Interrater Disagreement Resolution
A Systematic Procedure to Reach Consensus in Annotation Tasks

Yvette Oortwijn[*†] and **Thijs Ossenkoppele**[*] and **Arianna Betti**[*]
[*]University of Amsterdam, Institute for Logic, Language and Computation
[†]Eindhoven University of Technology, Algorithms, Geometry & Applications
{y.oortwijn, t.ossenkoppele, a.betti}@uva.nl

Abstract

We present a systematic procedure for interrater disagreement resolution. The procedure is general, but of particular use in multiple-annotator tasks geared towards ground truth construction. We motivate our proposal by arguing that, barring cases in which the researchers' goal is to elicit different viewpoints, interrater disagreement is a sign of poor quality in the design or the description of a task. Consensus among annotators, we maintain, should be striven for, through a systematic procedure for disagreement resolution such as the one we describe.

1 Introduction

A growing body of literature signals a thorny issue with assessing general progress in the field of natural language processing (NLP) as part of artificial intelligence. Benchmarks that are considered 'general', and are widely used as standards to assess NLP systems' performance, turn out to be rather specific, and hence of more limited significance than commonly acknowledged (Raji et al. 2020; Schlangen 2020). Good performance on specific benchmarks does not guarantee good performance across the board (Faruqui et al. 2016; Bakarov 2018; Ethayarajh and Jurafsky 2020): it only helps with gaining understanding of how certain systems work for those specific benchmarks. In order to claim progress across the board, one would need to evaluate system performance on a certain reasoned series of such specific benchmarks, that is, results on a host of "more focused and explicitly defined problems" (Raji et al., 2020, 1). To enact this, one would need a ground truth for the evaluation of each specific task-*cum*-dataset, including ground truths in expert domains.

Ground truth construction is challenging. In this paper we focus on the process of constructing ground truths via semantic annotations tasks. Recent studies stress the intrinsic difficulty of semantic annotation due to vagueness and ambiguity (Aroyo and Welty 2015; Kairam and Heer 2016; Pavlick and Kwiatkowski 2019). Importantly, some argue that interpretative disagreements due to different conceptualizations or perspectives cannot be seen as just 'mistakes' (Sommerauer et al. 2020; Herbelot and Vecchi 2016). It is our tenet that in ground truth construction differences in conceptualizations or perspectives can and must be explicitly specified as an integral part of annotation tasks; moreover, interrater disagreement is not necessarily due to inherent ambiguities in the data, but at least in part to the annotation task being underspecified, in particular as to the right context to consider.

Take annotation tasks involving relatedness or similarity judgments, which are key types of judgment for NLP evaluation. Similarity is not a property of two things by themselves in isolation: it is always judged by a specific standard, and by weighing properties of the things compared in different ways, according to a context (Goodman 1972; Batchkarov et al. 2016). When people judge by different standards[1], disagreement arises as a matter of course - and is especially likely when annotating texts of high conceptual density, as this requires a lot of prior knowledge and interpretation. In order to get comparable and meaningful annotations, judgment standards need to be aligned and made extremely transparent.

In this paper we propose a six-step systematic procedure for interrater disagreement resolution in which conceptual alignment figures as one of the steps. The procedure is designed to facilitate the resolution of interrater disagreement that fre-

[1]As Gladkova and Drozd (2016) point out, similarity is defined by Turney and Pantel (2010) as co-hyponymy (e.g. car and bicycle), whereas Hill et al. (2015) define it as "exemplified by pairs of synonyms; words with identical referents" (e.g. mug and cup).

Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pages 131–141
Online, April 19, 2021.

quently arises in annotation tasks in which multiple annotators participate. The emergence of disagreement in annotation tasks is valuable information, albeit of a negative type: barring cases in which the researchers' goal is none other than eliciting disagreement, interrater disagreement, we maintain, is a sign of poor quality in the design or the description of a task. In ground truth construction, consensus among annotators should be striven for. The procedure applies to a wide range of annotation tasks, namely every task involving the application of one or more concepts to a unit of annotation (a fragment of text, such as a paragraph or a sentence, or a more artificial unit, such as a string with a length of *n* characters). We hold that the benefit of a systematic procedure of resolving interrater disagreement is twofold: first, such a procedure leads to the construction of reliable and well-grounded datasets, and second, it ensures that the resolution proceeds in a non-arbitrary fashion allowing for proper documentation and replicability of the data.

2 Related work

Computational research: interrater agreement, dataset creation and ground truths Standard methods for measuring interrater agreement and reliability (Artstein and Poesio, 2008) such as (Cohen's) kappa (Cohen 1960; Landis and Koch 1977) and Krippendorff's alpha (Krippendorff, 2013) output a single score to represent the agreement between different raters. Methods such as the CrowdTruth framework (Aroyo and Welty 2014; Aroyo and Welty 2015) give a more detailed disagreement analysis, though only in post-annotation phase. Similarly, Kairam and Heer (2016) mention that disagreement cannot simply be treated as noise and propose a post-annotation method for identifying different valid interpretations annotators may use to come to different conclusions. By contrast, we take disagreement analysis and resolution as internal to the annotation procedure.

Sommerauer et al. (2020) stress difficulties with annotation due to ambiguity or vagueness in language while studying cases in which disagreement between different annotators is expected and multiple answers are legitimate. Our focus is datasets that are meant to be used as ground truths. In ground truth construction, we argue, it is necessary to resolve cases of disagreement (disagreement resolution phase, see *step 5* below), and, more importantly, dispel the ambiguities that cause disagreement (if ambiguity is the cause of the disagreement) by task specification, either by redesigning the task or by making the annotation guidelines more precise (conceptual alignment phase, see *step 2* below). We do recognize that genuine disagreement might exist due to e.g. ambiguity in language in existing datasets (see also, Palomaki et al. (2018)), but we see legitimate disagreement as having a specific meaning: it is either a signal that further resolution is needed (through annotation task redesign or guideline redefinition), or it is the possible result of a task specifically designed to chart or elicit instances of disagreement, as in Sommerauer et al. (2020) or Herbelot and Vecchi (2016).

We offer a procedure by which annotators can avoid disagreement due to unclarity of the task, accurately discern the reason for disagreement whenever it arises, and make a deliberate decision on how these cases should be annotated. Any differences between 'people's beliefs about the world' (or the data), we say, should be explicitly integrated in task design such that annotators are required to judge according to a certain perspective or set of beliefs, and not from an absolute point of view. We agree with Pavlick and Kwiatkowski (2019) that disagreement between annotators cannot simply be seen as noise in the data supposedly due to low-quality annotations. However, while they divide the annotations into consistent units to get sets of consistent gold labels, we argue that in ground truth construction the variety of human judgments can and should be narrowed down to exactly one type by specification of the task. In our case, the process of identifying reasons for disagreement is part of the annotation process, which allows for resolution of disagreement and thereby a dataset suitable for use as a ground truth for the task at hand.

In Betti et al. (2020), a general method for constructing expert-controlled ground truths for concept-focused domains is proposed, and the construction for an actual ground truth for a philosophical corpus is described. Disagreement resolution is mentioned, and one example of resolution is reported, but no explicit general methodology for disagreement resolution is offered.

It has been emphasized that the conditions under which a dataset has been created need to be properly documented to allow for reproducibility and replicability (Bender and Friedman 2018; Paullada et al. 2020; Hutchinson et al. 2021). Language models are known to pick up and reinforce exist-

ing biases in data (see, e.g., Bolukbasi et al. 2016; Zhao et al. 2017). Bender and Friedman (2018) offer instructions on how to document data using data statements to help reproducibility and replicability, bring existing biases to the surface and improve representation in future dataset creation. The procedure we propose asks for explicit decisions from raters after deliberation. This requirement makes the conditions of dataset creation clear, thus allowing proper documentation.

Philosophy Peer disagreement is a topic of investigation in philosophy, in particular in the subfield of social epistemology. A large amount of literature exists on issues concerning both peer disagreement (e.g. Goldman and Whitcomb 2011; Christensen and Lackey 2013) and group decision making in the face of such disagreement (e.g. List 2005), but resolution procedures that aid in moving from peer disagreement to unanimously agreed upon results are not proposed, and are in general '[...] at best rare in scientific contexts.' (de Ridder, 2014). One of the scarce examples is Gius and Jacke's (2017) procedure for resolving interrater disagreement in literary corpus annotation. Although similar in approach, our work improves on the latter in terms of applicability: we intend our procedure to be fit for all annotation tasks that involve the application of one or more concepts to units of annotation, while Gius & Jacke focus on tasks within literary analysis exclusively. Note that annotation tasks in which concepts are applied to units of annotation are frequent: any task involving the identifying of instances of any concept qualifies. For example, in our validation example in section 5.2 the annotation task requires annotators to identify wide-scope claims in the text of journal articles (that is, instances of the concept of *wide-scope claim*).

3 Ground truths and interrater agreement

In Pivovarov and Elhadad (2012) a Cohen's kappa of 0.68 is "accepted as representing a substantial amount of agreement between annotators". By contrast, in Betti et al. (2020) the initial interrater agreement of 0.65 was taken as a starting point to reach further consensus. When the aim of the annotation is e.g. to get an overview of the variety of ways in which people interpret statements, then interrater agreement need only be high on statements for which there is only one obvious interpretation and so agreement is expected. However, when the annotations are supposed to establish a ground truth, interrater agreement, we argue, should be 1.

One strategy used for getting the interrater agreement on the ground truth to 1, is to discard disputed annotation(s) (see, e.g., Kenyon-Dean et al. (2018)). But clearly this is loss of valuable information: for the purpose of training and evaluating a computational system we want to be as specific as possible as to what its output needs to be; by tossing out disputed annotation we underspecify what the right output on the matter is. Consider one of the examples in Herbelot and Vecchi (2016): "MISSILES EXPLODE received the labels SOME, MOST and ALL. It is likely that the SOME interpretation quantifies over missiles which actually explode, while the MOST/ALL interpretation considers the potential of a missile to explode". For ground truth construction, it is necessary to specify whether an annotator should e.g. take an actual or potential interpretation, to prevent annotators from making arbitrary choices or introducing unknown biases.

So, if an annotation data set is to be used as a ground truth, agreement should be the aim. When disagreement arises, it is important to identify why it arises, and make well-grounded decisions on how to deal with it. In the next section, we will outline a procedure for annotation through which different reasons for disagreement can be identified and which specifies directions for resolution of each of these types of disagreements. The procedure results in a reproducible dataset by forcing annotators to make well-grounded, and thereby traceable decisions on their annotations. Note that traceability makes the procedure relevant to all annotations, not just ground truth construction.

The annotation procedure supposes what we call an 'annotation toolbox' consisting of (i) the annotation task or question, (ii) the guidelines specifying the instructions for annotation and (iii) some kind of definition or characterisation of the key concepts involved (see *step 2*). Fixing the definitions and characterisations of these concepts is essential to the conceptual alignment of annotators and for subsequent use of the resulting annotations. The use of the annotation toolbox also facilitates disagreement resolution insofar as annotators can refer to elements of the toolbox to give a justification for their scoring. This also means that if disagreement cannot be resolved by referring to elements of the toolbox, the toolbox is incomplete, or in any case insufficient as a basis for annotation. In this

case, further expert research might be necessary to supplement the annotation toolbox. Based on the newly supplemented annotation toolbox, previous annotations might have to be redone, for there is no guarantee that these would end up receiving the same scoring. If such a resolution or supplementation is deemed impossible, the annotation cannot be completed and cannot lead to a dataset that is suitable as a ground truth.

4 The annotation procedure

What follows is a description of the steps of the annotation procedure (see flowchart in figure 1). Throughout this description we will talk of 'scoring' as the act of annotating a single unit. This is intended to also refer to types of annotation that are more adequately called 'categorizing', 'labelling' or otherwise. Note that with the exception of cases in which *step 0-2* is performed by the same group of researchers as *step 3-5* (see, e.g., section 5.1 in which the annotation procedure of Betti et al. (2020) is described), the annotators should be under close supervision of the researchers formulating the research question, and those setting the annotation task and guidelines, throughout all steps of the procedure.

4.1 The procedure

Step 0: Research setup and hypothesis forming In this initial phase, the prior research is done which indicates the need for an annotation task, research question(s) and hypotheses to be tested are formulated, and an annotation task is distilled to test these hypotheses. If at any point it is noticed that the research question or hypotheses are ill-defined or the annotation task does not match the research question, one should return to this step and start the process anew.

Step 1: Setting up annotation task and guidelines In this phase, the annotators are either presented with or set up themselves both 1) the annotation task, and 2) a set of annotation guidelines that guide 1). Ideally the annotators are already involved in the task and guideline set up since this improves the understanding of the task. 1) is immutable; if for some reason during the annotation procedure the task changes, the annotation procedure is reset and new guidelines must be set up that correspond to the new task. 2), however, is mutable; it can happen that new insights emerge during the annotation procedure that call for additional annotation guidelines or for an improvement of the existing ones. In case setting up the annotation task and guidelines requires additional research, one should return to *step 0*.

In developing the guidelines, researchers should consider how to score units that are ambiguous and therefore might endorse more than one interpretation. We recommend that instead of using, e.g., a simple binary scoring system, an "ambiguous" score is added to prevent forcing a decision. Forcing decisions could lead to arbitrariness, while ambiguity is still a real part of natural language that should be reflected in annotation. It should be ensured that this category won't mask unclarity in the task or the guidelines, by asking annotators to specify the source of unclarity (e.g. lexical ambiguity).

Step 2: Interrater conceptual alignment In this phase, the researchers identify the key concepts[2], and make sure that all annotators agree on the meaning or function of those concepts in the context of the task by specifying the definitions and characterisations for these concepts. In case researchers and annotators are two different sets of people, the annotators should be trained by the researchers in the concepts relevant to the task. The annotation procedure cannot move beyond this step if no interrater conceptual consensus is reached; this type of mismatch will almost certainly result in irresolvable conflicting annotations. Complex concepts, viz. concepts that involve many subconcepts when unpacked (e.g. philosophical concepts) require unpacking in the form of an interpretive model in the sense of Betti and van den Berg (2014). In these interpretive models, relations between subconcepts in the definition or characterisation of the concept modelled are made explicit. This facilitates the identification of instances of complex, rich concepts such as *epistemology* (see section 5.1). Such elaborate specification might not be required for simpler, or already well-defined concepts used consensually in different domains; in such cases, we expect less elaborate methods to suffice.

After consensus is reached on all key concepts that the annotators are aware of at this stage, the annotators can be expected to have an equal understanding of these concepts, which they can apply in

[2]By 'key concepts' we mean concepts mentioned in both task and guidelines. Note that settling on a definition for a concept at this step might require adding further new concepts to the guidelines in *step 1*, which should in turn be settled in *step 2*.

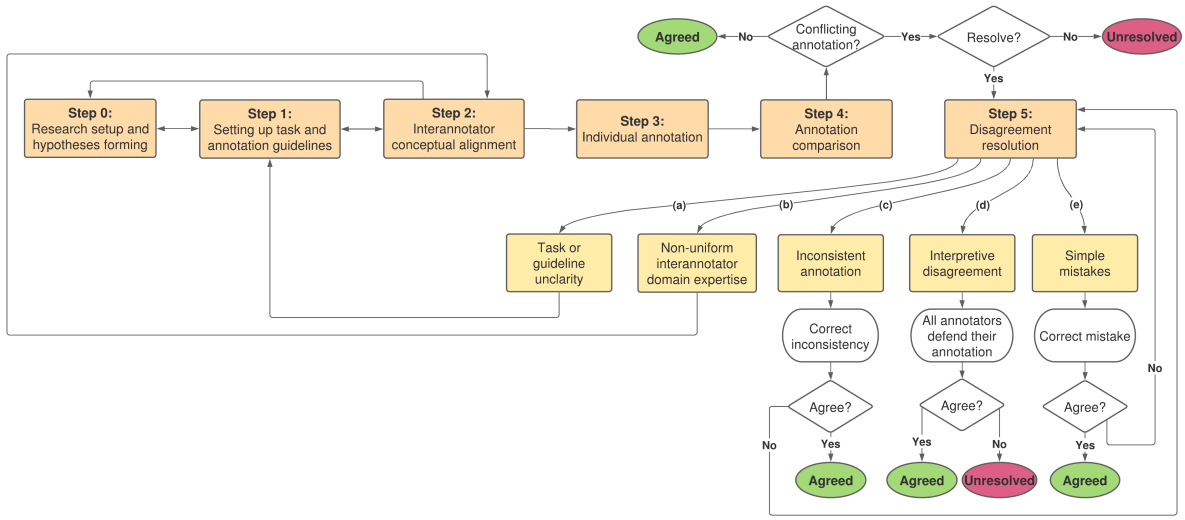

Figure 1: This flowchart serves as a summary of the annotation procedure detailed in section 4. The oval boxes contain the resulting annotations, green for agreed and pink for unresolved annotations. See https://github.com/YOortwijn/HumEvalDisRes to view the image separately.

annotating the units. As we observe in our second test case (section 5.2), questions for which there are issues with conceptual alignment receive lower interrater agreement than questions without such issues. The annotations for these questions should be redone after returning to this step for proper conceptual alignment.

Similar to *step 1*, it is possible that for the definition of concepts it is necessary to do further research, in which case one should return to *step 0*, or to further specify the task or guidelines, in which case one should return to *step 1*.

Step 3: Individual annotation Next, the annotations are performed according to the annotation guidelines specified in *step 1*. The manner in which the individual annotation proceeds depends on the guidelines, but as a general rule all annotators should score independently from each other to prevent being influenced by each other's scores.

Step 4: Annotation comparison After the individual annotation process, the annotations are compared. The comparison ideally yields a large set of agreed-upon annotations, but will likely also yield a set of conflicting annotations. For the latter, the disagreement resolution procedure should be put into operation. As mentioned in section 3, if conflicting annotations are simply discarded, we obtain an incomplete dataset which is not fit for use as a ground truth. Moreover, in such cases, hidden unclarities are likely to persist in the task or guidelines (see *step 5, a* below); as a consequence, we cannot trust previously agreed-upon annotations to reflect genuine agreement. We recommend in any case that it be specified whether the annotation procedure for the dataset under consideration has proceeded beyond this step; for, if not, then no attempt has been made to even check for inconsistent scoring by the same annotator (see *step 5, c*).

Step 5: Disagreement resolution We identify five main sources of interrater disagreement:

(a) *Task or guideline unclarity*. Among the possible reasons for interrater disagreement are 1) at least one annotator made a judgment based on a deviant interpretation of the nature of the task, and 2) the guidelines harbor residual unclarity as to the individual annotation procedure due to e.g. missing or vague instructions.

In case 1), the annotators should achieve a uniform understanding of the task through discussion. Different construal of the task can be due to poor or missing definition of the concepts involved in it. In this case the annotators should return to *step 2*. For other task unclarities the annotators should return to *step 1*. Recall that the task is immutable, so if it becomes apparent that the annotators cannot agree on what the task to be performed is, the whole annotation procedure should be abandoned; there is no justification for continuing an annotation task that is not equally clear for all annotators. The annotators will have to restart the procedure and redefine the task in such a way that all annotators

understand what is expected of them.

In case of 2), the annotators should return to *step 1* to reconsider the guidelines and, depending on the source of confusion, amend or supplement them. This should not be a controversial practice: it is not the task itself that is amended, but only the lines along which it is carried out most successfully. Note that in cases of drastic changes to the guidelines[3], the whole individual annotation process likely needs to be redone. This option should be duly considered since this situation casts doubts also on the cases of agreement in the dataset .

(b) *Non-uniform interrater domain expertise*. Despite having gone through *step 2*, there still may be differences in the amount of background knowledge that the annotators bring to the individual annotations. A difference in background knowledge used in annotating can cause diverging annotations. An example of divergence of this kind is when annotators align on the wrong width of some concept, i.e. a too narrow or too broad definition or characterisation of the concept, in which too many or too few aspects of that concept are considered. Mismatch in concept width among annotators is bound to lead to diverging annotations. In such a case, the annotators have to return to *step 2*.

(c) *Inconsistent annotation*. An annotator can have annotated inconsistently by scoring two units differently that should be given the same score (e.g. because the two units are functionally synonymous). In this case, the inconsistent annotator must decide whether they agree with the other annotators. If so, the scoring of the inconsistent units can simply be corrected and the disagreement is resolved. Reconsideration might however lead to rescoring such that the inconsistency is resolved, but the disagreement is not. In such cases, disagreement resolution won't be of type (c), though, and must be discussed under (a), (b), (d), or (e).

(d) *Interpretive disagreement*. Interpretive disagreement arises when, despite the fact that the annotators have reached conceptual alignment, there is disagreement about the purported meaning of certain terms in some unit. Annotators might hold a different interpretation of a certain unit even when they have an equal understanding of the concepts used in that unit, for example due to the use of an ambiguous term. The way these disagreements will have to be resolved is case-dependent. All annotators should defend their choice by stating the reasons for annotating the way they did. They should try to convince the other annotators by (rational) argumentation that their reading is the correct one. The annotators should then together weigh each others' reasons and see whether agreement can be reached. Whether the disagreement can be resolved or not depends on whether the annotators can settle for one interpretation that they all agree on. In some complex cases, deliberation might need to be postponed until research on the phenomenon encountered has sufficiently progressed.

(e) *Simple mistakes*. If it is suspected that an annotator has made a simple mistake somewhere (a typo, or disagreement about a unit that should not be controversial), this has to be pointed out to the annotator concerned. If they agree that they have made a mistake, the annotation can be corrected.

4.2 Unresolved Annotations

By identifying the source of disagreement and, if necessary, clarifying the task or guidelines for annotation, updating and repeating the (relevant parts of the) annotation procedure should result in a complete set of agreed-upon annotations. If there are structural unclarities in the task or annotation guidelines, it might be necessary to redo the individual annotations at *step 3*, and subsequent steps, after the task and guidelines have been clarified (*step 1-2*). Further research might also be needed to solve some disagreement (*step 0*) in which case the annotation process should be halted.

In case the resolution procedure has still failed to resolve all disagreements but the annotation process has to be finished, it is possible to settle for a deprecated dataset. Two strategies to complete the annotation process commonly used in current annotation dataset creation are: 1) the conflicting annotations remain disagreed upon, with the resulting data loss and problems with usage of the dataset as a ground truth mentioned in section 3 as its consequence, or 2) a pre-appointed 'dictator' has the last say and resolves the disagreements by force. The dictator does so by either forcing particular decisions of their own choosing (in which case this part of the dataset is a single-annotator portion), or by applying some judgment aggregation method, such as majority rule. The benefit of choosing 1) is having a fully peer consensus-based

[3] What "drastic changes" are depends on the nature of the task, and on whether the changes have any bearing on the scoring of other, previously completed annotations.

annotation dataset, but this option imposes limits on the applicability of the resulting dataset as a ground truth. If 2) is chosen, there will be no unresolved disagreements, but the epistemic status of the annotation procedure is significantly compromised, not to mention the risk of having a dictator that makes wrong or capricious decisions. These options are up to those responsible for the resulting dataset. We argue against keeping any disagreements essentially unresolved (see section 3); at the same time, we also advise strongly against appointing dictators, as persistent peer disagreements reflect poorly specified tasks or unclear guidelines, and the forced resolution of these disagreements obfuscate such defects. Instead, a higher degree of conceptual alignment or a better specification of the annotation task or guidelines should be aimed for. If this is not possible, both the dataset and the cases of interpretive disagreement should be flagged as such, and a report should be made.

5 Test cases

By way of illustration and validation, in this section we outline two different user applications of the procedure we have observed, by two non-overlapping teams of domain expert annotators. The first application concerns a study of a complex, rich philosophical concept in the complete corpus of the works of a specific author. In this case, the annotators worked through the entire procedure. The second application concerns a study of the methodological justification given to wide-scope claims in academic literature. Although the corpus used in the second case is also from the field of philosophy, the annotation task is generic, and could have been performed on any type of scholarly article. The second team set up the research (*step 0*), annotation task and guidelines (*step 1*), but they did not settle on the meaning of all key concepts (*step 2*) before annotation. For the first case we will give examples for each of the reasons for disagreement, while for the second case we will focus on an issue due to the lack of conceptual alignment.

5.1 Epistemology in Quine

In this task, the annotators scored paragraphs in the work of the philosopher W. V. O. Quine for relevance on his views on epistemology.[4]

[4]For more information about the dataset, see Betti et al. (2020) and https://github.com/YOortwijn/HumEvalDisRes

The annotators started by creating an initial interpretive model at *step 0*. The annotation task and guidelines, formulated as part of *step 1*, were as follows: The annotators have to score paragraphs based on the degree of evidence they contain with respect to a research question (RQ) concerning the nature of Quine's naturalistic epistemology.

Guidelines: The annotators have three scoring options:

1: the paragraph contains strong evidence for some answer to the RQ.

0: the paragraph contains mild evidence for some answer to the RQ, or the annotator is not sure whether the paragraph contains sufficient evidence to answer the RQ.

-1: the paragraph does not contain enough evidence to answer the RQ.

As part of *step 2* the annotators expanded the initial interpretive model to make sure they had a clearly defined, shared conception of all key concepts. Without this, the annotators might have started the individual annotation phase with diverging understandings of the concept of e.g. epistemology and would presumably fail to score the same way, leading to many disagreements.

After *step 3* (individual annotations), the annotators had an interrater agreement of about $\kappa \approx 0.65$. After *step 4* and *step 5*, the identification and resolution of all the cases of disagreement, an interrater agreement of 1 was reached. The following are examples of each of the possible reasons for disagreement and how they were resolved:

(a) *Task or guideline unclarity*: In some of the annotated paragraphs, Quine merely talks about the views of different philosophers on epistemology, instead of expressing his own. After discussion it was decided to add to the guidelines the rule that these paragraphs do not provide evidence for the research question and hence should be scored -1.

(b) *Non-uniform interrater domain expertise*: There was disagreement about a passage in which the term "first philosophy" occurred without an explanation of that term in the same passage. Not all annotators agreed on the degree of evidence the passage provided without an explication of "first philosophy". After further conceptual alignment, the annotators agreed that "first philosophy" expressed a concept of central importance, and that an equal understanding of the matter among annotators was thus essential to the task. A characterisa-

tion for the term was fixed, and the units containing "first philosophy" were re-annotated in unanimous agreement.

(c) *Inconsistent annotation*: Two paragraphs that had to be annotated indicated Quine's blurring of the boundary between ontological statements and (natural) scientific statements, only in different wording. One annotator scored the two passages differently, and corrected this after notice from and discussion with another annotator.

(d) *Interpretive disagreement*: One annotator scored 1, the other two 0. Upon discussion, the first annotator explained to have read the unit as if Quine defended a view mentioned as the "straightforward view". After discussion, the annotator became convinced that this cannot be clearly said from the fragment, and thus consensus was reached on scoring 0, resolving the disagreement.

(e) *Simple mistake*: An annotator noticed disagreement about a paragraph that should not be controversial. In that paragraph, Quine quite straightforwardly states that mathematical logic is an example of a hard science. The unit was rescored and the disagreement was resolved.

5.2 Literature Reviews in the History of Philosophy

In this annotation task, annotators scored articles from the *British Journal of History of Philosophy* between 2017 and 2019 by checking their abstracts, introduction and methodological information for clear statements of inclusion/exclusion criteria for the sources the authors take into account, the completeness of the sources consulted, and the scope of the claims authors made on this basis.[5]

The annotation task was as follows: for each article, the annotators answer the following questions:
Exclusion/Inclusion

1. Does the article use a reproducible methodology with explicit inclusion and exclusion criteria to identify and find primary literature?
2. Does the article use a reproducible methodology with explicit inclusion and exclusion criteria to identify and find secondary literature?

Completeness

3. Does the article explicitly attempt to identify all available primary literature relative to the research question?
4. Does the article explicitly attempt to identify all available secondary literature relative to the research question?

Wide-scope claims

5a. Does the article argue for wide-scope historical claims, i.e., claims spanning multiple decades or periods or intellectual movements?

5b. If 5a is answered positively, does the article qualify the wide-scope claims?

Guidelines: The annotators will annotate the article by scoring '1' for yes, otherwise, by scoring '0'. In case of a discrepancy between the abstract and body of the article, the body (represented by the introduction and methodology section) will be leading. The annotators will also check section and subsection headings in order to identify other relevant sections related to the finding and use of primary and secondary literature.

The annotators did not construct interpretive models for the key concepts in the task/guidelines. This is understandable, given the low complexity of concepts involved. The problem, though, is that the team did not fix definitions or characterisations of all relevant terms from the outset either, as will be clear below, and by contrast with the annotations in section 5.1. Missing this essential part of the annotation toolbox is a shortcoming that resulted in an interrater agreement unnecessarily lower than it should have been. We will highlight one case of task or guideline unclarity (*step 5, a*).

During discussion on specific disagreements on the basis of our flowchart, the annotators noticed that they used different construals of what constitutes a *wide-scope claim*. While the annotators were able to resolve these disagreements on a case-by-case basis, it cannot be guaranteed that the agreed annotations would still receive the same scoring by the new considerations on what constitutes a *wide-scope claim*. Therefore, when in *step 5* of the procedure it is discovered that the interpretation of key terms should be refined, it is necessary to revisit all annotations. By following the first three steps of the procedure before starting the individual annotations, annotators are forced to settle on an interpretation of terms such as *wide-scope claim* before annotating. This way disagreement on many passages and the need to redo all annotations can be avoided.

The interrater agreement on this task was $\kappa \approx 0.71$ before disagreement resolution. The annota-

[5] For more information about the dataset, see https://github.com/YOortwijn/HumEvalDisRes

tors resolved all cases of disagreement using *step 5* of the procedure. 62% of the disagreements were determined to be inconsistent annotations (*5, c*), 21% were due to guideline or task unclarity (*5, a*), 10% were due to non-uniform interrater expertise (*5, b*) and 7% were simple mistakes (*5, e*).

Note that the two questions about *wide-scope claims* have a much lower interrater agreement of $\kappa \approx 0.45$. This can be explained by the problems concerning the different construals of what constitutes a *wide-scope claim* discussed above and emphasized the need for conceptual alignment. Note also that no cases of interpretive disagreement were identified. This is likely because, after the interpretation of concepts has been settled in *step 2*, there is not much need for extensive interpretation of the units of annotation in this annotation task.

6 Further applications

We have shown how the procedure applies to the two test cases discussed in section 5. However, our procedure is not limited to cases of that type. Concepts are involved in any type of annotation task, and any concept necessitates both interpretation and conceptual alignment.

Consider the case of Herbelot and Vecchi (2016) again: "MISSILES EXPLODE received the labels SOME, MOST and ALL.". Suppose we want to construct a ground truth of property-object pairs. The example shows that the guidelines should specify whether to use an actual or potential interpretation of property possession. Note, though, that settling for an interpretation often won't be enough: while annotating under a potential interpretation, the issue may arise whether objects should have the potential to have a property *actually* (most do, but some are faulty) or *teleologically* (all). By our procedure, these ambiguities become apparent, and disagreement can be resolved.

The two test cases of section 5 both have academics as annotators, but this is no intrinsic requirement of our procedure. For some linguistic tasks, being a native speaker of the relevant language is enough expertise to be able to grasp and apply the concepts involved in the task. Another matter is the common practice of resorting to crowdsourcing platforms[6] to construct large, non-academic annotation datasets. The practice is useful, but ill-suited to accommodate the type of disagreement resolution we envisage. Our take is that even though it might not always be possible to adopt the entire procedure for ground truth construction, we see no fundamental, theoretical problems with its application in a wide variety of cases.

[6]See e.g. https://www.mturk.com/

7 Conclusion and further work

In this paper we proposed a six-step systematic procedure for annotation focused on disagreement resolution. We argued that disagreement is the result of poorly specified tasks or guidelines, or of insufficient conceptual alignment among annotators. To avoid incomplete datasets unfit for use as ground truths, we set up the procedure in such a way that the identification and non-arbitrary resolution of different types of disagreement is facilitated. Disagreement resolution by a clearly defined procedure results in more reliable and well-grounded datasets. By identifying the cause of disagreement and giving appropriate instructions for resolution for each type of disagreement, our procedure ensures that the resolution proceeds in a non-arbitrary fashion allowing for proper documentation and increasing replicability of the data.

We have validated the effectiveness and the importance of our annotation procedure by two test cases. The first case shows that conceptual alignment by itself does not guarantee that annotators make no mistakes or only come across clarified concepts, indicating the need for disagreement resolution after annotation. The second case emphasizes the importance of task clarification and conceptual alignment prior to annotation. Without this, the likeliness increases of having to redo annotations due to different construals of terms influencing both conflicting and agreed-upon annotations.

In further work we aim to collect more use cases to test the applicability of the procedure to more varied types of annotations. Moreover, we want to consider in more depth the interplay of *step 0-2* and further elaborate on the idea of *key concept* at *step 2*.

Acknowledgements

We thank the anonymous reviewers for their time and helpful comments. We thank the UvA e-Ideas team for their valuable discussion of a draft of this paper. This research was supported by grants *e-Ideas* (VICI, 277-20-007) and *CatVis* (314-99-117), funded by the Dutch Research Council (NWO), and by the Human(e)AI grant *Small data, big challenges* funded by the University of Amsterdam.

References

Lora Aroyo and Chris Welty. 2014. The three sides of CrowdTruth. *Human Computation*, 1(1).

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

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

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. arXiv:1801.09536.

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 7–12, Berlin, Germany. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Arianna Betti, Martin Reynaert, Thijs Ossenkoppele, Yvette Oortwijn, Andrew Salway, and Jelke Bloem. 2020. Expert Concept-Modeling Ground Truth Construction for Word Embeddings Evaluation in Concept-Focused Domains. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6690–6702, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Arianna Betti and Hein van den Berg. 2014. Modelling the History of Ideas. *British Journal for the History of Philosophy*, 22(4):812–835.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

David Christensen and Jennifer Lackey, editors. 2013. *The Epistemology of Disagreement: New Essays*. Oxford University Press, Oxford, UK.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jeroen de Ridder. 2014. Epistemic dependence and collective scientific knowledge. *Synthese*, 191(1):37–53.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. arXiv:2009.13888.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Evelyn Gius and Janina Jacke. 2017. The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis. *International Journal of Humanities and Arts Computing*, 11(2):233–254.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42, Berlin, Germany. Association for Computational Linguistics.

Alvin I. Goldman and Dennis Whitcomb, editors. 2011. *Social Epistemology: Essential Readings*. Oxford University Press, Oxford, NY.

N. Goodman. 1972. Seven strictures on similarity. In *Problems and Projects*, pages 437–450. Bobbs Merrill, Indianapolis, IN.

Aurélie Herbelot and Eva Maria Vecchi. 2016. Many speakers, many worlds: Interannotator variations in the quantification of feature norms. In *Linguistic Issues in Language Technology, Volume 13, 2016*. CSLI Publications.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA. Association for Computing Machinery.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1637–1648, New York, NY, USA. Association for Computing Machinery.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert

Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Klaus H. Krippendorff. 2013. *Content Analysis - 3rd Edition : An Introduction to Its Methodology*. SAGE Publications, Inc., Thousand Oaks, CA.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Christian List. 2005. Group Knowledge and Group Rationality: A Judgment Aggregation Perspective. *Episteme*, 2(1):25–38.

Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng. 2018. A case for a range of acceptable annotations. In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018), Zürich, Switzerland*, volume 2276 of *CEUR Workshop Proceedings*, pages 19–31. CEUR-WS.org.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Rimma Pivovarov and Noémie Elhadad. 2012. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. *Journal of Biomedical Informatics*, 45(3):471–481.

Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, Alex Hanna, and Amandalynne Paullada. 2020. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses (ML-RSA)*, Online.

David Schlangen. 2020. Targeting the benchmark: On methodology in current natural language processing research. arXiv:2007.04792.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? Evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.