**Epistemic Annotation Guidelines 9.9.19**

**Overview**

The goal of this work is to identify and classify epistemic statements in the scientific literature. An epistemic statement is a statement of unknowns, hypotheses, speculations, uncertainties, including statements of claims, hypotheses, questions, explanations, future opportunities, surprises, issues, or concerns (more clarifications in the task description below) within a sentence.

We identify these statements using epistemic cues – how these unknowns, hypotheses, speculations, and uncertainties are communicated in text. These epistemic cues will already be identified and highlighted in the text and it is your job to determine if the cue is an epistemic statement or not and then secondly, which category it belongs to. If the epistemic cue is more than one word and includes another epistemic cue, please take the largest all-encompassing epistemic cue possible. If you disagree with the automated category, please change to the category that best fits the epistemic statement. To help identify the epistemic statements, there are also general templates to see if the statement can fit in or not. Generally, statements of fact or methodology are not epistemic statements because the statements tend to be certain or explain the experimental design. This is not to say that there are never epistemic statements in the methods sections. The third part of the task is to identify the scope or subject of the epistemic cue. Here we are adopting and modifying the guidelines from the BIOSCOPE corpus. The scope should be the largest unit that makes a sentence, which can include the whole sentence.

The task at hand has 3 parts: Given a sentence with an epistemic cue highlighted (1) determine if it is an epistemic statement, 2) if yes, classify the statement to a specific category, and (3) identify and highlight the scope of the epistemic cue. The end goal of this work is to determine goals for knowledge (next steps or questions) based on the epistemic statements to provide an overview of the questions in a field and to drive literature-based discovery (finding relevant articles to the goals for knowledge).

For each sentence please complete all tasks at hand. In the full guidelines (word documents part 1 and part 2) with all examples, under each epistemic category are the exact epistemic cues along with positive and negative example sentences with yellow highlight and red highlight respectively, the scope/subjects are denoted by angle brackets <>, and an explanation in orange in square brackets [].

**EPISTEMIC CUE CATEGORIES:**

**Levels of Evidence: how much evidence do we have? How confident are we in that evidence?**

1. FULL UNKNOWN = a statement that indicates something is not known (a lack of information), or information is presented for the first time (new or novel) and a significant amount of research is needed; not a statement about the absence of something. The next actionable step is to explore the unknown further to gain any insights. The template is: “<subject> is unknown”.
2. EXPLICIT QUESTION = an explicit statement of inquiry (with a question mark or question word such as how, where, what, why). The next actionable step is to find answers to the questions and/or discover methodologies that will help answer the question. The template is: “a specific research question: <question/subject>”.
3. ALTERNATIVE OPTIONS = an explicit statement of multiple (at least 2) choices, actions, approaches, or methods that needs to be experimentally determined, with no stated disagreements, including statements with an implied second option, such as “whether”. Presenting multiple options or seeking other options which implies the best option has not yet been chosen. The next actionable step is to determine the correct option or a better option. The template is: “multiple options presented including: <subject>”.
4. INCOMPLETE EVIDENCE = a positive or negative statement proposing a possible/feasible explanation for a phenomenon on the basis of limited evidence as a starting point for further investigation OR a statement that information is needed to support an assertion or claim, including both positive and negative statements. Either a statement that some evidence already exists, explaining how current findings support previous work, adding confidence to a claim OR a statement that information is limited, more research is needed or is ongoing including limitations – biases or short comings related to the study design and execution. The next actionable step is togather more evidence to support the claim OR conduct more research to determine the validity of the claim; Complete the partial picture; Consider the short comings and biases for the next experiment and how it can be addressed. Due to this being the largest category, we split up the linguistic cues into 4 categories based on the stage of the research: early, middle, late, reflection. You do not need to classify these when you are annotating. This is only to provide a little more categorization when going through examples. The template is: “need more evidence for <subject>”.
5. PROBABLE UNDERSTANDING = a statement staking a claim to the most likely explanation, relationship, or phenomenon; there is a good chance this understanding is correct. The next actionable step is to determine if the most likely option is correct of if it’s another option. The template is: “staking a claim that <subject>”.
6. SUPERFICIAL RELATIONSHIP = a statement about a connection, link, or association between at least 2 variables; connectedness between entities and/or interactions representing their relatedness or influence. The next actionable step is to confirm the connection, link, or association between variables; determine the full underlying relationship between variables**.** The template is: “need to determine the full underlying mechanistic relationship between <subject>”.

**Future Opportunities: an explicit statement of future needs**

1. FUTURE WORK = a statement of extensions, including next steps, directions, opportunities, approaches, or considerations of the described work that may be implemented at some future time point. Also, this includes a statement of suggestion or a proposal as to the next best course of action, especially one put forward by an authoritative body; advice telling someone what the best thing to do is. The next actionable step is to determine the next course of action based on this future work proposal. The template is: “the next steps are <subject>”.
2. FUTURE PREDICTION = a statement of extrapolation of given data into the future and/or from past observations; no reference to next steps. The next actionable step is to run the simulation or experiment to determine if the prediction is correct; publicize the outcomes of the study to the correct people. The template is: “the prediction is <subject>”.
3. URGENT CALL TO ACTION = a statement calling for immediate attention including an action needed to be taken immediately or information that needs to be disseminated immediately OR critical: being in or verging on a state of crisis or emergency OR urgently needed OR absolutely necessary. The next actionable step is to take the urgent action ASAP or distribute the knowledge ASAP. The template is: “urgent call is to <subject>”.

**Anomaly/Curious finding** =a statement of a surprising result, conclusion, observation or situation; the researchers were not expecting the result, conclusion, observation or situation but are intrigued by it. The next actionable steps are to explore the surprising result, conclusion, or situation more and determine if the result, conclusion, observation, or situation are repeatable. The template is: “a curious finding that <subject>, (given that…)”.

**Barriers: an obstacle preventing research from moving forward that needs to be overcome**

1. DIFFICULT TASK = a statement of something not easily done, accomplished, comprehended, solved, or complicated with a multitude of underlying pieces or parts; heterogeneity; not medical complications. The next actionable step is to create methods to study the complicated system and to better understand any piece of the complicated system; potentially overall field development needed. The template is: “a difficult task to <subject>”.
2. PROBLEM OR COMPLICATION = a statement of issues, problems, mistakes, or medical complications that are cause for anxiety and/or worry. The next actionable step is to determine the gravity of the concern and determine if it needs to be dealt with before the next experiment or study. The template is: “problems/complications include <subject>”.
3. CONTROVERSY = a statement of disagreement amongst researchers OR a lack of consensus OR at least two possible answers are presented as results from different researchers - usually in reference to previous results and stated when results disagree with each other OR contradictions. The next actionable step is to determine the truth to break any disagreements. The template is: “need to resolve disagreements between/about <subject>”.

**Question answered by this work** = a statement of a goal or objective of a study that is attempted or completed during the study. The next actionable step is to find the answer(s) in the article; determine if the question(s) is (are) fully answered in the article. The template is: “the goal of the study is <subject>”.

**CLARIFYING CONFUSING DISTINCTIONS BETWEEN CATEGORIES**

1. **Incomplete evidence vs. Probable understanding**

Incomplete evidence a positive or negative statement proposing a possible/feasible explanation for a phenomenon on the basis of limited evidence as a starting point for further investigation. The statement is more proposing options and not staking a claim as to which is correct. This includes epistemic cues such as could, potential, it is possible, might, attracted increasing interest, has not been widely studied. In the sentence, “Small changes in the resistance to aqueous humor drainage may also contribute to diurnal differences in IOP”, the knowledge goal is to find more evidence that small changes in the resistance to aqueous humor drainage also contribute to diurnal differences in IOP.

Probable understanding is a statement staking a claim to the most likely explanation, relationship, or phenomenon where it has a good chance of being the case or of coming about. This includes epistemic cues such as it is becoming increasingly clear, likely, unlikely, promising, plausible, mainly. In the sentence “It is becoming increasingly clear that many forms of glaucoma have a genetic component” the researchers stake a claim that probably many forms of glaucoma have a genetic component.

1. **Superficial relationship notes**

Superficial relationship is a statement about a connection, link, or association between at least 2 variables; connectedness between entities and/or interactions representing their relatedness or influence. This category specifically captures the relationship nature of many statements where ideally, we want to know the full underlying mechanism of the relationship. Other categories will interact with this category and qualify it. In the sentence: “we hypothesized that <il23r may AFFECT the development of breast cancer>”, the “may” should be categorized as incomplete evidence and the “affect” is categorized as superficial relationship. The knowledge goal put together is: need more evidence to determine the underlying mechanistic relationship between il23r and the development of breast cancer. We want to call attention to the relationships separately due to the knowledge goal it has of getting at the mechanism.

1. **Controversy vs. Alternative options**

Controversy is a statement of disagreement amongst researchers; a lack of consensus; at least two possible answers are presented as results from different researchers; usually in reference to previous results and stated when results disagree with each other; contradictions. There are at least two possible answers already presented as results from different researchers. This is usually in reference to previous works findings and stated when findings disagree with each other. This includes epistemic cues such as do not agree, have been variably, some but not all, some but not other studies, have been implicated, and but. In the sentence “Some but not all human studies have reported a positive association between IOP and blood pressure” there is disagreement on the association between IOP and blood pressure.

Alternative options is a statement of multiple (at least 2) choices or actions with no stated disagreements, including statements with an implied second option such as "whether". These options have not yet been studied and so there cannot be disagreements yet about which one is correct. The statement only introduces the options. This includes epistemic cues such as alternatively, cannot be ruled out, though, either, and whether. In the sentence “An effect of anesthesia in these very old mice cannot be ruled out” the researcher indicates an anesthesia affect is an alternative option for explaining some phenomenon in very old mice.

1. **Future work vs. Urgent call to action**

Future work is a statement of extensions, including next steps, directions, opportunities, approaches, or considerations of the described work that may be implemented at some future time point. Also, this includes a statement of suggestion or a proposal as to the next best course of action, especially one put forward by an authoritative body; advice telling someone what the best thing to do is. This includes epistemic cues such as next steps, additional research is needed, warrant additional discussion, advocate, encourage, guideline, should be, require, must be, and it is important that. It is future work that does not necessarily need to be dealt with right away, there is no clear sense of urgency. For example, in the sentence, “Thus, to avoid effects of anesthesia on IOP, all measurements should be made within a window of up to 12 minutes after anesthetic administration.” The sentence recommends future work of taking measurements up to 12 minutes after anesthetic administration, without urgency, but useful for further experimentation.

Urgent call to action is a statement calling for immediate attention including an action needed to be taken immediately or information that needs to be disseminated immediately; critical: being in or verging on a state of crisis or emergency; urgently needed; absolutely necessary. This includes epistemic cues such as call to act, critical, cause for concern, crucial, essential, high priority, imperative, needed, influential, and vital. The statement has a clear sense of urgency. For example, in the sentence, “As PMTCT Option B+ scales up with regimens that require life-long antiretroviral treatment (ART) adherence, the need for HIV-infected pregnant and breastfeeding women to disclose their status to their male partners becomes more critical in order to maintain adherence to ART.” Here *the need* and *more critical* indicate a sense of urgency that needs to be dealt with sooner rather than later.

1. **Future work vs. Future prediction**

Future work is an explicit statement of extensions, including next steps, directions, opportunities, approaches, or considerations, of the described work that may be implemented at some future time point. This includes epistemic cues such as additional studies, more information… is needed, needs to be assessed, future investigation/strategies/opportunities/considerations, and candidate. For example, in the sentence “Mutations in two of the assessed candidate genes (*Lepr* and *Tyr*) result in increased IOP.” *Candidate* indicates that more work needs to be done to determine if these 2 genes actually increase IOP.

Future prediction is a statement of extrapolation of given data into the future and/or from past observations; no reference to next steps or future work. This includes epistemic cues such as if real, if confirmed, expect, supposed to, will, ultimately, would, and implication. For example, the sentence “If real, this sporadic sex difference was not dependent on age, sometimes occurring in a group of B6 mice at a particular age and sometimes not occurring in a separate group of the same age.” This sentence indicates that if the phenomenon was real, then multiple things follow (a prediction). Notice no next steps are stated.

**Scope/Subject Information: (from Bioscope)**

The scope of a speculative element can be determined on the basis of syntax. Generally, the scope extends to the biggest unit possible, that is, **annotated scopes always have the maximal size**.

The scope of **verbs, auxiliaries, adjectives and adverbs** usually starts right with the keyword. In the case of verbal elements, i.e. verbs and auxiliaries, it ends at the end of the clause (if the verbal element is within a relative clause or a coordinated clause) or sentence, thus, all complements and adjuncts are included.

*The presence of urothelial thickening and mild dilatation of the left ureter (<suggest> that the patient may have continued vesicoureteral reflux).*

*The presence of urothelial thickening and mild dilatation of the left ureter suggest that the patient (<may> have continued vesicoureteral reflux).*

*These findings that (<may> be from an acute pneumonia) include minimal bronchiectasis as well.*

*These findings (<might> be chronic) and (<may> represent reactive airways disease).*

The scope of **attributive adjectives** generally extends to the following noun phrase, whereas the scope of **predicative adjectives** includes the whole sentence.

*This is a 3 month old patient who had (<possible> pyelonephritis) with elevated fever.*

*(The demonstration of hormone receptor proteins in cells from malignant effusions is <possible>).*

**Sentential adverbs** have scope over the entire sentence, while the scope of **other adverbs** usually ends at the end of the clause or sentence.

*(The chimaeric oncoprotein <probably> affects cell survival rather than cell growth).*

*Right upper lobe volume loss and (<probably> pneumonia).*

**Conjunctions** generally have scope over the syntactic unit whose members they coordinate.

*A small amount of adenopathy cannot be completely excluded although there are no other findings of (adenopathy <or> pleural effusion).*

*Findings consistent with (viral <vs> reactive) airways disease with subsegmental atelectasis in the right lower lobe.*

Determiners, adjectives or any modifiers – as they modify the following noun, that is, belong to the noun phrase – are also included in the scope of the conjunction.

*Small focal opacity in right lower lobe, which represents (a small focus of pneumonia <or> atelectasis).*

The main exception that changes the original scope of the keyword is the **passive voice**. The subject of the passive sentence was originally the object of the verb, that is, it would be within its scope. That is why the subject also must be marked within the scope of the verb or auxiliary:

*(A small amount of adenopathy <cannot be> completely <excluded>). Doctors (<cannot> completely <exclude> a small amount of adenopathy).*

In a similar manner, **relative pronouns** such as *which* also have to be included in the scope of the auxiliary in the case of passive voice:

*These findings further support our previous observations on the distinct regulation of expression of the human HLA-DQ class II subset, (which <may> be thus controlled at the posttranscriptional level by a CIITA-independent mechanism).*

Another example of scope change is the case of **raising verbs** (*seem, appear, be expected, be likely* etc.). They can show two different syntactic patterns:

*It seems that the treatment is successful.*

*The treatment seems to be successful.*

In the first case, the scope of *seems* starts right after the verb. If this were the case in the second pattern, *the treatment* would not be included in the scope, however, it should be as it is indicated by the first pattern. Thus, in the second sentence, the scope must be extended to the subject as well.

*It <seems> (that the treatment is successful).*

*(The treatment <seems> to be successful).*

When the subject is not expressed in the second one of two coordinated clauses, the scope of the raising predicate starts with the main verb:

*This phenomenon, which is independent of tumour necrosis factor, is associated with HIV replication, and (is thus <likely> to explain at least in part the perpetuation of HIV infection in monocytes).*

The scope of **conjunctions** extends to all the members of the coordination. That is, it usually extends to both left and right. Complex keywords such as *either ... or* have got one scope. For the sake of simplicity, the following sentences are annotated only for conjunctions:

*Symptoms may include (fever, cough <or> itches).*

*Mild perihilar bronchial wall thickening may represent (<either> viral infection <or> reactive airways disease).*