IAA: Pairwise Joint Probability based on a multiclass or binary setup.



(a) Pairwise Cohen's Kappa using a multiclass setup.

(b) Pairwise Cohen's Kappa using a binary setup.

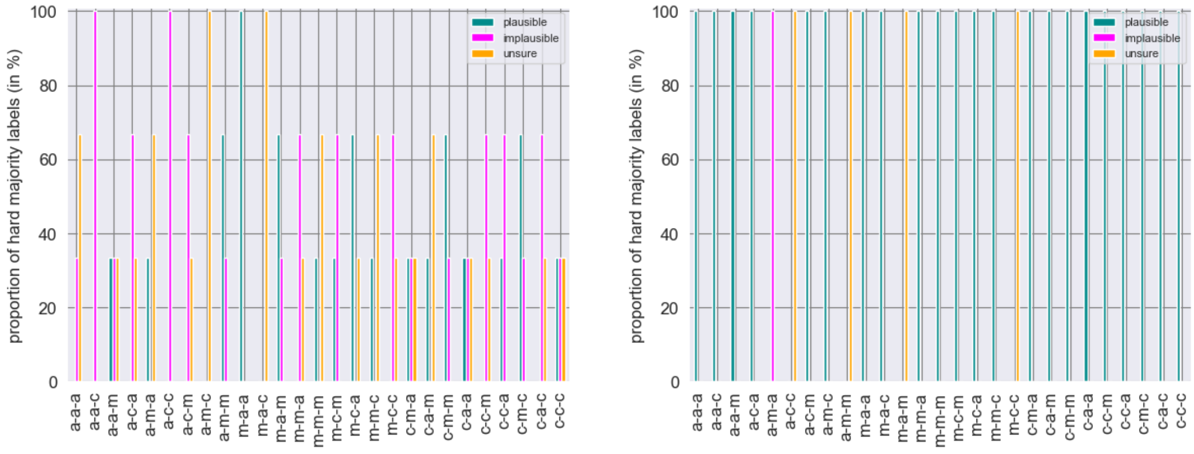
Figure 2: IAA: Pairwise Cohen's Kappa based on a multiclass or binary setup.



(a) Strict majority labels based on AMT annotation.

(b) Strict majority labels based on non-AMT annotation.

Figure 3: Strict majority classes per combination with (a) illustrating the distribution as annotated by AMT workers. On the right side, (b) shows the distribution of majority labels yielded from a non-AMT annotation. For each combination, we annotate 3 pseudo-implausible and 1 originally plausible events with a strict plausible majority as annotated by AMT workers. Original label distribution: 81 pseudo-implausible events, 27 plausible events.



(a) Strict majority labels based on non-AMT annotation for originally pseudo-implausible events.

(b) Strict majority labels based on non-AMT annotation for originally plausible events.

Figure 4: Strict majority classes per combination. On the left, (a) shows the label distributions for 81 *originally pseudo-implausible events* with 3 instances annotated per abstractness combination. All instances had been annotated plausible by AMT annotators, while the here presented non-AMT annotation is not as unanimous. On the right, (b) presents majority labels for 27 *originally plausible events* which are included as control instances. Here, only 1 instance per abstractness combination is annotated.

sents valid real-world annotations that are collected and thoroughly cleaned in a meaningful way. We thus recommend using our provided MACE aggregations in case raw annotations cannot be used.

Other aggregation which could be explored include the following: It might be helpful to weigh annotations based on extremes, i.e., weighing an annotated 1 more than a 2 as the former would be considered to more implausible than the latter. Furthermore, raw annotation distributions could be used directly to capture agreement and disagreement in the most direct way. Depending on the task, it can also be considered to use only a sample of PAP, namely events originally labeled plausible since the overlap between annotated and original label is quite high for these kinds of events. Lastly, the agreement threshold could be set more strictly thus considering only events with a very clear vote for either plausible or implausible.

4 Conclusion

We re-annotated a sample of the PAP dataset (Eichel and Schulte im Walde, 2023), focusing on originally pseudo-implausible events which were annotated as plausible by AMT annotators. Replicating the original study with non-AMT annotators, we find AMT annotations skewed towards higher plausibility while non-AMT annotations are less clear in their judgement. On the level of results, however, we still find event participant abstractness connected with plausibility and implausibility to be easier to catch given a concrete verb. Refining previous findings (Eichel and Schulte im Walde, 2023), we find disagreement being connected to mostly mid-range and abstract event participants, while more concrete participants seem to trigger less disagreement.

References

Annerose Eichel and Sabine Schulte im Walde. 2023. A dataset for physical and abstract plausibility and sources of human disagreement. In *Proceedings of the 17th Linguistic Annotation Workshop*, Toronto, Canada. Association for Computational Linguistics. Forthcoming.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. 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)*, pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Alexandra Uma, Dina Almanea, and Massimo Poesio. 2022. Scaling and Disagreements: Bias, Noise, and Ambiguity. *Frontiers in Artificial Intelligence*, 5.