

# SEMANTIC ANALYSIS OF IMAGE-BASED LEARNER SENTENCES

L. K.

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Department of Linguistics,

Indiana University

September or October 2018

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy.

Doctoral Committee

---

Markus Dickinson, PhD

---

Sandra Kuebler, PhD

---

David Stringer, PhD

---

Sunyoung Shin, PhD

Date of Defense: 10/10/2018

Copyright © 2018

*L. K.*

Dedicated to yo mama!

## **ACKNOWLEDGEMENTS**

I would like to acknowledge yo mama!

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

L. K.

## Semantic Analysis of Image-Based Learner Sentences

This is where I'd put my abstract if I had one. And right about here I'd start in on the tale of King and Dickinson (2014), because I know you'd want to hear about it. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

---

Markus Dickinson, PhD

---

Sandra Kuebler, PhD

---

David Stringer, PhD

---

Sunyoung Shin, PhD

## TABLE OF CONTENTS

<b>Acknowledgements</b> . . . . .	v
<b>Abstract</b> . . . . .	vi
<b>List of Tables</b> . . . . .	ix
<b>List of Figures</b> . . . . .	x
<b>Chapter 1: Introduction and Background</b> . . . . .	1
<b>Chapter 2: Related Work</b> . . . . .	4
2.1 The state of ICALL and content analysis . . . . .	4
2.2 Learner corpora . . . . .	5
2.3 Language assessment . . . . .	6
2.4 NLP tools and methods . . . . .	6
2.5 My previous work . . . . .	6
2.5.1 2013 . . . . .	6
2.5.2 2014 . . . . .	6
2.5.3 2016 . . . . .	7
2.5.4 2018 . . . . .	7
2.6 Image processing . . . . .	7

<b>Chapter 3: Data Collection</b> . . . . .	9
3.1 Picture Description Task . . . . .	9
3.2 Participants . . . . .	11
3.3 Responses . . . . .	13
<b>Chapter 4: Annotation</b> . . . . .	15
<b>Appendix A: Annotation Guide</b> . . . . .	17
<b>Curriculum Vitae</b>	



## LIST OF TABLES

1.1	This is an example Table. . . . .	3
3.1	PDT example images with their targeted questions. In the untargeted form, the question for each is <i>What is happening?</i> From left to right, the examples represent one intransitive, transitive and ditransitive item. . . . .	11
3.2	First and second response counts for the SAILS Corpus participant groups. Familiar (FNS) and crowdsourced (CNS) are subgroups of NS. NNS participants are not asked to provide a second response. . . . .	13
3.3	NS and NNS type-to-token ratios (TTR) for complete responses (not words), for the full corpus. . . . .	14

## LIST OF FIGURES

1.1	This is an example Figure. . . . .	2
2.1	This is an example figure from King and Dickinson (2013). . . . .	6
2.2	This is an example figure from King and Dickinson (2018). . . . .	8

# CHAPTER 1

## INTRODUCTION AND BACKGROUND

Yeah, yeah, Introduction and background...

Well, I reckon this is where I'd start writing my introduction n' such. I'd probably go on and on telling you about famous papers, like the earth-shattering King and Dickinson (2013). We'll also venture into discussions of spelling correction for learner responses and bag-of-dependencies approaches (King and Dickinson, 2016).

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat (?). Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur (?). Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum (?).

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat (?). Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur (?). Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat (?). Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur (?). Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

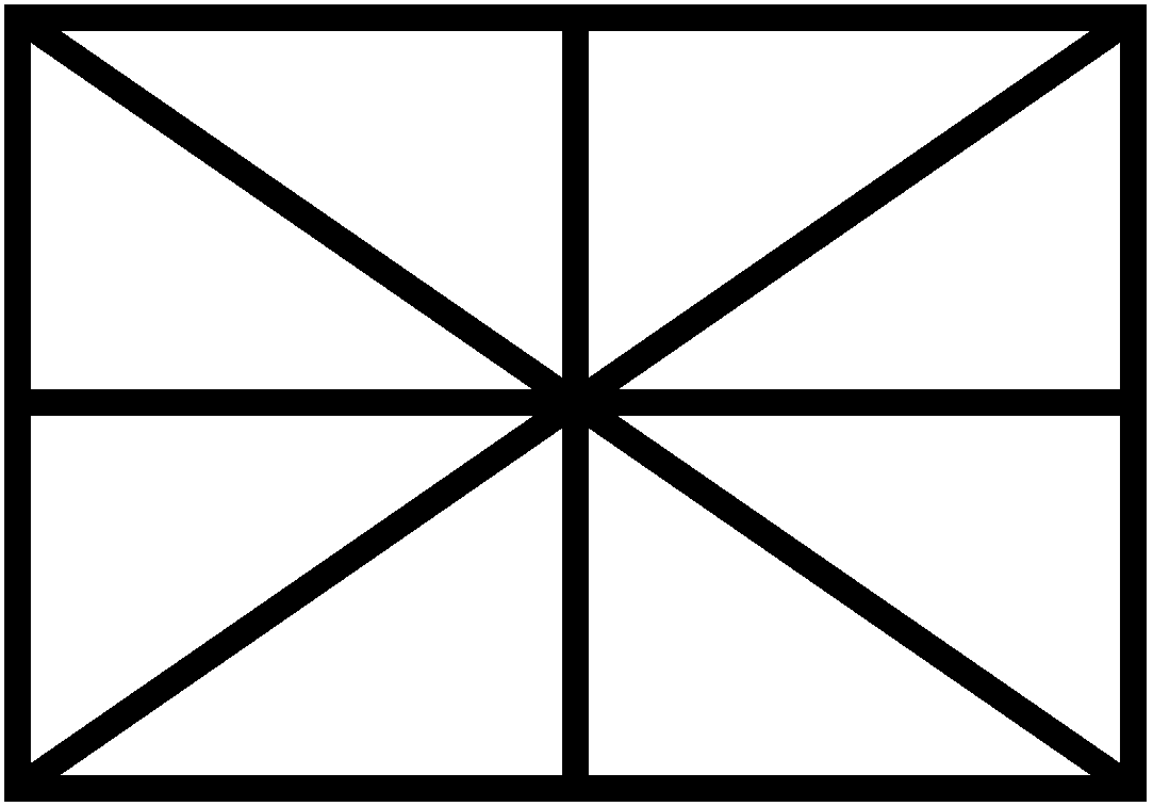


Figure 1.1: This is an example Figure.

Table 1.1: This is an example Table.

x	f(x)	g(x)
1	6	4
2	6	3
3	6	2
4	6	2

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat (?). Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur (?). Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat (?). Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur (?). Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat (?). Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur (?). Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

## **CHAPTER 2**

### **RELATED WORK**

This dissertation lies at the intersection of language testing, second language acquisition (SLA), intelligent computer-assisted language learning (ICALL), corpus linguistics and natural language processing (NLP). My work here is much indebted to related research in these areas, and this chapter will summarize some of the most relevant and comparable studies.

I begin in Section 2.1 with a look at the state of ICALL, its current abilities and limitations, and how it relates generally to this dissertation. In Section 2.2, I summarize research involving the collection, annotation or content analysis of task-based learner corpora. A brief overview of the NLP tools and methods used in my work is given in Section 2.4. Finally, in Section 2.5, I present a summary of my own previous related work.

#### **2.1 The state of ICALL and content analysis**

My dissertation bears much in common with language testing research, but it began as an experiment in bootstrapping NLP tools and learner data to achieve more meaning-based (and meaningful) ICALL. I do not attempt a full-fledged ICALL system, but I explore mechanisms for implementing content analysis that could be implemented in an interactive setting like a game or an interactive language tutor (ILT). This section examines the current state of ICALL, particularly with regard to visual references and meaning-based interactions.

## 2.2 Learner corpora

Here I will discuss task-based learner corpora research that relates to my work. This includes discussions of task design, data collection, annotation schemes, and automatic processing. I focus in particular on the learner content analysis research conducted by two clusters of researchers: one primarily associated with The Ohio State University and consisting of Detmar Meurers and colleagues, and the other primarily associated with Educational Testing Services (ETS) and consisting of Martin Chodorow, Swapna Somasundaran and Joel Tetreault and colleagues.

Here are some papers I discussed briefly in my BEA 2018 paper:

(Leacock et al., 2014)

(Kyle and Crossley, 2015)

(Weigle, 2013)

(Amaral and Meurers, 2007)

(Meurers and Dickinson, 2017)

(Heift and Schulze, 2007)

(Somasundaran et al., 2015)

(Bailey and Meurers, 2008)

(Meurers et al., 2011)

(Somasundaran and Chodorow, 2014)

(Cahill et al., 2014)

(Ragheb and Dickinson, 2014)

(Foster and Tavakoli, 2009)

(Cho et al., 2013)

(Landis and Koch, 1977)

(Artstein and Poesio, 2008)

(Tetreault and Chodorow, 2008a)



Figure 2.1: This is an example figure from King and Dickinson (2013).

(Tetreault and Chodorow, 2008b)

### **2.3 Language assessment**

### **2.4 NLP tools and methods**

### **2.5 My previous work**

Here I will discuss the work I have previously done in this area, including the papers given in the subsections below.

#### **2.5.1 2013**

(King and Dickinson, 2013)

#### **2.5.2 2014**

King and Dickinson (2014)



### **2.5.3 2016**

(King and Dickinson, 2016)

#### *Bag-of-dependencies*

Here we could discuss the switch to a bag-of-dependencies approach, the use of tf-idf and the use of vector cosine distance for ranking responses.

#### *Clustering*

Here we could briefly mention the clustering experiments we did in the 2016 paper. But really, I'd rather not, because I don't intend to repeat them in the dissertation.

### **2.5.4 2018**

(King and Dickinson, 2018)

## **2.6 Image processing**

We want to touch on image processing / automatic captioning / use of semantic primitives, etc. – linguistic annotation of images. NOT a deep discussion, but we need to acknowledge that there are other fields working on the relations between images and text, and give an idea of what some approaches are and how they work, and how they might relate to my work and the work discussed in my lit review.



Figure 2.2: This is an example figure from King and Dickinson (2018).

## CHAPTER 3

### DATA COLLECTION

This chapter will discuss the data collection task, participants and responses.

#### 3.1 Picture Description Task

The PDT is built around 30 images. Each image is a simple, cartoon-like vector graphic. These images were either purchased online or found in publicly available graphics libraries. In order to constrain response contents to the main action of each image, the images were modified to remove any non-essential detail or background. Vector graphics are ideal for this use, because they tend to have an illustrational style with very little detail, as compared to photographs or drawings. Moreover, most consist of layers of graphic objects, and these objects can be easily moved, resized, deleted, combined or otherwise modified to compose the desired stimulus. Example images are presented in Table 3.1 and the full set is contained in Appendix XYZ . To factor out the influence of previous linguistic context, images are intentionally devoid of any text or symbols. The only exceptions to this are two images with music notes, one with a legible analog clock, one with numerals in an arithmetic problem, and one with a question mark. LK: XYZ=?

Each image was chosen for its depiction of an ongoing or imminent action, performed by a person or an animal. The images are divided evenly into actions that are canonically intransitive, transitive and ditransitive in English.

Each PDT image is used in two different contexts: **targeted** and **untargeted**. An **item** consists of an image and a prompt question. For **targeted** items, questions take the form of *What is <subject> doing?*, with the subject provided (e.g., *the boy*, *the bird*). For all **untargeted** items, the question is *What is happening?* Thus, there are a total of 60 different items used in this study. Collecting these targeted and untargeted responses allows for the

examination of response variation with and without a subject constraint. This could help inform approaches to task design and automatic content assessment (Foster and Tavakoli, 2009; Cho et al., 2013).

Multiple versions of the PDT were necessary to collect roughly equal numbers of targeted and untargeted **responses** for each image. These versions vary in which images are presented as targeted items and which images are presented as untargeted items. Additionally, NSs were asked to provide two non-identical responses to each item, but NNSs were asked to provide only one response per item, so different PDT versions were used for these groups. The PDTs were hosted online via Survey Monkey, and all participants submitted their responses through this platform. Survey Monkey restricts the number of questions included in surveys used for purchasing crowdsourced responses, so the PDTs were split accordingly for this group.

In each (full-length) PDT, targeted items are presented in the first half, and untargeted items are presented in the second half. This targeted-untargeted ordering is intended to avoid the possibility that in an untargeted-targeted task, respondents might notice that the question for each untargeted item is always the same in the first half and finish the task hastily without noticing that later targeted items specify the subject. Each half is introduced with instructions, including an example item with sample responses. The instructions ask participants to focus on the main event depicted in the image and for each response to be one complete sentence. The PDT was presented as an online survey, and all participants typed their responses. Participants were instructed not to use any reference materials, but browser-based spell checking was not disabled, and participants are assumed to have used it as necessary.

A complete paper-based version of the PDT is included in Appendix XYZ and the full set of PDT versions is available for download with the SAILS Corpus.<sup>1</sup> The main task instructions are presented in (1). Additional instructions provided to NSs are presented in

LK: XYZ=?

---

<sup>1</sup><https://github.com/sailscorpus/sails>

(2).

- (1) 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 write a **complete sentence**, not a word or phrase.
- (2) 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.




		
What is the woman doing?	What is the woman doing?	What is the man doing?

Table 3.1: PDT example images with their targeted questions. In the untargeted form, the question for each is *What is happening?* From left to right, the examples represent one intransitive, transitive and ditransitive item.

### 3.2 Participants

This study involved a total of 499 PDT participants. Of these, 141 were NNSs, recruited from intermediate and advanced writing courses for English as a Second Language students attending Indiana University. These participants performed the task in a computer lab with

the researchers present. They were native speakers of Mandarin Chinese (125), Korean (4), Burmese (3), Hindi (2), and one native speaker each of Arabic, Indonesian Bahasa, German, Gujarati, Spanish, Thai and Vietnamese. Because nearly 90% of these recruits were native speakers of Chinese, care should be taken when drawing conclusions from the corpus; patterns observed among the NNSs here might not apply broadly to all NNSs.

Of the 358 NS participants, 29 were personally known and recruited by the researcher. Throughout this work, these participants will be referred to as the **Familiar Native Speakers (FNSs)**. Responses from the remaining 329 NSs were purchased via Survey Monkey, an online survey platform. These participants will be referred to as the **Crowdsourced Native Speakers (CNSs)** where participants earn credits they can redeem for gift cards and prizes. Due to length restrictions for purchased surveys, these NSs each completed only half of the task, so their data is equivalent to that of 164.5 full participants.

All participants completed a background questionnaire at the beginning of the PDT. This included questions about first and second languages, gender, age, national origin, amount of English language instruction and length of residency in English-speaking locations. This questionnaire is included as part of the PDT in Appendix XYX, and the background information provided by participants is included in the SAILS Corpus files. LK: XYZ

In previous similar work (King and Dickinson, 2013), NSs were found to produce less variation than NNSs. Many NSs provided identical responses or responses that hew very closely to the most canonical way of expressing the main action. A major motivation for collecting the current corpus was the notion of assessing NNS response content by comparing it against the NS responses. For this reason, NSs were asked to provide two non-identical responses, in the hopes that this would result in a wide range of examples of native-like responses for the NNS responses to be compared against.

### 3.3 Responses

A total of 13,533 responses were collected. The response counts for each participant group are presented in Table 3.2. The overwhelming majority of responses appear to be given in good faith, but a small number of responses (primarily from the CNS group) are problematic in this regard. These may contain gibberish or obscenities or are otherwise inappropriate for the task. Such responses would also be expected in an ICALL environment, so they were not removed from the corpus. Instead, these responses were simply annotated like all others (see Chapter 4). Indeed, automatically assigning low scores to inappropriate responses is a central challenge and goal in this project (see Chapter XYZ).

LK: XYZ

Group	Response Counts		
	First	Second	Total
NNS	4290	0	4290
NS (all)	4634	4609	9243
FNS	642	641	1283
CNS	3992	3968	7960
Total	8924	4609	13,533

Table 3.2: First and second response counts for the SAILS Corpus participant groups. Familiar (FNS) and crowdsourced (CNS) are subgroups of NS. NNS participants are not asked to provide a second response.

To examine the degree of variation among the NSs and NNSs in the current study, type-to-token ratios (TTR) were calculated on the response level for the entire set of items, shown in Table 3.3. For this calculation, final punctuation was ignored, and all responses were converted to lowercase. To illustrate, the three response *tokens* in (3), (4) and (5) would constitute a single response *type*.

(3) The woman is holding a dog

(4) the woman is holding a dog!

(5) The Woman is holding a Dog.

For each data point in the table, the corpus contains 10 items, for each of which there are roughly 150 NS responses and 70 NNS responses. To control for this imbalance and its effect on the likelihood of seeing new responses, the TTR was calculated for each item based on a random sample of 50 responses. This was repeated 10 times and then averaged to produce a final TTR each item. Then, for intransitives, transitives and ditransitives, the TTR was calculated as the average TTR of the 10 items in each set. The scores in in Table 3.3 show that in all cases, the NS set shows a greater degree of response variation, meaning that asking for two responses appears to be an effective way of collecting a broader range of NS responses.

Set	Targeted		Untargeted	
	NS	NNS	NS	NNS
Intrans	0.628	0.381	0.782	0.492
Trans	0.752	0.655	0.859	0.779
Ditrans	0.835	0.817	0.942	0.936

Table 3.3: NS and NNS type-to-token ratios (TTR) for complete responses (not words), for the full corpus.

**\*Let's separate 1st and 2nd responses and calculate the TTRs this way also\***

The ratios show the direct relationship between the complexity of the event portrayed (represented here as intransitive, transitive and ditransitive) and the degree of variation elicited. In all cases, TTR increases with this complexity. Interestingly, this trend seems more pronounced in the NNS responses; in the targeted NNS responses, the TTRs for intransitive and ditransitive items are 0.381 and 0.817, respectively, compared to 0.628 and 0.835 for NS responses. **\*Might this be explained by the inclusion of NNS 2nd responses? Let's investigate\*** The ratios also show that in all cases, as expected, variation is greater for untargeted items than it is for targeted items. In other words, asking about a particular subject in the prompt question does constrain the variety of responses.

LK: 1st  
vs 2nd  
responses

LK: 1st vs  
2nd?



## **CHAPTER 4**

### **ANNOTATION**

Let's talk about annotation.

# **Appendices**

## **APPENDIX A**

### **ANNOTATION GUIDE**

The following pages consist of the annotation guide. This guide was produced through an iterative process of annotation and discussion between the researchers, annotators and outside linguists. This is the final version of the guidelines, which was used to produce the annotations included in the SAILS Corpus.

# Semantic Analysis of Image-Based Learner Sentences (SAILS)

## Annotation Guide

Version 1.0, December 19, 2017

## Contents

<b>1</b>	<b>Task Background</b>	<b>3</b>
1.1	Overview . . . . .	3
1.2	Participants . . . . .	4
1.2.1	Non-native speakers . . . . .	4
1.2.2	Native speakers . . . . .	4
1.2.2.1	Familiar NSs . . . . .	4
1.2.2.2	Crowd-sourced NSs . . . . .	5
1.3	Instructions . . . . .	5
1.4	Item Examples (Targeted and Untargeted) . . . . .	6
<b>2</b>	<b>Annotating Features</b>	<b>9</b>
2.1	Core event . . . . .	9
2.1.1	Contextuality of core event . . . . .	9
2.1.2	Defining <b>core event</b> . . . . .	9
2.1.2.1	Subjects . . . . .	9
2.1.2.2	Verb forms . . . . .	10
2.1.2.3	Content . . . . .	11
2.1.3	Alternative interpretations & inaccurate information . . . . .	12
2.1.4	Language problems . . . . .	12
2.1.5	Imprecise language . . . . .	13
2.1.6	Slang . . . . .	13
2.1.7	Intransitive vs. transitive core events . . . . .	13

2.1.7.1	Intransitive core events . . . . .	13
2.1.7.2	Transitive core events . . . . .	14
2.1.8	Pronouns . . . . .	15
2.1.9	Targeted items and passive responses . . . . .	15
2.1.10	Untargeted item leniency . . . . .	16
2.2	Verifiability . . . . .	16
2.2.1	Contextuality of verifiability . . . . .	17
2.2.2	Reasonable inferences . . . . .	17
2.2.3	Subject and object variation . . . . .	17
2.2.4	Language problems . . . . .	19
2.2.5	Incomplete responses . . . . .	19
2.2.6	Alternative interpretations . . . . .	19
2.2.7	Responses in the form of a question . . . . .	19
2.2.8	Modality . . . . .	20
2.2.9	Unverifiable inferences . . . . .	21
2.2.9.1	Participant opinions . . . . .	21
2.2.10	Irrelevant information . . . . .	22
2.3	Answerhood . . . . .	22
2.3.1	Contextuality of answerhood . . . . .	22
2.3.2	Defining <b>answerhood</b> . . . . .	22
2.3.3	Accuracy . . . . .	24
2.3.4	Targeted vs. untargeted items . . . . .	24
2.3.5	Verb forms . . . . .	24
2.3.5.1	Progressive verbs . . . . .	24
2.3.6	Events and activities . . . . .	26
2.3.7	Imminent actions . . . . .	26
2.3.7.1	Targeted subject variations and pronouns . . . . .	27
2.3.7.2	Misspellings . . . . .	28
2.4	Interpretability . . . . .	28
2.4.1	Semi-contextuality of interpretability . . . . .	29
2.4.2	Defining <b>interpretability</b> . . . . .	29
2.4.2.1	Verb arguments . . . . .	29
2.4.2.2	Content and composition . . . . .	30
2.4.3	Common interpretability concerns . . . . .	31

2.4.3.1	Grammar and spelling . . . . .	31
2.4.3.2	Incomplete sentences . . . . .	32
2.4.3.3	States and actions . . . . .	32
2.4.3.4	Questions and modals . . . . .	33
2.4.3.5	First and second person . . . . .	33
2.4.3.6	Slang . . . . .	33
2.4.3.7	Impossible or unknowable information . . . . .	34
2.5	Grammaticality . . . . .	34
2.5.1	Non-contextuality of grammaticality . . . . .	34
2.5.2	Defining <b>grammaticality</b> . . . . .	35
2.5.3	Incomplete sentences . . . . .	36
2.5.4	Punctuation and capitalization . . . . .	36
2.5.5	Common grammaticality concerns . . . . .	36
2.5.5.1	Events and activities . . . . .	36
2.5.5.2	Non-propositional responses . . . . .	37
2.5.5.3	Bare nouns . . . . .	37
2.5.5.4	Missing <i>be</i> verbs . . . . .	37
2.5.5.5	Misspellings . . . . .	38
2.6	Example items . . . . .	39

# 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 browser-based spelling 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. Participants were instructed to provide one 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. The researchers intend to study the many ways in which responses vary, and to compare these variations for NS and NNS responses. Ultimately, the researchers intend to use the NS responses to derive a kind of answer key or **gold standard (GS)**, which can be used to automatically evaluate the content of NNS responses.

## 1.2 Participants

The following section describes the different participant groups. It is provided for informational purposes only. While annotating, annotators do not need and are not given any information about the source of the responses.

### 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. Roughly 140 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 **crowd-sourced** 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 150 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 (e.g., *the girl*, *the bird*) 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	
	
<b><i>What is the man doing?</i></b>	
<b>Your sentence:</b>	<i>The man is shouting.</i>
<b>Your second sentence:</b>	<i>He is yelling.</i>
<b>There is not a single correct response. Many responses may be possible. Other responses might be:</b> <i>The man is yelling something.</i> <i>He is speaking loudly.</i>	

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

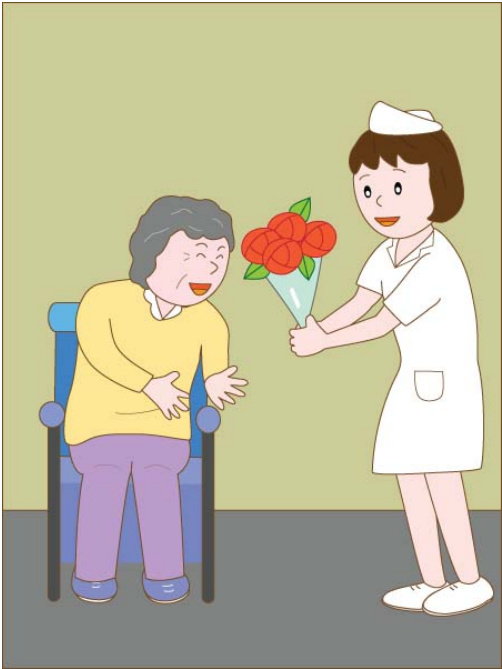
Example 2	
	
<b><i>What is happening?</i></b>	
<b>Your sentence:</b>	<i>The nurse is giving a patient roses.</i>
<b>Your second sentence:</b>	<i>A woman is getting flowers from a nurse.</i>
<b>There is not a single correct response. Many responses may be possible. Other responses might be:</b> <i>The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

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 ***core event***, ***verifiability***, ***answerhood***, ***interpretability*** and ***grammaticality***. 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] would represent a response that was annotated *no* for grammaticality and *yes* for all other features.

### 2.1 Core event

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

#### 2.1.1 Contextuality of core event

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

#### 2.1.2 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 4, for example, the response *He is a dancing machine* does not capture the core event; it describes a characteristic of the boy, but does not describe what is actually taking place in the image.

##### 2.1.2.1 Subjects

The form of a core event is generally similar to that of a *predicate* in traditional grammar. The core event describes what the subject (or agent) is doing. Thus, when annotating for core event, the predicate of the sentence is the most important consideration. However, there are some rules pertaining to the subject. The sentence must include a subject. In the case

of targeted items, the subject may be omitted if it can be understood from the question. Annotators should be quite flexible with regard to the subject, with a few restrictions. Even for targeted items, the subject in the response does not need to be identical to the subject provided in the question. For example, in response to *What is the boy doing?*, responses that restate the subject as *guy* or *kid* or proper names like *Peter* should be accepted. Much flexibility with regard to age should be given as well; infants aside, *man/boy* should be treated interchangeably, as should *woman/girl*. Crucially, the meaning of the subject in the response should not be in conflict with what is shown in the image. Thus, a response that restates the male subject as female or assigns an exclusively female name should not be accepted. More flexibility is allowed for number; a response that depicts a singular subject as plural or vice versa is still acceptable. The rationale for this decision is that the core event feature should avoid penalizing responses for concerns covered by other features. Concerns about number would primarily be covered with the grammaticality and verifiability features. Moreover, while a subject is necessary to fulfill the core event, the focus of this feature is the event itself. In short, responses that assign an incorrect number to the subject are acceptable, but those that change a subject's gender are not.

### 2.1.2.2 Verb forms

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 past or future terms might be acceptable; annotators should use their own best judgment. Responses that use modality or hedging (e.g., *He must be dancing*; *I think he's dancing*), and those that are formed as questions (e.g., *Is he dancing?*) are also acceptable, as long as the core event is present and clearly tied to the appropriate subject (or agent).

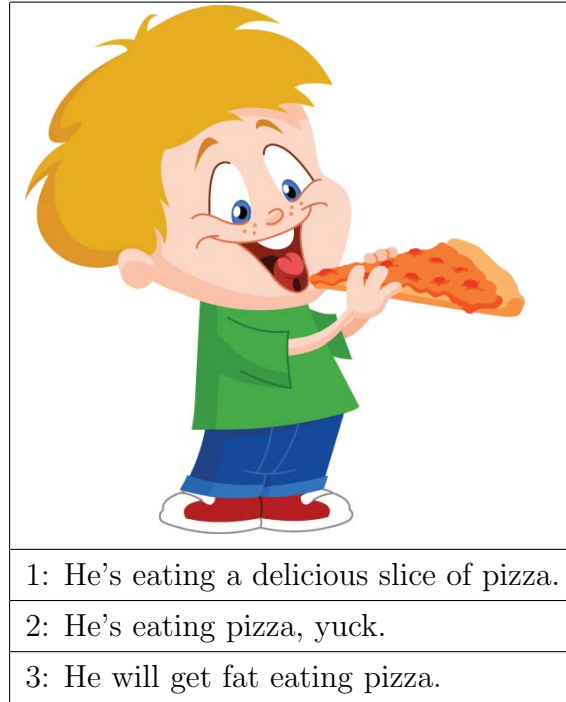


Figure 3: Item 2 (targeted: *What is the boy doing?*) and example responses.

### 2.1.2.3 Content

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 a cause or 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.

### 2.1.3 Alternative interpretations & inaccurate information

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 7 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. An even smaller number of participants describe the scene as a student giving a gift to his teacher. However, the “student” here is wearing a work uniform and holding a brown parcel with a visible shipping label, so this interpretation should be rejected. Annotators should use their own judgement in annotating responses that contain variations in interpretation.

As long as the core event is present and linked to a reasonable subject (or agent), inaccurate information in a response should be ignored and the response should be accepted. For Figure 4, for example, *A boy is dancing at a birthday party* should be annotated *yes*. Although we see no evidence of a party, the response nonetheless covers the core event, which is *(boy) is dancing* or something equivalent. Likewise, the (hypothetical) response *The guy is dancing on the moon* should be accepted, because the core event and a reasonable subject are present.

### 2.1.4 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 3, for example, the responses *He’s eating a **peice** of pizza* and *The boy’s **eatting** pizza* should be annotated *yes*, because the core event in these responses remains intact and interpretable, despite the misspellings. Misspellings or other language problems that lead to ambiguity about the meaning of the core event should be annotated *no*. Annotators should use their best judgment in determining when language problems obscure the core event.

### 2.1.5 Imprecise language

Responses that use imprecise language should be evaluated for how well they convey the core event. Consider, for example, Figure 4, which depicts a boy dancing, and Figure 3, which depicts a boy eating pizza. For Figure 3, the response *A boy is enjoying pizza* should be annotated *yes* because to *enjoy* pizza almost certainly means to *eat* pizza. For Figure 4, however, *A boy is enjoying music* should be annotated *no* because the meaning leaves too many possible interpretations. To *enjoy* music could mean to dance to music, but it could also mean to perform music, to listen to a record or to attend a concert.

### 2.1.6 Slang

Responses that describe the event using slang should be annotated as *yes* for the core event if the language used can be readily understood as equivalent to a more canonical description of the core event. For example, Fig 4 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.

### 2.1.7 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.

#### 2.1.7.1 Intransitive core events

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





Figure 4: Item 1, for which the core event is roughly *boy dancing*.

#### 2.1.7.2 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 5 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 the first response, the subject (*the cyclist*) 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.

### 2.1.8 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 5, 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*. Additionally, as discussed in Section 2.1.2.1, the incorrect use of plural or singular forms to describe subjects (and objects) is not penalized under the core event annotation, and this applies to pronoun forms as well.

### 2.1.9 Targeted items and passive responses

In targeted items, a subject is provided in the question. This provided subject (or its replacement) 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 (or its replacement) 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, the targeted item in Figure 5 asks *What is **the boy** doing?* A passivized response may move this subject to a textitby 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. 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.

### 2.1.10 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* (for Figure 5). 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. Responses that do not somehow convey the notion of the core event, however, should still be rejected. For example, *The man is crossing his arms* and *The boy is gesturing with his hands* do not cover the core event and should be rejected.

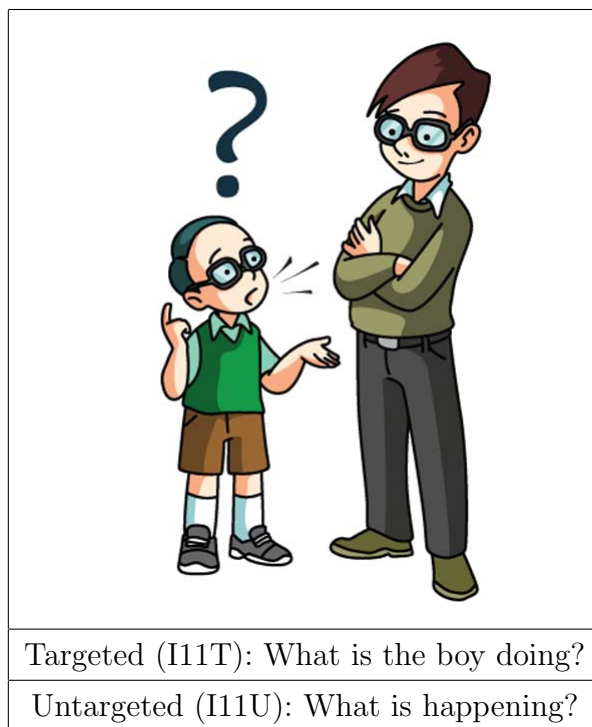


Figure 5: Item 11, for which the core event is roughly *boy asking question*.

## 2.2 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.

### 2.2.1 Contextuality of verifiability

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

### 2.2.2 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 6 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 guesses and reasonable inferences.


### 2.2.3 Subject and object variation

Because verifiability focuses on the truthfulness of information presented in responses, there are few restrictions regarding subjects for this feature. Even for targeted items, responses that omit or change the supplied subject may nonetheless be considered verifiable. Even responses that ignore the question entirely but present information that is verifiably true based on the image should be accepted. For this feature, participants are free to refer to subjects (and other entities) in the images as they wish, so long as they do so accurately and clearly. Responses to a targeted item that asks about *the girl*, for example, may refer instead

to *the lady, the young woman, the short girl*, etc.; if the annotator believes such references are accurate, the responses should be annotated *yes* for verifiability.

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 be 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 acceptable. 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 to women* or *selling cars to women*.



Response	Acceptable 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 6: Example inference judgments (*verifiability*) for *What is the boy doing?*

#### **2.2.4 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 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.

#### **2.2.5 Incomplete responses**

Responses that do not present a complete proposition should be annotated *no* for verifiability. 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.

#### **2.2.6 Alternative interpretations**

Although every effort was made to produce unambiguous PDT images, reasonable alternative interpretations are seen among the responses for some items. For example, Figure 7 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 judgement in annotating responses that contain variations in interpretation.

#### **2.2.7 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. For the verifiability feature, the content of the question is not taken as an assertion of facts and cannot be compared against the facts of the image.

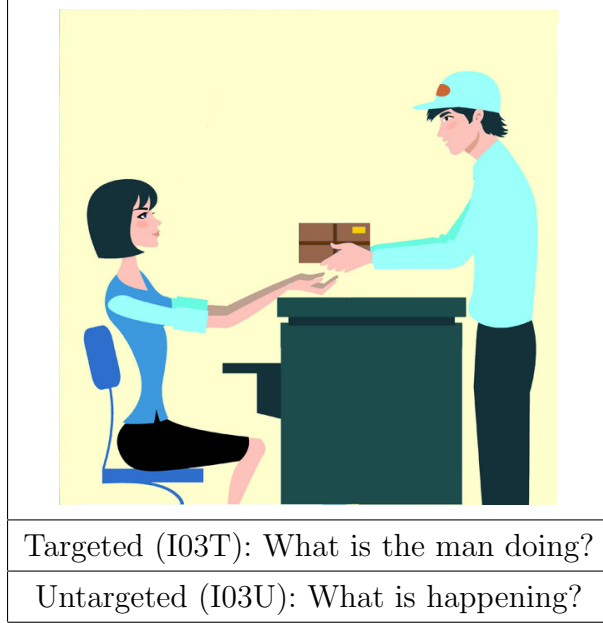


Figure 7: Item 3, in the targeted and untargeted versions.

### 2.2.8 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.

### 2.2.9 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 3, 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 5 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 5). This information is not verifiable, so the response is annotated *no* for this feature.

#### 2.2.9.1 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 3, 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 reasonable 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.



### 2.2.10 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 3, 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.

## 2.3 Answerhood

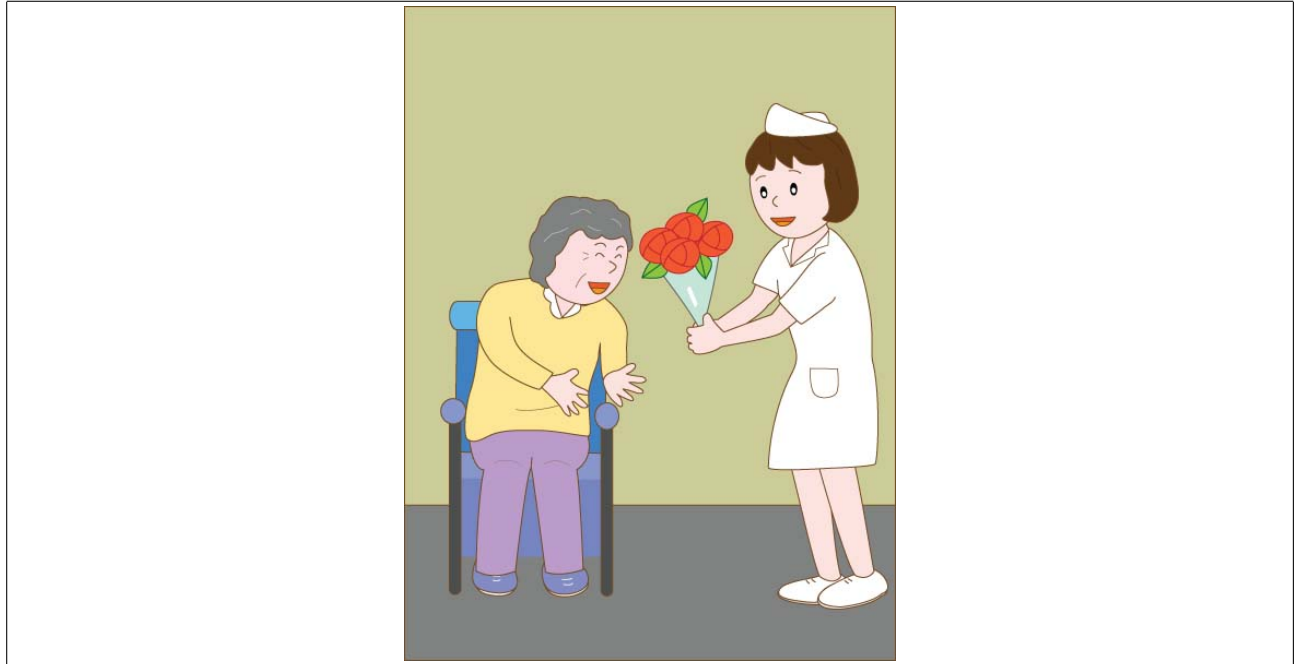
The answerhood feature primarily considers the following question: *Exactly as written, does the response make an attempt to answer the specific question asked?*

### 2.3.1 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.

### 2.3.2 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.*) In general, because all of the PDT questions use a present progressive verb, responses should either use a present progressive verb *or* indicate an imminent action; see Section 2.3.5. Figure 8 presents a number of example responses and answerhood annotations.



	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?

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.

### 2.3.3 Accuracy

Answerhood should be annotated without regard to the accuracy of the response. Consider Figure 3 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.

### 2.3.4 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 7 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.

### 2.3.5 Verb forms

The PDT items ask what *is happening* or what a particular figure in the image *is doing*, and these present progressive verb forms limit the range of acceptable responses. For the purposes of answerhood, acceptable responses should either employ a progressive verb form, indicate imminent action, or present an appropriate event. These forms and related considerations are explained below.

#### 2.3.5.1 Progressive verbs

The majority of responses use a dynamic verb in the progressive form. 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 (present) progressive form

((*is*) *eating*, (*is*) *dancing*). This is in contrast with stative verbs, which 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*). Responses that rely on a stative verb should be annotated *no* for answerhood. These responses (and any others) that simply describe a state of affairs in the image should be annotated *no*, because they do not directly answer the question. For example, *The boy loves pizza*, a response to Item 2 (Figure 3) is annotated *no* for answerhood, because it does not directly answer the question. Likewise, *The nurse seems happy*, shown in Figure 8, should receive a *no* annotation (for both the targeted and untargated versions) because it describes a state depicted in the image but does not directly answer the question of what the nurse is *doing*.

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.

With the exception of *event* responses (see Section 2.3.6) and *imminent action* responses (see Section 2.3.7), responses that lack a progressive verb should be annotated *no*, even if this is the only problem with the response. For example, *The boy is hold a pizza* and *The boy seems to eat pizza* would both be annotated *no*. The mere appearance of a progressive form verb in a response does not automatically satisfy the answerhood feature, however. The necessary progressive verb must appear in a linguistic context that indicates that the verb directly responds to the question. For *What is the dog doing?*, for example, the response *The dog likes to chase the running cat* contains a progressive verb form, but not in a context that satisfies the answerhood feature.

Responses that omit a *be* verb but include a progressive verb form in an otherwise appropriate context (e.g., *The boy holding a pizza*) should generally be annotated *yes* for answerhood. (The grammatical concerns are covered with the grammaticality feature.)

For handling misspelled verbs, see Section 2.3.7.2

### 2.3.6 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, with any necessary determiners or articles. Grammar is not strictly considered for answerhood, but because these responses tend to be very short, proper form is used to differentiate between low-effort responses and those that appear to offer a thoughtful answer to the question. Such low-effort responses may simply describe some element of the image without considering the question. For example, *a baseball game* should be accepted in response to the question *What is happening?*, but *baseball* and *baseball game* should not.

### 2.3.7 Imminent actions

Some responses describe the item in terms of an imminent action rather than a progressive action, e.g., *The boy is about to eat the pizza*. Such imminent action responses are common among the responses from both native and non-native speakers. Some items elicit more of this type of response than others; Figure 3, for example, shows a boy holding a slice of pizza near his mouth. Perhaps because the *eating* action has not yet begun here, many responses indicate this as an imminent action rather than a progressive action. In general, responses that describe the subject's state in relation to an imminent action should be accepted, provided they otherwise fulfill the requirements for answerhood. However, responses that use a future aspect to describe the actions (e.g., *The boy will eat the pizza*) do *not* meet the requirements for answerhood.

Some responses do use a progressive form to indicate an imminent action, such as *The boy is fixin' to eat the pizza* and *The doctor is preparing to treat the patient*. Such responses should be annotated *yes*, and annotators should be flexible in accepting variations and informal forms; for example, *preparing*, *fixin'*, *fixin*, and *gonna* are all acceptable here.

In general, responses that describe the subject’s state in relation to an imminent action are acceptable, with or without a progressive form. This includes responses that use these phrases (or others like them) followed by an action: *is ready to*, *is getting ready to*, *is preparing to*, *is fixing to*, *is about to*, *is gonna*, etc. In the case of *ready to* and *about to*, because these expressions lack an actual verb, they must be preceded by a copular verb (*is*, *seems*, etc.), which cannot be dropped. Likewise, the subject cannot be dropped. For example, *preparing to eat the pizza* is acceptable in response to the question, *What is the boy doing?*, but *about to eat the pizza* is not acceptable.

### 2.3.7.1 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 be 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 blond nurse*, *the nurse who is standing*, or *this nurse* is the same as *the nurse*. Additionally, a targeted subject should not be expanded to include other persons or entities; in response to *What is the man doing?*, *The man is greeting the woman* is acceptable, while *The man and woman are saying hello* is not.

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. Exceptions may be made for babies and animals 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.

### 2.3.7.2 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.

## 2.4 Interpretability

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

### 2.4.1 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 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 2.4.3.2 for more discussion of incomplete sentences.)

### 2.4.2 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.

#### 2.4.2.1 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*.

#### 2.4.2.2 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 vague 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.

### 2.4.3 Common interpretability concerns

#### 2.4.3.1 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.

### 2.4.3.2 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.

### 2.4.3.3 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.

#### 2.4.3.4 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.)

#### 2.4.3.5 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*.

#### 2.4.3.6 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.

### 2.4.3.7 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.

## 2.5 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?*

### 2.5.1 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.

Moreover, the item question may be taken into consideration when it is necessary for assessing

the grammaticality of a particular response. Responses to targeted questions (*What is the X doing*), for example, commonly drop the subject. Such responses can be grammatical; see Section 2.5.3.

### 2.5.2 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 4: *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 (presumably *A boy listens to music and (he) dances*), 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*. This also commonly applies to responses that use an incorrect (but grammatical) pronoun. For example, *The boy is talking to her brother*, in response to Figure 5 (where no female is pictured or otherwise indicated as a potential antecedent to *her*), should be annotated *yes* for grammaticality.

### 2.5.3 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 2.5.5.2, 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.

### 2.5.4 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.

Sentence internal punctuation should be considered for this feature, but annotators should be lenient and keep in mind that many punctuation decisions may simply be a matter of style rather than grammar. Punctuation (or lack thereof) that results in ambiguity or leads the annotator to question the overall grammaticality of the sentence should result in a *no* annotation for the response. Annotators should use their own best judgment in assessing such cases.

### 2.5.5 Common grammaticality concerns

#### 2.5.5.1 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 considered grammatical. 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, with any necessary determiners or articles. For example, *a baseball game* should be accepted in response to the question *What is happening?*, but *baseball* and *baseball game* should not.

#### 2.5.5.2 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*.

#### 2.5.5.3 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*.

#### 2.5.5.4 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 (and most others), 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 2.5.3.)



#### 2.5.5.5 Misspellings

Misspellings generally result in a *no* for grammaticality. 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.

Some responses use proper names for persons, places or objects in the images. When a proper noun appears to be misspelled, annotators should be less strict. If the proper noun is reasonably interpretable, the response should still be annotated *yes*, provided it has no other disqualifying problems. Annotators should use their own judgment in assessing such cases.

## 2.6 Example items

	
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, including *targeted* questions. The question for all *untargeted* items is *What is happening?*

## BIBLIOGRAPHY

- Luiz Amaral and Detmar Meurers. 2007. Conceptualizing student models for ICALL. In *Proceedings of the 11th International Conference on User Modeling*, Corfu, Greece.
- Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4).
- Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In *The 3rd Workshop on Innovative Use of NLP for Building Educational Applications (ACL08-NLP-Education)*, pages 107–115, Columbus, OH.
- Aoife Cahill, Binod Gyawali, and James Bruno. 2014. Self-training for parsing learner text. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 66–73, Dublin, Ireland. Dublin City University.
- Yeonsuk Cho, Frank Rijmen, and Jakub Novák. 2013. Investigating the effects of prompt characteristics on the comparability of toefl ibt integrated writing tasks. *Language Testing*, 30(4):513–534.
- Pauline Foster and Parvaneh Tavakoli. 2009. Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language learning*, 59(4):866–896.
- Trude Heift and Mathias Schulze. 2007. *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.
- Levi King and Markus Dickinson. 2013. Shallow semantic analysis of interactive learner

- sentences. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–21, Atlanta, Georgia.
- Levi King and Markus Dickinson. 2014. Leveraging known semantics for spelling correction. In *Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning*, pages 43–58, Uppsala, Sweden.
- Levi King and Markus Dickinson. 2016. Shallow semantic reasoning from an incomplete gold standard for learner language. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 112–121, San Diego, California.
- Levi King and Markus Dickinson. 2018. Annotating picture description task responses for content analysis. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana.
- Kristopher Kyle and Scott A Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners*, second edition. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- Detmar Meurers and Markus Dickinson. 2017. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning. Special Issue on Language learning research at the intersection of experimental, corpus-based and computational methods*, 67(S1):66–95.

- Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011. Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9. Association for Computational Linguistics.
- Marwa Ragheb and Markus Dickinson. 2014. The effect of annotation scheme decisions on parsing learner data. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, Tübingen, Germany.
- Swapna Somasundaran and Martin Chodorow. 2014. Automated measures of specific vocabulary knowledge from constructed responses (‘Use these words to write a sentence based on this picture’). In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, Baltimore, Maryland.
- Swapna Somasundaran, Chong Min Lee, Martin Chodorow, and Xinhao Wang. 2015. Automated scoring of picture-based story narration. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–48, Denver, Colorado. Association for Computational Linguistics.
- Joel Tetreault and Martin Chodorow. 2008a. Native judgments of non-native usage: Experiments in preposition error detection. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 24–32, Manchester, UK. Coling 2008 Organizing Committee.
- Joel Tetreault and Martin Chodorow. 2008b. The ups and downs of preposition error detection in ESL writing. In *Proceedings of COLING-08*, Manchester.
- Sara Cushing Weigle. 2013. English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, 18(1):85–99.

## **VITA**

Levi King was born in 1818 on the Oregon Trail in modern day Nebraska. He spent his youth hunting direwolves and learning the songs of the Skrull people. Vita may be provided by doctoral students only. The length of the vita is preferably one page. It may include the place of birth and should be written in third person. This vita is similar to the author biography found on book jackets.