Annotation Guidelines

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.

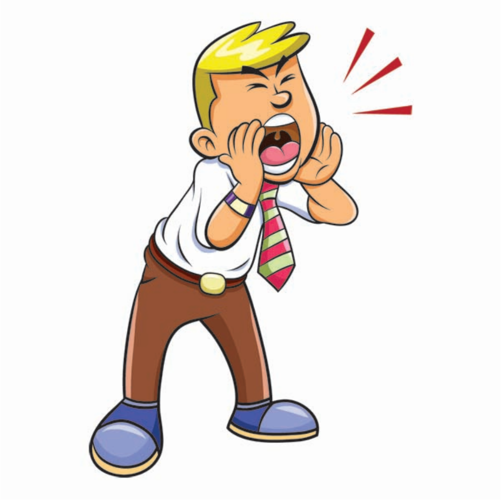
1.4 Item Examples (Targeted and Untargeted)

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.

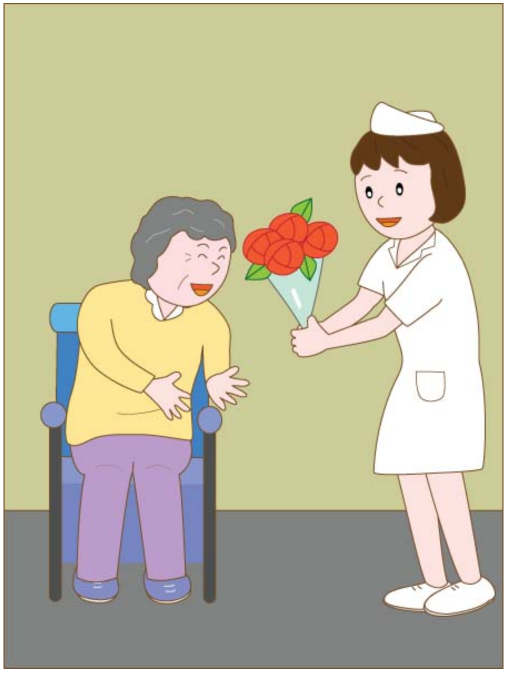
Example 2

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 six dimensions, or *features*. These features, explained below, are referred to as ***grammaticality, native-likeness, 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, 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 can 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
Native-likeness	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

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, responses must be interpretable *as events.*

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 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 would not know the subject. In other words, annotation of interpretability for untargeted items should not depend on any information in the question or image.

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 a person reading it could illustrate its meaning with a drawing, without the need to guess at any arguments required for the sentence. For example, *A man is delivering a package to a woman* is interpretable; a reasonable reader could read the sentence and draw a corresponding image (See Figure 5). *The man is delivering a package* should also be considered interpretable; a reader could draw a man leaving a package in mailbox or on a doorstep. However, *A man is delivering* is not interpretable, because in order to draw this, a reader would have to guess about the identity of the object being delivered. In linguistic terms, this generally means that any necessary arguments must be specified.

This means that responses that lack a clear semantic agent of the event are not interpretable, because there is no clear way to illustrate them without inferring the agent. Note that this does not include targeted item responses where an omitted agent can be understood as the subject of the question (see Section 2.3.1). This does include any other forms of incomplete sentences that lack a clear or understood agent (see Section 2.3.3) as well as passive sentences that do not include the agent in a "by" phrase. For example, *A package is being delivered* is not interpretable, whereas *A package is being delivered by a man* is interpretable.

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. As a rule, a noun phrase that omits the head noun (i.e., a determiner without a noun) such as *The is asking a question* should result in a "no" for interpretability. Beyond this specific rule, annotators should use their own best judgment in annotating interpretability.

2.3.3 Incomplete sentences

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

Incomplete sentences among the responses commonly take the form of a verb phrase. As noted in Section 2.3.1, for targeted items, when appropriate, the subject of a response that lacks an explicit subject may be understood to be the subject provided in the question. For untargeted items, however, it is uncommon for verb phrases alone to satisfy the interpretability criteria, because they typically omit the agent. For example, *Delivering a package* would be annotated "no" for interpretability.

Another common type of incomplete sentence in the data is a noun phrase. Many of these take the form of a sentence that is missing a copular verb (*be*), such as *A man delivering a package* (as opposed to *A man* ***is*** *delivering a package*). This response should be annotated "yes" for interpretability, because 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.

2.3.4 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 interpretable; the content of the question is not an assertion and thus the response could not be illustrated without some level of abstraction or symbolic representation.

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 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 equally convey the same concept in different forms should be judged as equal. 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 5 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.

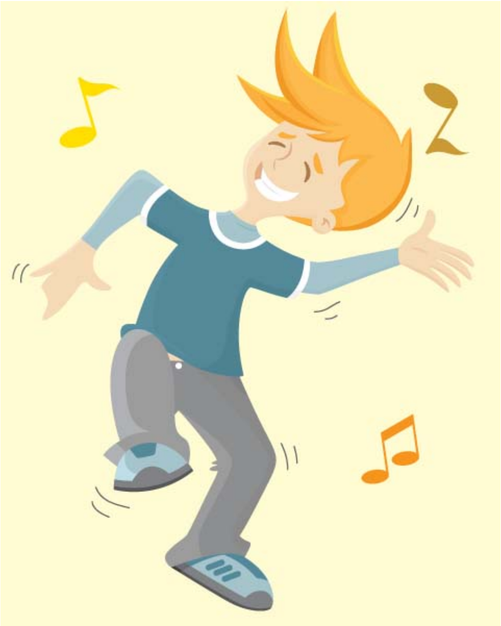
2.4.4 Language problems

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 “yes.” Such cases are typically very minor errors. For Figure 6, 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. 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.* 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.*

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

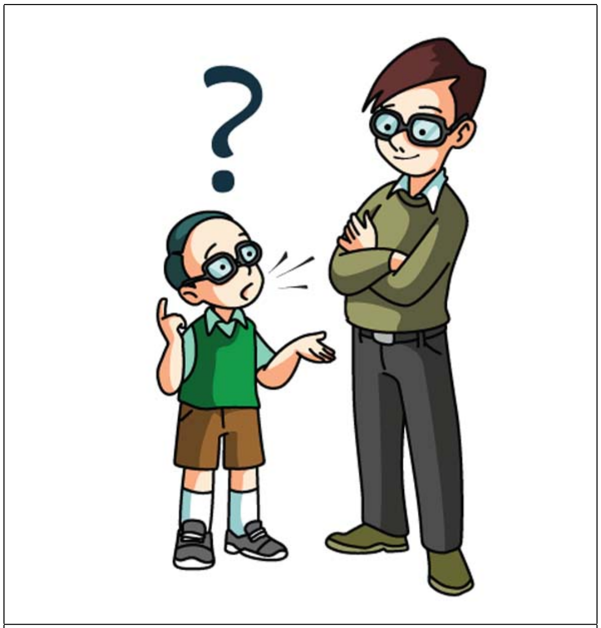
Pronouns as subjects are acceptable in responses to both targeted and untargeted items. A pronoun that clearly assigns the wrong gender to a subject or object should result in a "no" for the core event feature. Otherwise, annotators should retain a high degree of flexibility with regard to pronouns. 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 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 2 and 3 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 and relevant based on the image?*

This feature is mainly concerned with identifying four issues in responses: modality, inaccurate information, unverifiable inferences and irrelevant information.

2.5.1 Contextuality of verifiability

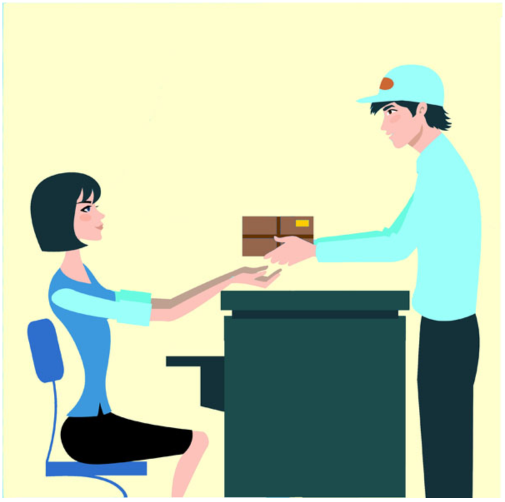
Annotation for the verifiability feature is contextual; it must consider the image presented in the item.

2.5.2 Unintelligible responses

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.

2.5.3 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 5 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 verifiability. Annotators should use their own in judgement in annotating responses that contain variations in interpretation.



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

Figure 5: Item 3, for which the core event is roughly *man delivering a package.*

2.5.4 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.

2.5.5 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.

2.5.6 Unverifiable inferences

Responses containing unverifiable inferences are common among the data. Any such response should be annotated as "no" for this feature. For example, Figure 4 depicts a male child asking a question of a male adult. Although the two figures may bear a resemblance, the image contains no verifiable information about their relationship. Therefore, any response that refers to either person as "son", "brother" or "father" should receive a "no" annotation for this feature.

A similar situation arises for the item in Figure 6, which shows a boy eating a slice of pizza. Some responses to this item refer to the pizza as "sausage", "pepperoni" or "cheese" pizza. Much like the inference of a father/son relationship in Figure 4, these pizza descriptions seem plausible but are not explicitly verifiable based on the image.

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.

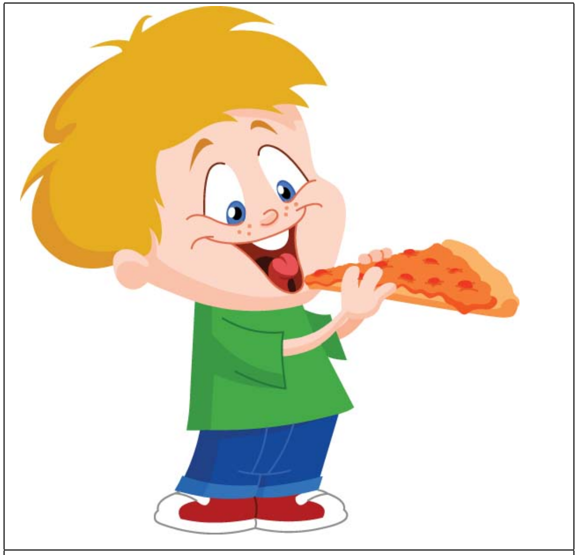
2.5.6.1 Participant opinions

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 6, 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.

2.5.7 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 6, 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 6: 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 question?*

2.6.1 Contextuality of answerhood

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

2.6.2 Accuracy

Answerhood should be annotated without regard to the accuracy of the response. Consider Figure 6 for example. 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. The accuracy of the response is accounted for with the core event and verifiability features.

2.6.3 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 5 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 core event for either version, while *The woman is receiving a package* would be annotated “yes” only for the untargeted version.

2.6.3.1 Verb forms

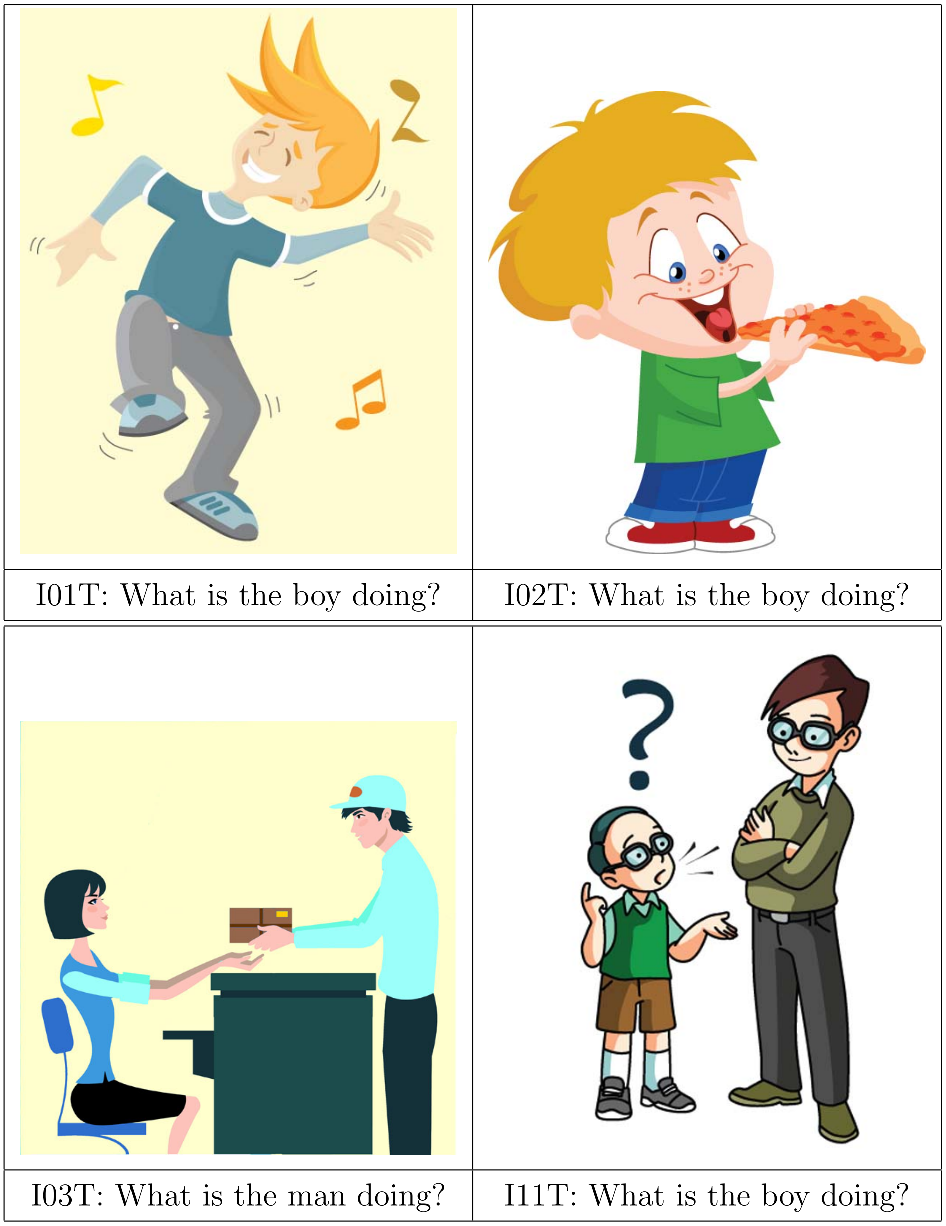
As all items depict some core event or action and ask about that action in the present progressive, responses should contain a dynamic verb to describe that action. This is a key consideration. Even targeted items receive a number of responses that contain only a stative verb describing the state of the situation 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 boy is confused,” a response to Item 11 (Figure 4), 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. Likewise, “The boy is hungry,” a response to Item 2 (Figure 6) is annotated “no” for answerhood, because it does not directly answer the question.

The PDT items ask what *is happening* or what a particular figure in the image *is doing.* 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 generally 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 science*) and verbs of relation (*Alex* ***resembles*** *his father*).

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.* 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. In annotating this feature, however, this is not a hard rule. While it may not seem native-like, a response with the present simple form of a verb may still be annotated as "yes" for this feature if it could reasonably be interpreted to describe the particular event shown in the image. Thus, an annotator should annotate "He eats pizza" as "yes", because the response could be interpreted to refer to this particular event, rather than a habitual action. (The oddness of such a response is captured under the *native-likeness* feature.)

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 7: Example items used in Table 2 and Table 3. The question for all untargeted items is *What is happening?*

	Response	G	N	I	C	V	A
1	dancing	1	1	1	1	1	1
2	the boy is dancing along the music	0	0	1	1	1	1
3	the man is dancing with song	0	0	1	1	1	1
4	the boy dances with music on	1	0	1	1	1	1
5	a boy is dancing around with his smile on his face	1	0	1	1	1	1
6	the kid is eating the pizza	1	1	1	1	1	1
7	the boy enjoys his pizza	1	0	1	1	1	1
8	pizza is this boy's favorite food	1	1	0	0	0	0
9	he's pigging out	1	1	1	0	1	1
10	the boy is eating pepperoni pizza	1	1	1	1	0	1
11	the man is giving a box to a woman	1	1	1	1	1	1
12	a woman is receiving a box from a man	1	1	1	0	1	0
13	he is handing a package	0	0	0	0	1	1
14	the man is putting the package to the woman	0	0	1	0	1	1
15	he is giving a gift to a young woman	1	1	1	0	0	1
16	the man is sending a package to a woman	1	1	1	0	0	1
17	the delivery man is picking up a woman's package	1	1	1	1	1	1
18	the boy is asking a question	1	1	1	1	1	1
19	questioning	1	1	1	1	1	1
20	the boy is asking the older guy a question	1	1	1	1	1	1
21	he's asking his dad a question	1	1	1	1	0	1
22	the little is asking a question	0	0	0	0	0	0
23	the boy asks a question to his father	1	0	1	1	0	1
24	the man is listening to the boy	1	1	1	0	1	0
25	a man is being asked questions by a boy	1	1	1	1	1	0

Table 2: Targeted example responses and annotations for the items shown in Figure 7.

	Response	G	N	I	C	V	A
1	dancing	1	1	0	0	1	1
2	the boy is dancing along the music	0	0	1	1	1	1
3	the man is dancing with song	0	0	1	1	1	1
4	the boy dances with music on	1	0	1	1	1	1
5	a boy is dancing around with his smile on his face	1	0	1	1	1	1
6	the kid is eating the pizza	1	1	1	1	1	1
7	the boy enjoys his pizza	1	0	1	1	1	1
8	pizza is this boy's favorite food	1	1	0	0	0	0
9	he's pigging out	1	1	1	0	1	1
10	the boy is eating pepperoni pizza	1	1	1	1	0	1
11	the man is giving a box to a woman	1	1	1	1	1	1
12	a woman is receiving a box from a man	1	1	1	1	1	1
13	he is handing a package	0	0	0	0	1	1
14	the man is putting the package to the woman	0	0	1	0	1	1
15	he is giving a gift to a young woman	1	1	1	0	0	1
16	the man is sending a package to a woman	1	1	1	0	0	1
17	the delivery man is picking up a woman's package	1	1	1	1	1	1
18	the boy is asking a question	1	1	1	1	1	1
19	questioning	1	1	0	0	1	1
20	the boy is asking the older guy a question	1	1	1	1	1	1
21	he's asking his dad a question	1	1	1	1	0	1
22	the little is asking a question	0	0	0	0	0	0
23	the boy asks a question to his father	1	0	1	1	0	1
24	the man is listening to the boy	1	1	1	0	1	0
25	a man is being asked questions by a boy	1	1	1	1	1	1

Table 3: Untargeted example responses and annotations for the items shown in Figure 7.