Detailed BJHP Interrater Disagreement Cases and Their Resolutions

Author(s): Yvette Oortwijn, Thijs Ossenkoppele, Arianna Betti, Annotator 1 (A1), Annotator 2 (A2)
Version: 1.1
Date of creation: March 30th 2021
Document origin:
https://docs.google.com/document/d/1OLVuV0bkZFa_jrWM1uupeXBOLOqcy1TuB_ogIt8JG6s/edit
Aim: To describe the interrater disagreement cases for the BJHP annotation task in detail, as well as their resolutions.
Released together with: The HumEval2021 paper 'Interrater disagreement resolution; A systematic procedure to reach consensus in annotation tasks' (2021), by Oortwijn et al.
On: Forthcoming

Disagreement Resolutions

General comment

We, A1 & A2, resolved 25 cases of disagreement annotations out of 80. We noticed that the first and major cause of the disagreements was inconsistent annotation: we could solve 15 disagreements through solution (c). The second cause of the disagreements was guideline unclarity: we could solve 6 disagreements through solution (a). 3 disagreements were cases of non-uniform interannotator expertise and solved through solution (b), while 1 disagreement was a simple mistake and solved through solution (e). We did not have cases of interpretive disagreement, probably because this type of annotations did not leave space for much interpretive task.

All in all, we think that the guidelines not only help summarizing what we do, but that they also are a useful tool for cataloguing annotator disagreement and coming to a solution. Moreover, the guidelines help us to think about the annotations scheme and correcting it. Hence, we recommend that in future annotation tasks this guideline is adopted.

Specific cases:

2017-1-5. Discussing the case we discerned guideline unclarity (solution (a)). A2 and A1 had different conceptions of what a qualification is. A2 read it sometimes as a general justification of the paper, sometimes as a solution to the problem introduced in the paper, and sometimes as a specific qualification of the specific wide-scope claim. By contrast, A1 read it is qualifying the scope of the specific wide-scope claim, that must refer to the specific wide-scope claim transcribed in 5b. In short, the term 'qualification; is ambiguous and we should revisit the guidelines. We decided to reformulate the guidelines along A1's lines as follows: " does the article qualify the scope of the specific wide-scope claims transcribed in 5b".

2017-1-7. There is still some unclarity about what constitutes a *wide-scope claim* (guideline unclarity (solution (a)). A2 annotated "a comparison will [...] show the strength and philosophical appeal of Sellars's views, which

A1 believes still hasn't been sufficiently appreciated by mainstream analytical philosophy. (p.108)" as a wide-scope claim. A1 was not sure whether this claim cannot be based on a representative sample of secondary works, and thus does not qualify as a wide scope claim. We thus were unsure about how to apply the concept *wide-scope claim* in specific contexts. After discussion, we agreed that a thorough justification of the claim in question implies that a lot of data has been surveyed (if we take the author seriously) so we opted for A2's interpretation.

2017-1-9. We had a case of non-uniform interannotator expertise (solution (b)). We had different interpretations of the following long quote: "Schelling maybe viewed as addressing and resolving a problem which faces Kant's theory of freedom and transcendental idealism, deriving from the challenge posed by Spinozism. (p.133)if an argument can be constructed from broadly Kantian premises to Schellingian conclusions, then that is all to the good from the point of view of elucidating an undoubtedly profound but also very puzzling work, one which moreover may readily seem, particularly in Hegelian eyes, to be leading the post-Kantian development away from its rationalist core into wilder, Schopenhauerian and proto-Heideggerean territory." (p.134) A2 read this as referring to different traditions. A1 reads this as a common way of talking in German idealism studies to refer to individual authors and to describe Kant's views as Kantian, Spinoza's views as a variety of Spinozism, Schelling's views as Schellingean. Because of A1's greater expertise in the period, we opted for his interpretation (i.e., not a wide-scope claim), even if we certainly agreed on the fact that the statement is obscure and difficult to interpret.

2017-1-9. This was a case of Inconsistent annotation (solution (c)) by A1. A1 did not interpret "Most interpretations of the Freiheitsschrift, however, concentrate on only one of these approaches, thus foreshortening their understanding of Schelling's enterprise. [...]" as a wide-scope claim, whereas he did annotate similar claims as wide-scope for different articles.

2017-2-3. These were simple mistakes (solution (e). A1 took the reference to the *Dialogues* to mean the paper is only about that book, but it is not and the research question cannot be answered by just appealing to the *Dialogues*.

2017-2-4. A2 doubted but scored the article 0 for wide-scope claim. This was a case of inconsistent annotation (solution (c)). The article makes a wide scope claim ("In this paper I examine Kant's account of science in the context of the experimental tradition of philosophy, particularly in relation to the generation dilemma of the eighteenth century. (abstract)"). So we went with A1's annotations.

2017-3-2. This was the same case of guideline unclarity discussed in 2017-1-5, (guideline unclarity (solution (a)). A2 and A1 had different conceptions of what a qualification is. See 2017-1-5 for further discussion. We decided, again, to go with A1's interpretation.

2017-3-4. Here, for A1 it was an inconsistent annotation (solution (c)). He did not annotate a qualification to a wide-scope claim as a qualification. So we went for A2's annotation.

2017-3-5. This a case of guideline unclarity (solution (a). There is, as in 2017-1-7, still some unclarity about what constitutes a *wide-scope claim*. A1 interpreted the research question ("I aim to elucidate the meaning and scope of Spinoza's vocabulary related to 'consciousness'.") as a wide scope claim because it concerns one term throughout the entire oeuvre of Spinoza. A2 did not interpret this as wide-scope because the claim was restricted to Spinoza. We decided to treat claims concerning one author as *not* being wide-scope, and went with A2's interpretation.

2017-3-8. We had a case of non-uniform interannotator expertise (solution (b)). A1 read this as a standard paper in Kant studies about a comparison between three authors (Cudworth, Rousseau and Kant) about some selected aspects of Kant's thought, without any attention to minor authors or broad historical developments. A2 took this as a wide-scope claim (since the three authors live in different time-periods). We decided to go for A1's annotation, given his greater expertise on these kinds of paper.

2017-4-4. This was a case of Inconsistent annotation (solution (c)) by A1. A1 did not interpret "Smith is in turn revealed as generating a major break with Hume – a break which, if based on a superior theory of moral foundations (as Smith thought it to be) has important consequences for how we treat Smith and Hume in both the history of philosophy and contemporary moral theory. (p.681)" as a wide-scope claim, whereas he did annotate similar claims as wide-scope for different articles.

2017-4-8. This was the same case of guideline unclarity discussed in 2017-1-5, (guideline unclarity (solution (a)). A2 and A1 had different conceptions of what a qualification is. See 2017-1-5 for further discussion. We decided, again, to go with A1's interpretation.

2017-5-2. Here, for A1 it was an inconsistent annotation (solution (c)). He did not annotate a qualification to a wide-scope claim as a qualification. So we went for A2's annotation.

2017-5-3. This was a case of Inconsistent annotation (solution (c)) by A1. A1 did not interpret "in spite of the relevance of the subject, there has not been much discussion of either of these two readings, in particular, as concerns Descartes' final answer on the controversy, contained in his second letter to More, of 15 April 1649, and there still remains a considerable gap in scholarship." as a wide-scope claim, whereas he did annotate similar claims as wide-scope for different articles." as a wide-scope claim, whereas he did annotate similar claims as wide-scope for different articles.

2017-5-5. This was Inconsistent annotation (solution (c)) by A2. A2 took the article to be only about two books, but the article is unclear and also mentions other works. Hence, there is no procedure for the identification of all works (condition 3), even if A2 annotated that there was. Other articles of the same type were annotated differently by A2.

2017-5-6. Here, for A1 it was an inconsistent annotation (solution (c)). He did not annotate a qualification to a wide-scope claim as a qualification (the qualification being: "we will draw on the scholarship of the German Egyptologist Jan Assmann in order to reassess the significance of Cudworth's theory of religion for later philosophical developments. (p.932) the apparently antiquarian interest in Egyptian theology can be shown to form a link between Cudworth and one of the key controversies of the late eighteenth century and thereby, one might add, the emergence of nineteenth century idealism. (p.934)). So we went for A2's annotation.

2017-5-8. This was the same case of guideline unclarity discussed in 2017-1-5, (guideline unclarity (solution (a)). A2 and A1 had different conceptions of what a qualification is. See 2017-1-5 for further discussion. We decided, again, to go with A1's interpretation.

2017-5-9. This was a case of Inconsistent annotation (solution (c)) by A1. A1 did not interpret "Many scholars have since pointed to the ways that Conway's system anticipates or prefigures Leibniz's, sometimes characterizing Conway as a proto-Leibnizian." as a wide-scope claim, whereas he did annotate similar claims as wide-scope for different articles.

2017-5-10. This was Inconsistent annotation (solution (c)) by A2. A2 took the article to be only about three books, but the article is unclear and does not specify a reproducible methodology for the identification of primary literature with explicit inclusion and exclusion criteria (condition 1). Other articles of the same type were annotated differently by A2.

2017-6-3. This was Inconsistent annotation (solution (c)) by A2. A2 did not annotate the claim "From the time of Augustine to the late thirteenth century, leading Christian thinkers agreed that freedom requires the ability to make good choices, but not the ability to make bad ones" as a wide-scope claim, even if she did annotate similar passages as wide-scope.

2017-6-4. This was Inconsistent annotation (solution (c)) by A2. A2 annotated a non-wide scope claim as wide-scope, whereas similar passages of non wide-scope claims were annotated as non wide scope.

2017-6-6. This was a case of Inconsistent annotation (solution (c)) by A1. A1 did not interpret "It is argued that contrary to what is often assumed, Wollstonecraft's conception of the imagination is not primarily characterized by its Romantic features, but rather by the close affinity she posits between reason and the imagination. (p.1138)" as a wide-scope claim, whereas he did annotate similar claims as wide-scope for different articles.

2017-6-7. We had a case of non-uniform interannotator expertise (solution (b)). A1 read this as a standard paper in German studies about a comparison between three authors (Fichte, Kierkegaard, and Sartre) about some selected aspects of Fichte's thought, without any attention to minor authors or broad historical developments.

A2 took this as a wide-scope claim (since the three authors live in different time-periods). We decided to go for A1's annotation, given his greater expertise on these kinds of paper.

2017-6-8. This was Inconsistent annotation (solution (c)) by A2. A2 took the article to be only about two books, but the article is unclear and does not specify a reproducible methodology for the identification of primary literature with explicit inclusion and exclusion criteria (condition 1). Other articles of the same type were annotated differently by A2.

2017-6-10. This was Inconsistent annotation (solution (c)) by A2. A2 took the article to be only about a number of articles, but the paper is unclear and does not specify a reproducible methodology for the identification of primary literature with explicit inclusion and exclusion criteria (condition 1), nor does it identify all the primary literature. Other articles of the same type were annotated differently by A2.