Semantic Analysis of Image-Based Learner Sentences (SAILS)

Annotation Guide

Levi King

Last updated:

November 16, 2017

Contents

1 Task Background

1.1 Overview

In order to best annotate the data, annotators should have a basic understanding of the task used to collect it. The task is a picture description task (PDT), implemented as an online survey. The PDT consists of 30 items. An *item* is one image and corresponding question. Each item is displayed on a single page of the online survey, and participants type a response into the provided field before clicking ahead to the next page. The task was conducted with default web browser settings, so spelling correction and grammar correction tools were available to participants.

The images used are simple digital drawings. No two images are related, and nothing appears in more than one image. Each image was chosen or created to depict a single event or action.

In order to focus attention on the main action, images contain very little background or other detail. Each question is intended to elicit a complete sentence capturing the main action in the image.

The data collected in the task will be used to analyze the differences in English native speaker (NS) and non-native speaker (NNS) language use. Specifically, this process will use language tools and NS responses to derive an "answer key" or "gold standard" (GS), which can be used to automatically evaluate the language and content of NNS responses.

1.2 Participants

1.2.1 Non-native speakers

NNS participants were recruited from intermediate and advanced level English as a Second Language (ESL) courses in the English Language Improvement Program at Indiana University. 141 NNS students completed the PDT. These participants all performed the task independently in a computer lab, with the researchers present. Responses from this group appear to be given in good faith.

1.2.2 Native speakers

Two different groups of NSs participated: "familiar" NSs and crowd-sourced NSs. All NSs performed the task remotely, without the researchers present.

1.2.2.1 Familiar NSs

40 "familiar" NS participants completed the full task. They were recruited among friends, family and acquaintances of the researchers. Responses from this group appear to be given in good faith.

1.2.2.2 Crowd-sourced NSs

Responses were also collected from roughly 330 different NSs through the online platform, Survey Monkey. The researchers purchased survey responses from the platform's pool of

users, who may win prizes or earn donations for charities in exchange for completing surveys. These participants all performed the task remotely, without the researchers present.

Crowd-sourced participants are less likely to complete a lengthy task, so the PDT was divided into four smaller tasks, and each crowd-sourced NS completed only one of these. Additionally, a sizable number of these participants completed only part of their task before abandoning it. The resulting data set is equivalent in size to roughly 100 completed familiar NS PDTs. Responses from the crowd-sourced group are of varying reliability; The majority are legitimate and in good faith, but some responses clearly are not. Some crowd-sourced NSs simply typed random characters in the response fields in order to move on to the next item and complete the task with minimal time and effort. Others responded with jokes, sarcasm or profanity.

1.3   Instructions

Before beginning the task, respondents read a short page of instructions including an example item and possible responses. The instructions are as follows:

> In this task, you will view a set of images. For each image, please write **one sentence** to answer the question provided with the image. It is important to answer with a **complete sentence**, not a word or phrase.

English native speakers (NSs) and non-native speakers (NNSs) complete slightly different versions of the task. The items are identical in both versions, but whereas NNSs provide one response to each question, in the NS version, respondents are asked to provide two responses to each question. They are given the following additional instructions:

> Then, you will be asked to write a second, *different* answer, which is also a complete sentence. This might involve rewording or reorganizing your first sentence. It does not need to be *completely* different; some words may be the same. If you cannot think of another way to answer the question, you may leave the second answer space empty, but any second responses you provide will be greatly appreciated.

1.4 Item Examples (Targeted and Untargeted)

The first half of the task consists of 15 **targeted** items, and the second half consists of 15 **untargeted** items. Targeted and untargeted items differ only in the question. All targeted items take the form of *What is X doing?*, where *X* varies but is specified in the question, always as the subject (or one of the subjects) of the main action in the image. For all untargeted items, the question is always the same: *What is happening?*.

For each image used in the task, a roughly equivalent number of targeted and untargeted responses were collected. Multiple versions of the task were administered; a given image is used in the targeted section for some versions, and in the untargeted section for other versions. In all versions, the targeted items precede the untargeted items. This ordering is intended to avoid the possibility that a participant encounters the question *What is happening?* consistently in the initial items, assumes that this question applies to the entire task, and responds to the later targeted items without reading the questions.

The terms *targeted* and *untargeted* are never used in the task, and participants are not explicitly informed of these differences. They are, however, provided with an example of each type immediately following the instructions, as seen in Figures 1 and 2 below.

Example 1
What is the man doing?
Your sentence: *The man is shouting.*
Your second sentence: *He is yelling.*
There is not a single correct response. Many responses may be possible. Other responses might be: *The man is yelling something.* *He is speaking loudly.*

Figure 1: An example *targeted* item, as presented in the task instructions. The "second sentence" portion is presented to native speakers only.

Example 2
What is happening?
Your sentence: *The nurse is giving a patient roses.*
Your second sentence: *A woman is getting flowers from a nurse.*
There is not a single correct response. Many responses may be possible. Other responses might be: *The nurse is giving a lady some red flowers.*   *A patient is receiving flowers from a nurse.*

Figure 2: An example *untargeted* item, as presented in the task instructions. The "second sentence" portion is presented to native speakers only.

2 Annotating Features

Each response is annotated according to five dimensions, or *features*. These features, explained below, are referred to as **grammaticality, interpretability, core event, verifiability** and **answerhood**. Annotations for each feature have only two possible values, *yes* or *no* (or *1* or *0*). The annotation for each response is thus an ordered list (i.e., a vector) of zeros and ones. For example, [1, 1, 1, 0, 1] would represent a response that was annotated *no* for verifiability and *yes* for all other features.

Some features are non-contextual; these features should be annotated without consideration of the PDT image or question (See Table 1). The annotation for these features should be the same for both targeted and untargeted versions of an item. Other features are contextual and must be annotated with consideration of the image and question; for these features, targeted and untargeted items must be handled separately.
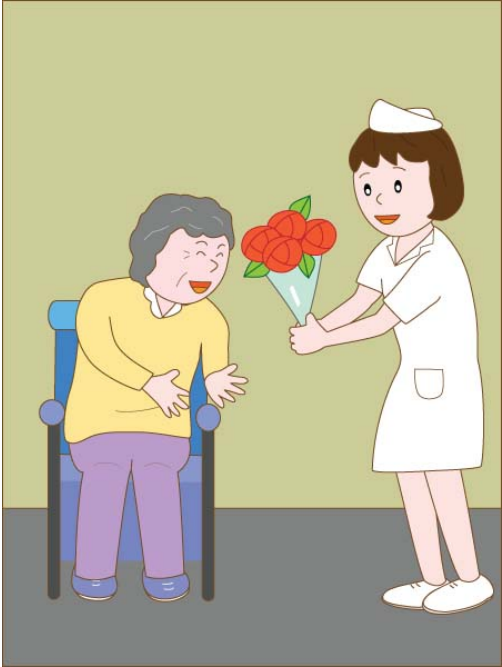
Feature	Contextual?	Targeted v. Untargeted Annotation
Grammaticality	no	identical
Interpretability	semi	may vary
Core Event	yes	may vary
Verifiability	yes	may vary
Answerhood	yes	may vary

Table 1: Contextuality of annotation features.

2.1 Grammaticality

The grammaticality feature primarily considers the following question: *Exactly as written, does the response convey a proposition and does it lack any grammar or spelling errors?*

2.1.1 Non-contextuality of grammaticality

This feature considers only the response, regardless of the item or question. In other words, a response that is grammatical but irrelevant given the specific item image and question should still be annotated as "yes" for this feature.

However, grammaticality should be annotated within the bounds of the very general context of the task; the PDT elicits descriptions of common events, so responses should convey a proposition and be grammatical when interpreted accordingly.

2.1.2 Defining *grammaticality*

For the current annotation purposes, a *grammatical* response is one that is free from grammar errors or misspellings, and conveys a reasonable meaning (given the very general context of the task). Grammar errors come in many forms, including omitted words, out-of-place words, incorrect word forms, and syntactic disagreement, among others. This feature does not directly consider *meaning*. However, the events depicted in the PDT images are all common, unsurprising events that might occur under normal circumstances, and a response that requires an unreasonable interpretation in order to be grammatical should be annotated "no" for grammaticality. For example, *The boy is dancing on music* is probably not grammatical without resorting to a fairly unusual interpretation – perhaps involving a boy dancing on a floor covered with sheet music or vinyl records.

Annotators will need to make judgment calls, but should be lenient in judging grammaticality and the necessary interpretation of meaning. If there is a reasonable reading of the sentence under which it is grammatical (and has none of the specific grammaticality problems outlined below), it should be annotated as "yes". (Annotators should keep in mind that concerns other than grammar are likely to be captured under the annotation of other features.) For example, consider this response to the item in Figure 3: *A boy listens to music and dancing.* Given the image, one could point out that the meaning conveyed by the response is not the intended meaning, and thus argue that the response is ungrammatical. However, because the response is not ungrammatical without the item context, and it conveys an arguably reasonable meaning, such a response should be annotated "yes".

2.1.3 Incomplete sentences

Although the task asks participants to provide a complete sentence, incomplete sentences (which are mostly verb phrases among the data) may nonetheless be annotated as "yes" for grammaticality, so long as the content of the response is indeed grammatical. For example, "eating pizza" is an incomplete sentence but a grammatical response. This also applies to any one word responses, but as explained in Section 2.1.5.1, a grammatical response should

be interpretable as a proposition. For example, "eating" should be considered a grammatical response, because it conveys some propositional meaning, but "pizza" is not grammatical here because it does not indicate any action or event. Incomplete sentences are subject to all of the same grammaticality considerations as complete sentences.

2.1.4 Punctuation and capitalization

Responses have been converted to all lowercase letters. Final punctuation has been removed from most responses. Annotators should ignore these concerns when annotating grammaticality. Any sentence internal punctuation, however, should be considered.

2.1.5 Common grammaticality problems

2.1.5.1 Non-propositional responses

A response that lacks a grammatical interpretation *as a proposition* should be annotated "no" for grammaticality. A proposition typically requires a verb and a subject; for the current task, a response may be judged as grammatical if it lacks a subject so long as it indicates an action or event. Non-propositional responses do not fit the general context of the task. These responses typically lack a verb and some appear to be well-formed noun phrases, such as *A boy with pizza*.

2.1.5.2 Bare nouns

A bare noun that is missing a determiner should result in a "no" for grammaticality. Examples include *Boy is eating pizza* and *A man is delivering package*.

2.1.5.3 Missing *be* verbs

Common among the data are responses that omit a necessary copula (or "be" verb). These often result in what could be interpreted as well-formed noun clauses, such as *A little boy eating pizza.* If, as in this case, one can reasonably assume that the apparent noun clause is an ungrammatical expression of a copular sentence (*A little boy **is** eating pizza*), the response should be annotated "no" for grammaticality.

Note that incomplete sentences that omit the subject may also omit a "be" verb. In other words, while *A little boy eating pizza* should be annotated "no" for grammaticality, simply *eating pizza* may be annotated as "yes" if appropriate. (See Section 2.1.3.)

2.1.5.4   Misspellings

Misspellings sometimes result in real but unintended words, so it is not always clear if a word is in fact a misspelling. A response containing a suspected real word misspelling should be annotated "no" for grammaticality only if it results in a grammar error.

2.1.6   Open questions

[This section should be removed in the final version of the guidelines.]

1. **Misspelled proper nouns.** For now, we're marking misspellings of proper nouns (e.g.,"lambergini") as "maybe".

2. **Activity/event noun phrases as responses.** The instructions clearly ask participants to respond using complete sentences. Nonetheless, many participants ignore this. We decided to accept responses that simply drop the subject provided in the question, as the subject is understood. Such responses are verb phrases, like "dancing" and "delivering a package". However, there are other reasonable and arguably grammatical responses that take the form of a noun or noun phrase. For example, if a participant is asked "What is the woman doing?", "origami" might be considered a reasonable and grammatical response (if we ignore the task instructions). "Origami" is of course a noun phrase. However, origami can be "done"; a person can "do origami". The untargeted items face a similar situation, where the prompt is "What is happening?" and noun phrases that can "happen" also seem acceptable. **Such activity/event noun phrases should be marked "maybe" for the time being.**

2.2   Native-likeness

NOTE: As of 8/7/2017, the Native-likeness feature has been suspended. The feature is under consideration, but it will likely be scrapped. Only a few items have been annotated for this feature as it has proven troublesome. The "no simple present verbs" rule was scrapped, because for some simple present responses,

it proved nearly impossible to judge whether the response should be interpreted as referring to a specific event or to a general fact. Compare: *The kid is enjoying pizza*, *The kid enjoys pizza*, *Kids enjoy pizza*. Moreover, such verb forms are in fact common among the native speaker responses, anyway. Upon removing this rule, native-likeness became a rather vague measure of awkwardness (distinct from grammaticality) or linguistic appropriateness/pragmatics, etc. Effectively, it was left to make (very difficult) decisions about the register of the response. Initially, only overly-complex, "highfalutin" responses were marked "no" for this revised feature (e.g., *The young man has become filled with the joy of music and he flails to its rhythm*; the register is too formal for the task. Subsequently, in consideration of fairness, "no" judgements were extended to also include responses in a register too informal for the task (e.g., *He's getting his groove on*). (Many of these responses contained what many speakers would consider slang.) Defining the appropriate register proved to be too problematic, and judging responses on the feature was too subjective to be applied consistently.

The native-likeness feature primarily considers the following question: *Exactly as written, is the response native-like?*

2.2.1   Non-contextuality of native-likeness

This feature considers only the response, regardless of the item or question. In other words, a response that is native-like but completely irrelevant given the context should still be annotated as "yes" for this feature.

2.2.2   Defining *native-likeness*

For annotation purposes, a response is considered native-like if a native speaker could produce the response exactly as written under reasonable circumstances. Because the feature is judged without regard for the context, a response is considered native-like if it does not internally contain any non-native-like characteristics. A "no" for native-likeness should be given when the annotator believes it would be very unlikely for a native speaker to produce the utterance under common, reasonable circumstances.

In general, grammaticality is a requirement for a native-like response. However, if a response is deemed ungrammatical in Standard English but seems to be grammatical in another

(native) dialect or variety, the response may still be annotated "yes"; annotators should exercise their best judgment in such cases.

2.2.3 Simple present verbs

Responses that use the simple present verb form are common among the data, e.g., *The boy dances with music on* and *The boy enjoys his pizza*. These sentences might be native-like under certain circumstances, such as the narration of a nature film, for example. However, for the current task, such responses should be annotated "no" for native-likeness.

2.2.4 Incomplete sentences

Incomplete sentences may be annotated as native-like, so long as they fulfill the criteria for this feature. For example, *A little boy eating pizza* contains no non-native-like characteristics, so it is considered native-like. Likewise, *Hungry* is annotated as native-like, although generally speaking it may not be a desirable response. *Him hungry*, however, is not native-like.

2.3 Interpretability

The interpretability feature primarily considers the following question: *Exactly as written, is the response interpretable enough to evoke a clear image?*

2.3.1 Semi-contextuality of interpretability

This feature is largely non-contextual, but because the task asks participants about events, responses must convey a proposition. In other words, a response must be interpretable as an event, or as a statement about the state of affairs in the image. Annotators may find it useful to view the PDT image, but interpretability should be judged without regard to its contents; to meet the criteria for this feature, a response should evoke *an image*, regardless of how similar that image is to *the image* in the PDT.

For targeted items only, when the subject of the response is omitted, it should generally be understood to be the same subject given in the targeted question. (This is not appropriate for *all* responses that lack a subject, and annotators should use their judgment to decide if

14

the respondent intended the subject to be understood.) For example, *eating pizza* should be annotated as interpretable (according to the criteria below) as a response to the targeted question, *What is the boy doing?*

In contrast, for the untargeted question (*What is happening?*), a response like *eating pizza* would not be interpretable, because a reader could not confidently conjure an image of the subject. (See Section 2.3.3.2 for more discussion of incomplete sentences.)

2.3.2 Defining *interpretability*

The interpretability feature is concerned with whether or not a response can be adequately understood and visualized. Because a response is based on an image, its interpretation should evoke a concrete image. A response should be considered interpretable if it A) includes any arguments that are syntactically required by the verb, and B) provides enough semantic content to derive a reasonably specific, unambiguous illustration.

2.3.2.1 Verb arguments

For this first requirement, *A man is delivering a package to a woman* is interpretable. *Delivering* is used as a ditransitive verb here, and all syntactically required arguments are specified; the sentence has a subject, direct object and indirect object. *The man is delivering a package* should also be considered interpretable. This sentence does not include an indirect object, but in this transitive use of *deliver*, the syntax does not require one. However, *A man is delivering* is not interpretable, because the verb *deliver* is missing one or more syntactically necessary arguments. This consideration requires a grammaticality judgment on the part of annotators. Annotators may have differing judgments with regard to the arguments required by given verbs; this is expected. Native speakers would likely agree that *The man is cooking* is grammatical as is (without an object), and that *The girl is telling* is not grammatical, because it requires an object (or more context). However, native speakers may disagree on the grammaticality of sentences like *The boy is washing* or *The woman is buying*.

2.3.2.2 Content and composition

Interpretable responses are statements that could be illustrated with a canonical composition, without the need to infer any critical elements. Responses that provide only a broad description are likely to fail this criterion. A sentence like "The man is working" is not specific enough to evoke a clear image. An illustrator could show a man picking fruit, building a bridge, typing at a computer, etc., so long as the image contained a man doing some kind of work. A significant amount of information concerning the action in the image would need to be inferred.

Likewise, a sentence that uses semantically empty references ("someone"/"something"/unspecified "it", etc.) for essential elements or simply leaves them out is not interpretable. Such a response could not be illustrated as a canonical, representational painting, because some essential elements would have to be guessed or inferred. The response could, however, be represented as an abstract painting.

It may be helpful for annotators to think of this as "The Norman Rockwell Rule." That is, "Would Norman Rockwell illustrate this response?" Straightforward composition and a clear representational style are hallmarks of Rockwell's paintings. A response like "The man is delivering a package to a woman" fits this style of illustration. "A man is delivering a package" also fulfills the Rockwell Rule, because a painting of a delivery man leaving a package in a mailbox or on a doorstep could easily be imagined as a Rockwell painting. (Annotators should keep in mind that interpretability annotation should not be influenced by the PDT image and the image evoked by the response is not judged here for how well it matches the actual PDT image.) For a response like "Someone is delivering things to a woman," a Rockwell painting simply would not fit; both the deliverer and the thing being delivered would have to be out of frame, obscured, somehow abstracted, or purely guessed at. Annotators should rely on their own judgment when considering these content and composition concerns.

2.3.3 Common interpretability concerns

2.3.3.1 Grammar and spelling

Grammar and spelling problems do not automatically result in a "no" here; these concerns are covered by the grammaticality feature. Major or multiple grammar or spelling problems are

likely to result in an uninterpretable sentence, but minor grammar or spelling problems may leave a sentence's interpretation intact. Annotators will vary in judging the severity of such problems, but in general, an annotator should mark a response as "yes" for interpretability only when he or she can be reasonably confident in the intended meaning. In other words, a grammar or spelling problem that could be corrected in multiple ways to result in multiple reasonable corrected sentences should be marked "no" for interpretability. As a reminder, for this feature, responses should be judged blindly, without influence from the image or previously seen responses.

For example, *The boy is danceing* contains a spelling error, but a reader can be quite confident that the intended meaning is *dancing*. *The boy is dacing*, however, would likely be judged uninterpretable, because without more context, the error has numerous plausible candidates for correction – *racing, pacing, daring*, etc.

Responses that contain contradictory information should generally be marked "no" for interpretability, but annotators should use their own discretion in handling these cases. Such problems often take the form of a noun phrase containing disagreement. For example, in *The man is giving the package to a women*, it is impossible to determine if the indirect object would be illustrated as one woman or multiple women. If an annotator feels confident that other information in the response disambiguates the intended meaning, the annotator may rate the response "yes" for interpretability. For example, in *A young girls feeds a tasty carrot to her pony*, the determiner, the verb form and the later singular pronoun all indicate that *girls* should be singular here.

Annotators should be lenient with subject-verb disagreement, unless they feel that such disagreement derails the interpretation of the response. For example, *The children is playing ball* is unambiguous, despite the error.

2.3.3.2   Incomplete sentences

Incomplete sentences should be annotated "yes" for interpretability, so long as they fulfill the requirements explained above.

In general, responses may rely on information understood from the question. This means that for targeted items, where the question is of the form *What is X doing?*, *X is* may be understood for responses like *washing the car* or *jogging*. For certain responses, like *the laundry* or *the foxtrot*, *X is doing* can be understood instead. In these cases, note that

the response must be an action or event that is commonly described as being *done*; *do the laundry* is common expression, while *do the baseball game* is not.

Untargeted responses may also rely on information understood from the question, *What is happening?* In these cases, *is happening* may be understood when appropriate. This means that noun phrases that can *happen* as events may be judged as interpretable, provided they otherwise fulfill the requirements of the feature. Therefore, *A fight between a cat and a dog* would probably be marked "yes" for interpretability, because it can *happen* and it contains adequate information about the event participants. However, *A fight*, which can also *happen*, would be marked "no", because it cannot be illustrated confidently without more information.

Also common among the data are noun phrases resulting from a sentence with an omitted copular verb (*be*), such as *A man delivering a package* (as opposed to *A man **is** delivering a package*). An omitted copula generally does not affect comprehension, so such a response should be annotated "yes" for interpretability, provided it meets the above requirements for this feature.

Other forms of incomplete sentences appear in the data. Annotators should use their best judgment for these, but keep in mind that it is difficult for incomplete sentences to satisfy the criteria, especially for untargeted items, where very little information can be understood from the question.

2.3.3.3 States and actions

The PDT is designed to elicit responses that describe an action; as a result, most responses contain an active verb. Some responses, however, describe a state of affairs in the image, such as "The boy is wearing a green shirt" or "The boy is ready to eat his pizza". Responses that describe a state are nonetheless interpretable, so long as they fulfill the remaining criteria.

2.3.3.4 Questions and modals

A small number of responses among the data take the form of a question. Some of these responses nonetheless present an assertion. For example, *Why is the baby crying?* indicates that *the baby is crying*. This response should be annotated "yes" for interpretability, because the assertion it contains meets the criteria for interpretability.

Some responses in the form of a question lack an assertion that can be judged for interpretability, e.g., *Do you think the boy likes pizza?* Such responses are not interpretable.

Responses that use modality may be considered interpretable if the modality does not effect information crucial to producing a visual representation. For example, in *The boy is eating so much pizza he may get fat*, it is stated as fact that a boy is eating pizza, so this could be visually represented. The modal part of this sentence contains unnecessary detail and could be ignored. In contrast, in *The man may be proposing marriage to the woman* the modality has scope over the whole predicate, so this response should be marked "no" for interpretability. (The man *may* be proposing marriage to the woman, but there is no limit to the number of things he *may* be doing.)

2.3.3.5 First and second person

All entities in the PDT items should be represented in the third person. Responses that use the first or second person to indicate a participant in the image should be considered uninterpretable. For example, *A young man will mail a package for you* should be marked "no".

2.3.3.6 Slang

Some responses contain what may be considered slang. Such responses are interpretable if they meet the other requirements for interpretability. For example, *The boy is getting his groove on* would probably be taken to mean that the boy is dancing intensely and could thus be considered interpretable. A response that contains unclear or unknown slang should be considered uninterpretable. Annotators must rely on their own judgment regarding slang.

2.3.3.7 Impossible or unknowable information

All PDT items consist of a single image. They present information in a straightforward manner and are almost completely devoid of any text, signs or symbols. Thus all responses should present information that can be learned from such an image. Responses that present important information (not details) that could not be known from or represented with a single image should be marked "no" for interpretability. For example, *He is sending a box*

to a woman could not be easily represented in a single image, as the man sending the box and the woman receiving the box would be in different locations. Moreover, the man and woman (and box) are arguably equally important arguments, so choosing whether to omit the subject or indirect object when illustrating the image would be problematic.

Responses that present an interpretable proposition but embellish it with unknowable details should be considered interpretable. (Note that concerns about unverifiable information are captured under the verifiability feature.) For example, *As the man hands the package to the woman, their eyes meet and a passionate romance ensues* presents a simple, illustratable event – a man handing a package to a woman, perhaps while making eye contact. The remaining details are unnecessary for assessing interpretability. Annotators must use their own judgment in such cases.

2.4 Core event

The core event feature primarily considers the following question: *Exactly as written, does the response capture the core event of the item?*

2.4.1 Contextuality of core event

Annotation for the core event feature is contextual; it must consider the image and question presented in the item.

2.4.2 Defining *core event*

Each image depicts a single *core event* that could be captured by a simple sentence or verb phrase. Each core event involves an action; responses that merely describe a state or feature of the image do not capture the core event. Considering Figure 3, for example, the response *He is a dancing machine* does not capture the core event; it describes a characteristic of the boy, but does nothing to describe what is actually taking place in the image.

The core event is best fulfilled with a present progressive verb form, but responses that use other verb forms may be acceptable. Crucially, the response should allow for an interpretation in which the verb refers to the specific event displayed in the image. For example, in most contexts, *He enjoys dancing to music* would be interpreted to mean that *in general,*

the subject enjoys the activity of dancing to music. However, in this context, it could refer to the event displayed in the image; the sentence could be intended as a narration of the image. Likewise, responses that describe the event in the past tense should be accepted.

For targeted items, the subject provided in the question must be the subject (or agent; see Sec 2.4.8) of the response. Beyond this, the core events are not predefined; annotators should decide what each core event is and whether or not a response captures it. Moreover, a core event should be conceived of abstractly rather than as a particular phrase or expression. Two responses that convey the same concept in different forms should be judged as equally acceptable. For example, *The man is shouting* and *He is yelling*, as seen in Figure 1, convey the same core event using different words.

Given the simplicity of the images, the core event should be clear for each. None of the images depicts any background events that are unrelated to the core event. Any non-core event that could be described either supports the core event or is an effect of the core event. In Figure 2, for example, the untargeted question (*What is happening?*) could be answered with *The patient is smiling*, but this is clearly an effect of the core event, in which a nurse is giving the patient flowers. Thus, *The patient is smiling* should be annotated "no" here.

2.4.3 Alternative interpretations

Although every effort was made to produce unambiguous PDT images, reasonable alternative interpretations are seen among the responses for a very small number of items. For example, Figure 6 shows a woman seated behind a desk and a man holding a package in front of the desk. Most participants interpret the scene as the man delivering a package to the woman. However, a small number of participants interpret this scene as a man picking up a package from the woman – a reasonable alternative. Such reasonable alternatives should be annotated "yes" for core event. Annotators should use their own in judgement in annotating responses that contain variations in interpretation.

2.4.4 Language problems

Grammatical and spelling problems do not automatically result in a "no" for the core event feature. Responses with errors that do not obscure the core event may still be annotated as "yes." In other words, if, despite a language problem, the necessary elements of the core event are intact and their relationship is reasonably interpretable, the response is annotated

21

"yes." Such cases are typically very minor errors. For Figure 7, for example, the response *He's eating a **peice** of pizza* should be annotated "yes", because the core *boy eating pizza* event remains intact and interpretable, despite the misspelling.

However, *He's **eatting** a piece of pizza* should be annotated "no", because a misspelling directly obscures the core event; one would not be able to find *eating* or some equivalent in the response.

2.4.5   Slang

Responses that describe the event using slang should be annotated as "yes" for the core event if the language used can readily be understood as equivalent to a more canonical description of the core event. For example, Fig 3 depicts a boy dancing. The responses *The boy is **getting down*** and *He is **grooving*** could be understood to mean *dancing* by most annotators, so they should be annotated as "yes" for core event. The response *He's **going bananas*** however, cannot be easily understood as equivalent to *dancing*, so it should be annotated as "no" for core event. Annotators will need to use their own judgement in handling slang responses.

| Targeted (I01T): What is the boy doing? |
| Untargeted (I01U): What is happening? |

Figure 3: Item 1, for which the core event is roughly *boy dancing*.

2.4.6 Intransitive vs. transitive core events

The PDT was created using a variety of images intended to cover intransitive, transitive and ditransitive events in equal numbers. These categories are not given for each item; if it becomes necessary to explicitly determine the category for a core event, annotators should use their own judgement. In general, an intransitive event is described without an object, a transitive event is described with a direct object, and a ditransitive event is described with a direct object and an indirect object.

2.4.6.1 Intransitive core events

For intransitive events, the response should link the subject and the verb of the core event.

2.4.6.2 Transitive core events

Predicates. For transitive events (including ditransitives), the response should link the subject with the verb and direct object (i.e., the *predicate*) of the core event. Where appropriate, indirect objects are desirable but not not required for the fulfillment of this feature.

A direct object may be omitted when it is sufficiently indicated through either the subject or the verb. For example, consider the image in Figure 4 and the corresponding questions for the targeted and untargeted items. Here the core event predicate could be described as *asking a question*, or some equivalent, e.g., *posing a query* or even simply *questioning* (without an object). While *questioning* alone is acceptable here, *asking* alone is not an acceptable equivalent for *asking a question*, because it is not comparably precise. *Questioning* can be seen as meaningfully equivalent to *asking a question*, but simply *asking* leaves the object ambiguous; one can ask many things besides questions, such as *for help* or *for money*.

As another example, in response to a targeted item *What is the professor doing?*, both *She is lecturing* and *She is teaching a lesson* are acceptable. Similarly, for an untargeted item *What is happening?*, *The cyclist is riding* and *The man is riding a bike* both satisfy the core event feature. In this case, the subject sufficiently indicates the bicycle.

Omitted subjects. For the targeted version, a response may omit the subject, because the subject is included in the question and may thus be understood to be the subject of the response. Such cases most often involve only a verb phrase, e.g., "asking a question"

or "asking the man a question". For the untargeted version, a response must indicate the subject of the core event, because it is not included in the question and thus cannot automatically be understood to be the subject of the response.

2.4.7 Pronouns

Pronouns as subjects are acceptable in responses to both targeted and untargeted items. A pronoun that clearly assigns the wrong gender to a subject or object should result in a "no" for the core event feature. Otherwise, annotators should retain a high degree of flexibility with regard to pronouns. The item in Figure 4, for example, depicts an *ask* action involving two males, one as the subject and the other as an object. The pronoun "he" could thus lead to ambiguity, but nonetheless the response "He is asking him a question" should be annotated as "yes". In other words, with regard to pronouns, ambiguity is acceptable, but inaccuracy is not.

2.4.8 Targeted items and passive responses

In targeted items, a subject is provided in the question. For example, the targeted item in Figure 4 asks *What is **the boy** doing?* This provided subject will be the subject of most responses. However, this is not a hard requirement for annotating a targeted response as "yes" for the core event. The crucial requirement is that the provided subject be indicated as the agent of the core event predicate, even if it is not expressed as the syntactic subject in the response. For example, a passivized response may move this subject to a "by" phrase, as in *The man is being asked a question by a boy.* Because the provided subject *(the) boy* can be understood as the agent of the core event, this response should be annotated as "yes" here. Omitting this "by" phrase (i.e., *The man is being asked a question*) would result in a "no" annotation, however, because the provided subject is lost. Likewise, a response that reframes the event like *The man is listening to a boy's question*, is annotated "no", because *boy* is not expressed as the agent of the core event.

2.4.9 Untargeted item leniency

In general, with regard to the core event feature, a greater variety of responses may be annotated as "yes" under the untargeted version of an item than under the targeted version,

because the untargeted question is less specific than the targeted question. This may include passivizations, such as *A man is being asked a question.* Likewise, responses that simply cast the core event from a different angle may be appropriate and may be annotated as "yes" for an untargeted item. For example, *The man is listening to the boy's question* would be annotated as "yes" for the untargeted version of this item. See Tables **??** and **??** for more examples of annotated responses for the targeted and untargeted versions of this item.

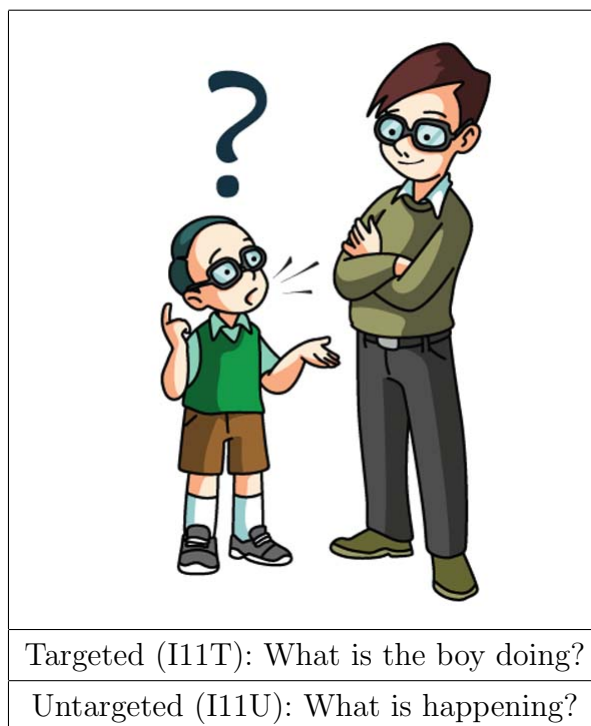| Targeted (I11T): What is the boy doing? |
| Untargeted (I11U): What is happening? |

Figure 4: Item 11, for which the core event is roughly *boy asking question.*

2.5 Verifiability

The verifiability feature primarily considers the following question: *Exactly as written, is all information in the response verifiable (or reasonably inferred) based on the image?*

This feature is mainly concerned with identifying inaccurate information and unverifiable inferences.

2.5.1 Contextuality of verifiability

Annotation for the verifiability feature is contextual; it must consider the image presented in the item.

2.5.2 Reasonable inferences

Responses that contain reasonable inferences should be considered verifiable. For this feature, an inference that can be assumed to be true for an overwhelming majority of situations like the one depicted in the image should be taken as "reasonable". Inferences that posit a degree of information that cannot safely be assumed (i.e., a *guess*) should not be considered reasonable and should be annotated "no" for verifiability. For example, the image in Figure 5 depicts a boy carrying a bag of groceries alone. The first example infers that the destination for the boy and his groceries is "home". This is taken as a reasonable inference because a person carrying a bag of groceries is almost certainly taking the groceries home. The second example describes the boy's action as "helping carry" the groceries. This is also taken as a reasonable inference, because the small boy is very unlikely to be doing his own grocery shopping. The third example states that the boy is "helping his mother" carry the groceries. Annotators should give this a "no" for verifiability because the inference posits an unnecessary and unknowable level of detail; "mother" is a fair guess here, but it is indeed a guess. Annotators must use their own best judgment in distinguishing between reasonable inferences and reasonable guesses.

2.5.3 Language problems

Responses that are unintelligible should be annotated "no" for verifiability; if the information in the response cannot be clearly understood, then it cannot be verified. However, grammar

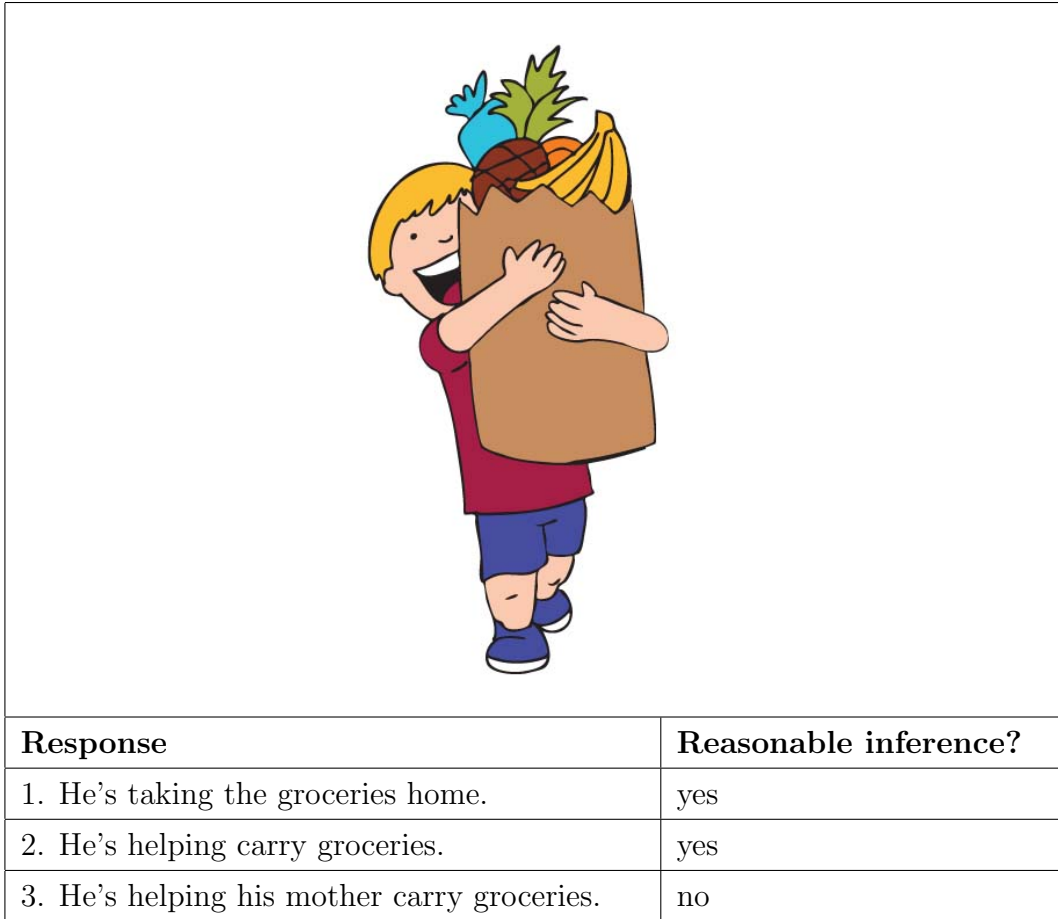| Response	Reasonable inference?
1. He's taking the groceries home.	yes
2. He's helping carry groceries.	yes
3. He's helping his mother carry groceries.	no

Figure 5: Example inference judgments for item 6, targeted: *What is the boy doing?*

and spelling problems do not automatically result in a "no" for verifiability. Responses that contain errors but remain reasonably clear and interpretable should be judged for verifiability like any other response.

Many responses incorrectly describe a singular subject as plural or vice versa. In cases where the subject's number is clearly incorrect or too ambiguous to discern, the response should be annotated "no" for verifiability. Some responses may indicate an incorrect number but still contain enough evidence that the correct number is intended, as in "The two little kid are playing." Given the "two" and "are", this response should annotated "yes", despite the fact that "kid" should be "kids". Annotators should use their best judgment in such cases. With regard to objects, annotators should use their best judgment to determine if similar changes in number are reasonable inferences. For example, a hunter shown shooting a single bird might nonetheless reasonably be described as "hunting birds" or "fowl", but a salesman shown handing car keys to a lone female customer would not be reasonably described as

"selling a car (or selling cars) to women".

2.5.4 Incomplete responses

Responses that do not present a complete proposition should be annotated "no" for verifia-
bility. For example, untargeted responses that contain only a verb or verb phrase should be
annotated "no" for verifiability because they cannot be verified if the subject of the verb is
unknown.

2.5.5 Alternative interpretations

Although every effort was made to produce unambiguous PDT images, reasonable alternative
interpretations are seen among the responses for a very small number of items. For example,
Figure 6 shows a woman seated behind a desk and a uniformed man standing across from her
holding a package. Most participants interpret the scene as the man delivering a package to
the woman. However, a small number of participants interpret this scene as a man picking
up a package from the woman – a reasonable alternative. Such reasonable alternatives
should be annotated "yes" for verifiability. Annotators should use their own in judgement
in annotating responses that contain variations in interpretation.

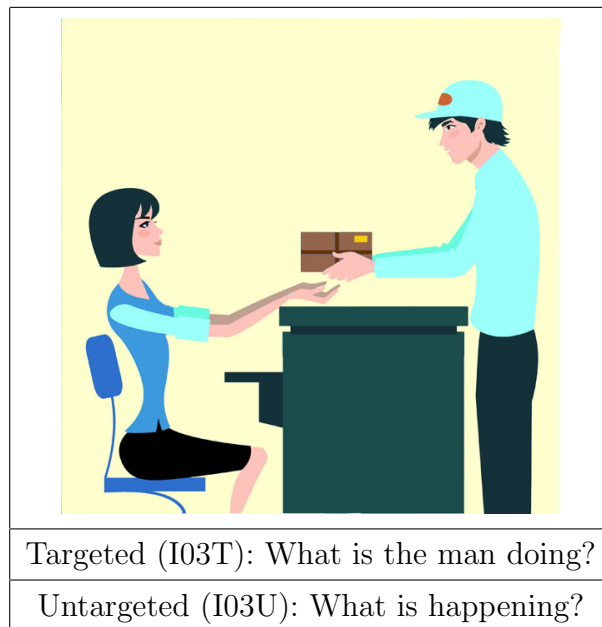| Targeted (I03T): What is the man doing? |
| Untargeted (I03U): What is happening? |

Figure 6: Item 3, in the targeted and untargeted versions.

2.5.6 Responses in the form of a question

A small number of responses among the data take the form of a question. In general, such responses are not considered verifiable; the content of the question is not an assertion of facts and cannot be compared against the facts of the image.

2.5.7 Modality

Modality in a response can impact the verifiability. For annotation purposes, a sentence is *modal* if it conveys the speaker's belief about the possibility of that sentence, using a modal verb (*may, should,* etc.), or a modal adverb (*maybe, perhaps,* etc.). (This is known as epistemic modality, because it involves the speaker's belief about the facts of the world.)

In a response where modality allows for doubt about the facts, the modal portions should be ignored, and the remainder of the response should be annotated for verifiability. For example, *The man is smiling as he hands the woman a package, maybe he likes her* would still be annotated "yes" for verifiability, because removing the modal portion (*maybe he likes her*) leaves a verifiable statement based on the image (*The man is smiling as he hands the woman a package*).

If, after removing the modal portions, a response is not verifiable, it should be annotated as "no" for this feature. For example, in *Perhaps the boy is asking a question*, the modal adverb has scope over the entire sentence, so removing the modal portion would leave no verifiable information.

2.5.8 Unverifiable inferences

Responses containing unverifiable inferences are common among the data. Unverifiable inferences that embellish a response with unnecessary detail should result in a "no" annotation for the response. For example, consider the item in Figure 7, which shows a boy eating a slice of pizza. Some responses to this item refer to the pizza as "sausage", "pepperoni" or "cheese" pizza, and the image is ambiguous enough that one might argue for any of these descriptions. However, as these inferences cannot be confidently verified and they merely contribute detail, they should be annotated "no" for verifiability.

Similarly, some creative responses assign names or other unknowable descriptors to persons in the PDT images. Such responses should be annotated "no" for verifiability.

Some unverifiable inferences are arguably unavoidable based on the PDT item. For example, Figure 4 depicts a male child speaking to a male adult. Few participants could be expected to describe these figures as "a male child" and "a male adult" or something similarly unnatural. Instead, the image lends itself to reasonable inferences that describe the figures based on a relationship: a father and son, a big brother and little brother, or a student and teacher would all be reasonable and practically unavoidable inferences.

Responses may contain other "creative" inferences, like "He is asking the man where babies come from" (Figure 4). This information is not verifiable, so the response is annotated "no" for this feature.

2.5.8.1 Participant opinions

For annotation purposes, unverifiable information also includes statements that seem to derive only from the opinion of the participant, and not from the content of the image. To illustrate, consider Figure 7, which depicts a boy eating a slice of pizza. In the first example response, *He's eating a slice of delicious pizza*, the word "delicious" is an expression of opinion, but based on the pleased expression on the boy's face, we can consider this verifiable and not solely dependent on the participant's opinion.

In the second example response, *He's eating pizza, yuck*, the word "yuck" can only be explained as the respondent's judgement about pizza, because there is nothing in the image to indicate that the pizza is "yucky" or undesirable.

2.5.9 Irrelevant information

A less common problem to be considered under this feature is the presentation of irrelevant information. A response should be annotated "no" for verifiability if it contains mostly irrelevant information, given the item. In Figure 7, the third response, *He will get fat eating pizza*, should be annotated "no" because the event described is not relevant based on the PDT image and question.
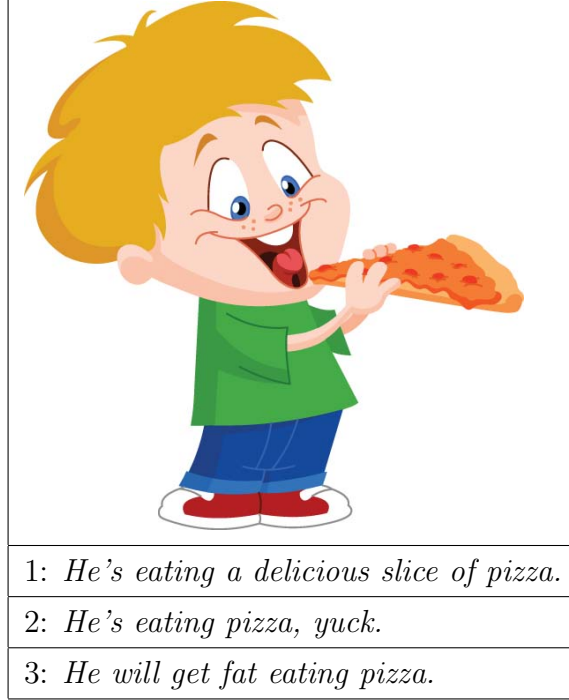
| 1: *He's eating a delicious slice of pizza.* |
| 2: *He's eating pizza, yuck.* |
| 3: *He will get fat eating pizza.* |

Figure 7: Item 2 (targeted: *What is the boy doing?*) and example responses.

2.6 Answerhood

The answerhood feature primarily considers the following question: *Exactly as written, does the response make an attempt to answer the specific question asked?*

2.6.1 Contextuality of answerhood

Annotation for the answerhood feature is contextual; it must consider the question presented in the item. The image is mostly irrelevant and is only used for targeted items to confirm that when a response replaces the subject with a pronoun, an appropriate pronoun is used.

2.6.2 Defining *answerhood*

As noted above, responses should address the specific question in the prompt. In other words, the response must answer the exact question given; merely answering a *similar* or *related* question is not adequate. Responses should make a positive assertion; responses that merely point out a negative fact are not acceptable (e.g., *The boy is not wearing a helmet.*) Figure 8 presents a number of example responses and answerhood annotations.

2.6.3 Accuracy

Answerhood should be annotated without regard to the accuracy of the response. Consider Figure 7 for example. The targeted version asks *What is the boy doing?*; the response *He's eating a sandwich* should be annotated "yes" because it does attempt to answer the question, even though the boy is clearly eating pizza. Moreover, *The boy is riding a bicycle* would also be annotated "yes", despite the fact that no bicycle appears. The accuracy of the response is accounted for with the core event and verifiability features.

2.6.4 Targeted vs. untargeted items

The answerhood feature, like *core event*, is dependent on the differences in the targeted and untargeted versions of the items. In other words, a sentence that may receive a "no" annotation as a targeted response could receive a "yes" annotation as an untargeted response. (The opposite should not be possible, as the targeted version of an item always asks a more specific question than its untargeted counterpart.) For example, consider Figure 6 and the targeted and untargeted questions: *What is the man doing?* and *What is happening?* The response *The man is delivering a package* would be annotated "yes" for answerhood for either version, while *The woman is receiving a package* would be annotated "yes" only for the untargeted version.

2.6.5 Common answerhood concerns

2.6.5.1 Verb forms

The PDT items ask what *is happening* or what a particular figure in the image *is doing*. Therefore, responses should contain a dynamic verb to describe the action in the image. (Some acceptable responses may contain only noun phrases; see Section 2.6.5.2.) This is a key consideration. A significant number of responses merely describe the state of affairs depicted in the image. Such responses leave the action of the item for the reader to infer and do not directly answer the question, so they receive a "no" annotation for this feature. For example, "The nurse is happy," shown in Figure 8, should receive a "no" annotation (for both the targeted and untargeted versions) because it describes a state depicted in the image but does not directly answer the question of what the nurse is *doing*. Likewise, "The

32

boy loves pizza," a response to Item 2 (Figure 7) is annotated "no" for answerhood, because it does not directly answer the question.

Dynamic verbs are appropriate for responses because they describe an event or action that happens and typically has a beginning and end. Dynamic verbs often take the progressive form (*is eating*, *was dancing*), and the majority of responses use progressive forms. Stative verbs are inappropriate for this task as they describe a state or condition. Stative verbs cannot be used in the progressive form (with rare and arguably non-stative exceptions). Roughly speaking, stative verbs can be categorized as verbs of cognition (*Susan **knows** karate*; *Sabrina **believes** in the team*) and verbs of relation (*Josh **resembles** his father*).

Sentences that use only a copula and a complement do not satisfy the answerhood requirements; *The girl is tall* and *The nurse seems happy* are examples. However, responses that use verbs like *seem* or *appear* followed by an action are acceptable, e.g., *The boy appears to be dancing* and *He seems to be waking up*. Responses that hedge with some form of modality should also be accepted, e.g., *The boy may be dancing* and *He is probably waking up*.

Although most responses use a present progressive verb (e.g., "He *is eating* pizza"), responses using the simple present form of a verb ("He eats pizza") are also common among the data. This form is commonly used to describe general truths or habitual actions, like *The horse eats grass* or *The river flows east*. Responses that use the simple present should be annotated "no" for answerhood. In most situations, in English the simple present would not be used to describe the actions in the PDT items, and particularly not in response to the present progressive questions in the PDT.

Responses that omit a "be" verb but include a progressive form verb (e.g., *The boy holding a sandwich*) should generally be annotated "yes" for answerhood.

For handling misspelled verbs, see Section 2.6.5.4

2.6.5.2   Events and activities

In some cases, a noun phrase may be an adequate and natural response to the PDT questions. For targeted items (*What is the X doing?*), a response in the form of a noun or noun phrase that can be *done* should be accepted. For example, *gymnastics*, *origami* and *the laundry* are acceptable in response to *What is the woman doing?*. Likewise, for untargeted items, a response in the form of a noun or noun phrase that can *happen* should be accepted. For

example, *an interview, a volleyball game* and *a math class* are acceptable responses to *What is happening?*.

For targeted and untargeted items, such event and activity responses should be properly formed as a grammatical response to the question. Grammar is not strictly considered for answerhood, but because these responses tend to be very short, proper form is used to ensure that the response was intended as a grammatical response to the question and not simply a sign of low effort. For example, *a baseball game* should be accepted in response to the question *What is happening?*, but *baseball* and *baseball game* should not.

2.6.5.3   Targeted subject variations and pronouns

All targeted questions take the form of *What is the X doing?*. Responses should use the same subject provided in the question, or an appropriate pronoun. This subject should be in the subject position of the response; if the response contains only a predicate, the subject of the question should be understood as the subject of the response. Responses should not alter the subject in any way, or move it from the subject position (as in passivization). This is in keeping with the requirement to answer the question exactly as it is asked. Several relevant examples are presented in Figure 8.

To put this concisely, responses to targeted items must either repeat the subject exactly as presented in the question, or use an appropriate pronoun, or drop the subject so that it is understood from the question. To clarify, the subject should not be altered in terms of definiteness, number, specificity, role or any other characteristic. Such responses add context to the question, and in order to evaluate answerhood, this new information would need to verified to ensure that the subject presented in the response is indeed the subject provided in the question. Verifying information for the sake of answerhood adds noise and complication, so verifiability is left to its own feature. For answerhood purposes, *a nurse* is not the same as *the nurse*. Likewise, neither *nurse, the young nurse, the nurse who is standing*, or *this nurse* is the same as *the nurse*.

Regarding pronouns, all humans presented in the PDT images are clearly male or female, and any targeted response that replaces the subject with a pronoun should use a pronoun that matches the subject's gender. One exception may be for babies portrayed in the PDT; the gender is not evident, and any third person singular pronoun is acceptable. For many items, the gender of the subject is clear from the question (*What is the man/woman/boy/girl*

doing?). Some items present a human subject in non-gendered terms, however, such as *the nurse*, *the teacher* and *the doctor.* In these cases, annotators should check the image to ensure that appropriate gender pronouns are used. Pronouns should also match the subject in number, and all subjects in the PDT are singular. When a response presents a subject with a non-matching pronoun, annotators should mark this as "no" for answerhood, because it is not possible to know if the response was indeed an attempt to answer the question asked. Some PDT items present an animal as the subject; gender is not indicated in these items, so any singular subject pronoun is acceptable.

2.6.5.4 Misspellings

The answerhood feature addresses whether or not a response *makes an attempt* to answer the PDT question, so misspellings do not automatically result in a "no" annotation.

Annotators should be strict in handling misspelled subjects for targeted items. The subject is provided on screen for the participant, so misspellings should be avoidable. Only misspellings that are very clearly typos should be accepted here, such as *t.he girl.* Misspellings that change the subject or leave it ambiguous in any way should be rejected. Pronouns must be properly spelled, but pronoun contractions that simply omit or misuse an apostrophe (e.g., *Its* for *It is*) should be accepted.

Verbs, even when misspelled, should appear to have the appropriate form (i.e., progressive). Annotators should be lenient with regard to misspelled verbs when a response appears to attempt to answer the question, even if the intended verb is not obvious. For example, *The boy is steeaching his arms in bed* should be accepted, despite the badly misspelled attempt at *stretching.*

When other elements of a response are misspelled, annotators should be lenient. The key consideration should be whether or not the response attempts to answer the question.
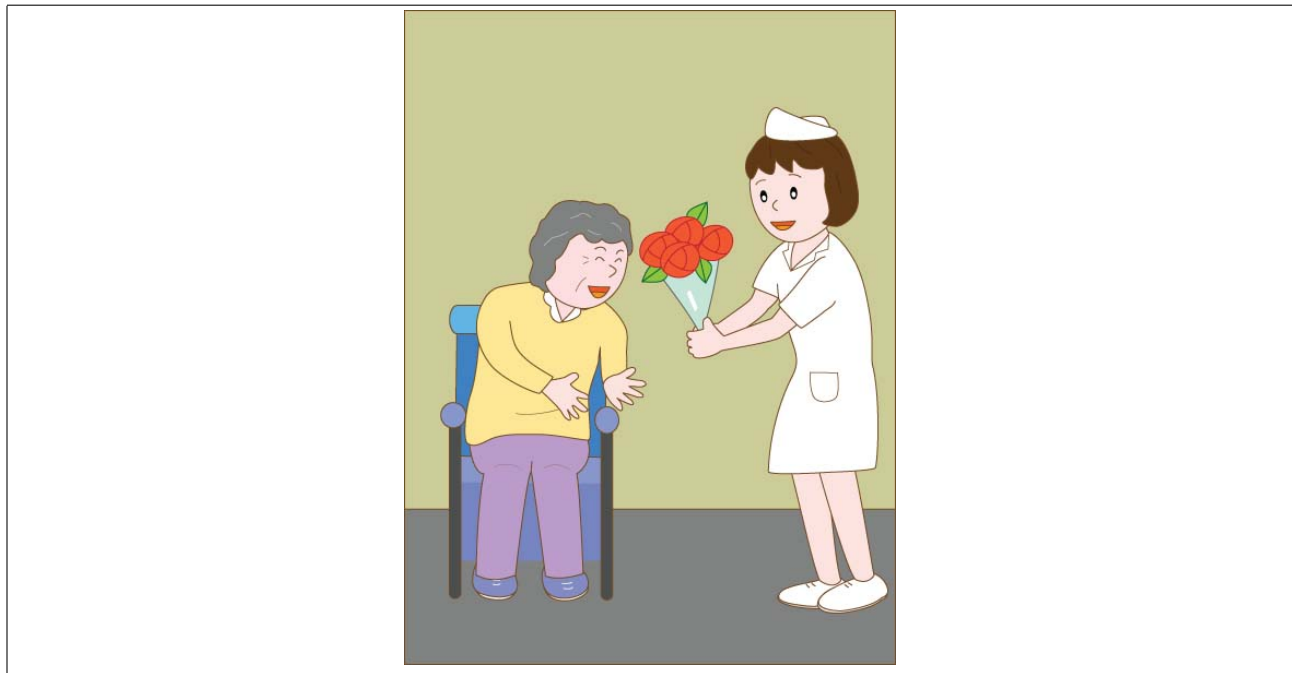
	Response	An.	Appropriate question
1	Giving a patient flowers.	yes	(prompt)
2	She's giving flowers to a patient.	yes	(prompt)
3	The nurse is giving away flowers.	yes	(prompt)
4	A nurse is giving away flowers.	no	What is happening?
5	A young nurse is giving away flowers.	no	What is happening?
6	The woman is giving the patient flowers.	no	What is the woman doing?
7	The nurse is happy.	no	How is the nurse?
8	The nurse is smiling.	yes	(prompt)
9	The nurse gives flowers away.	no	What does the nurse do?
10	The nurse gave the patient roses.	no	What did the nurse do?
11	The young nurse is giving out flowers.	no	What is the young nurse doing?
12	The smiling nurse is giving away roses.	no	What is the smiling nurse doing?
13	This nurse is giving away flowers.	no	What is this nurse doing?
14	That nurse is giving her patient flowers.	no	What is that nurse doing?
15	Nurse is giving away flowers.	no	What is Nurse doing?
16	The patient is receiving roses from the nurse.	no	What is the patient doing?
17	The nurse is about to give the woman roses.	no	What is the nurse about to do?

Figure 8: Example responses to targeted Item 2 (*What is the nurse doing?*) and their answerhood annotations ("An."). A particular response could be appropriate for multiple questions, but a likely example is given for each.

2.7 Appendix: Annotated examples

I01T: What is the boy doing?	I02T: What is the boy doing?
I03T: What is the man doing?	I11T: What is the boy doing?

Figure 9: Example items used in Table **??** and Table **??**. The question for all untargeted items is *What is happening?*