SEMANTIC ANALYIS OF IMAGE-BASED LEARNER SENTENCES

Levi King

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the

requirements for the degree of Doctor of Philosophy.

Doctoral Committee

<div style="text-align: right">

_____

Markus Dickinson, PhD

_____

Sandra Kuebler, PhD

_____

David Stringer, PhD

_____

Sunyoung Shin, PhD

</div>

Date of Defense: Month/Day/2020

XYZ

ACKNOWLEDGEMENTS

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.

Levi King

Semantic Analyis of Image-Based Learner Sentences

King and Dickinson (2014)... 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

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND BACKGROUND

A way out west there was a fella, fella I want to tell you about, fella by the name of Jeff Lebowski. At least, that was the handle his lovin' parents gave him, but he never had much use for it himself. This Lebowski, he called himself the Dude. Now, Dude, that's a name no one would self-apply where I come from. But then, there was a lot about the Dude that didn't make a whole lot of sense to me. And a lot about where he lived, like- wise. But then again, maybe that's why I found the place s'durned innarestin'...

They call Los Angeles the City of Angels. I didn't find it to be that exactly, but I'll allow as there are some nice folks there. 'Course, I can't say I seen London, and I never been to France, and I ain't never seen no queen in her damn undies as the fella says. But I'll tell you what, after seeing Los Angeles and thisahere story I'm about to unfold– wal, I guess I seen somethin' ever' bit as stupefyin' as ya'd see in any a those other places, and in English too, so I can die with a smile on my face without feelin' like the good Lord ripped me.

Now this story I'm about to unfold took place back in the early nineties– just about the time of our conflict with Sad'm and the Eye-rackies. I only mention it 'cause some- times there's a man–I won't say a hee-ro, 'cause what's a hee-ro?–but sometimes there's a man.

And I'm talkin' about the Dude here– sometimes there's a man who, wal, he's the man for his time'n place, he fits right in there–and that's the Dude, in Los Angeles.

...and even if he's a lazy man, and the Dude was certainly that–quite possibly the laziest in Los Angeles County..which would place him high in the runnin' for laziest worldwide– but sometimes there's a man. . . sometimes there's a man.

Wal, I lost m'train of thought here. But–aw hell, I done innerduced him enough.

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 studies.

I begin in Section 2.1 with a discussion of the importance of transparency and interpretability in ICALL and language testing. In Section 2.3, I examine approaches to ICALL that relate to and inform this dissertation. In Section 2.4, 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 2.6. Finally, in Section 2.7, I present a summary of my own previous work related to this dissertation.

2.1 On interpretability for learner applications

(See also Section 2.2; this should address the use of sentence encoders like BERT and their use in conjunction with image recognition technology).

This work would be remiss without discussing the role of machine learning (ML) in current NLP, given that such technologies are largely absent from this dissertation. Recent years have seen the rapid development of ML technologies like neural networks and deep learning. These technologies have been widely implemented in areas like NLP and computer vision, often with impressive gains in performance. They can also lead to reductions in the amount of human expertise needed to automate tasks like syntactic labeling, voice recognition and synthesis, and object and facial recognition. Naturally, this also means significant reductions in the cost of such systems. A major drawback with many such ML technologies, however is the loss of interpretability. For example, *word embeddings* (such

LK: Maybe it's better to have a P here summarizing my "ethos" (low resource, content focused, interpretable) instead of spreading that out below.

LK: What is ML good at? citations!

LK: cite stuff

as Word2Vec) are ML based NLP tools suitable for many tasks involving the processing of linguistic meaning or structures. Word2Vec essentially "learns" an approximation of the meanings and grammatical usage of words by observing them in context. Instead of relying on expert annotation of features like part-of-speech, sentence structure and morphology to train a model, the system needs only large amounts of raw text. From this text, it observes large numbers of features, such as the average distance between instances of *Word A* and *Word B*. It reduces these raw features to a (still quite large) number of abstract features, or "latent variables", which form a vector of numeric values; this vector then serves as a representation of a word's "meaning". In a classic example, if one takes the vector for "queen", subtracts the vector for "female" and adds the vector for "male", the resulting vector is roughly equivalent to that of "king".

LK: cite

LK: this is all very cf Brian Riordan's alumni talk... similar sources would be ideal

For many applications, the capabilities and cost reduction of ML make it an attractive and suitable choice; this is certainly the case with many NLP tasks. The use of ML tools with learner language is problematic on at least two fronts, however. First, such tools are typically designed for and trained on well-formed, native-like text (or speech). As mentioned, these tools generally do not rely on annotation in their training data; instead, they make up for this lack of expertise by the sheer volume of training data they consume. Including real learner data on the scale required by ML tools would be impractical if not impossible for most researchers. As a result of ML tools' training on mostly native-like data, they are ill-equipped for processing the variability and ambiguity of learner language. For example, native English trained NLP tools expect regular sentence punctuation; text from a beginning English learner lacking in punctuation could therefore be misconstrued as having longer sentences and thus higher proficiency (Meurers and Dickinson, 2017a). Second, and perhaps more importantly, tools that rely heavily on ML are inherently less interpretable than "classical" NLP tools. Because classical NLP tools are trained on expert annotation, their output is generally determined by the kinds of features that are annotated in the training corpus. This means linguistic researchers can design NLP tools and pipelines

that produce output precisely suited for their research questions, so long as they have the resources to produce adequate training corpora. This is not the case with ML based tools, however. Due to their reliance on abstract features and latent variables, these newer tools are largely "black box" technologies; raw data goes in and processed data comes out, but even the architect of such a system cannot explain exactly how or why the analysis was produced. In a language learning application, this is problematic because it means the development of a pedagogically sound feedback system for the learner is not possible; the features underlying the analysis are not accessible or interpretable. The outcomes of language testing can have a tremendous impact on a test taker's future, and in such a high stakes application, the lack of interpretability can be even more problematic. Arguably, it is far better for all stakeholders if a language test can deliver not only a score, but also a rationale for that score, such as which kinds of errors a test taker makes and in what contexts. This need for interpretable features was one of a few major factors in the decision to choose classical NLP over newer ML tools in this dissertation, and most of the related studies discussed here take similar approaches.

2.2 Image processing

This should probably include discussion of ML approaches at image "encoding/decoding" and their use in tandem with sentence encoders (BERT, etc.). Ask Ben S. for reading suggestions?

We want to touch on image processing / automatic captioning / use of semantic primatives, 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.

2.3 An overview of ICALL and content analysis

This dissertation 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 performing the core content analysis that could be implemented in a setting like a game, an interactive language tutor (ILT) or a language test. I see this work as a push toward relatively low-cost, extendable ICALL with an emphasis on content over form. Each of these points is an attempt at a more interdisciplinary and pedagogically sound approach to ICALL. In keeping with this ethos, this section focuses on ICALL research that primarily uses existing NLP tools and allows for free user input (as opposed to menu-based input).

One relatively well-documented ICALL system is TAGARELA, an application for adult learners of Portuguese (Amaral, 2007; Amaral and Meurers, 2007). In more recent updates to the system, the authors describe TAGARELA's "Unstructured Information Management Architecture (UIMA)," which is effectively a collection of text relevant to the tasks that are enriched with multiple annotations relating to both form and meaning. TAGARELA relies on a set of NLP modules implemented in a flexible, task-based, free input ICALL system (Amaral et al., 2011). The system includes six different activity types: reading, listening, description, rephrasing, vocabulary and fill-in-the-blank. These different tasks require different types of (textual) learner input as well as different subsets of the NLP modules for input processing and the generation of feedback.

Before developing anything, the TAGARELA team began their work with the creation of a "taxonomy of expected errors," gleaned from their analysis of a corpus of written assignments from learners of Portuguese. The authors describe their approach as "data-driven rather than process-driven". In many ways, this bucks a tendency among many ICALL developers to simply address the kinds of errors NLP tools can readily identify. What is needed instead is ICALL development informed by second language acquisition

LK: cite DuoLingo? etc.

research and by the kinds of challenges learners face, as borne out by real data. These errors annotated in the learner data and handled by TAGARELA cover both form and meaning, with error types consisting of, for example, *spelling*, *agreement* and *word choice*.

One major advantage of TAGARELA's approach to ICALL is its ability to accommodate multiple activities with only a handful of existing and custom NLP modules along with a small number of carefully chosen and annotated model responses. The current dissertation found inspiration in TAGARELA's prioritization of the handling of errors related to meaning and its reliance on ordinary and interpretable NLP tools – primarily a tokenizer, part of speech tagger and syntactic parser.

2.4 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 Tetrault 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, 2017b)

(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)

(Tetreault and Chodorow, 2008b)

2.5   Language assessment

2.6   NLP tools and methods

2.7   My previous work

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

2.7.1   2013

(King and Dickinson, 2013)

2.7.2   2014

King and Dickinson (2014)

2.7.3   2016

(King and Dickinson, 2016)

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

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.

2.7.4 2018

(King and Dickinson, 2018)

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

CHAPTER 3

DATA COLLECTION

In Chapter 1 I explained that a major motivation of this work is to investigate relatively low-resource mechanisms for content analysis that can help shift the focus of ICALL from form to meaning. In Chapter 2, I examined related work in testing and ICALL. While numerous creative approaches to contextual content analysis are discussed in the literature, the data they rely on is typically not available to other researchers. With these considerations in mind, I decided to collect a corpus of picture description task responses for use in my experiments. This chapter will discuss the data collection task, participants and responses.

3.1 Picture Description Task

The picture description task (PDT) is built around 30 images. Each image is a simple, cartoon-like vector graphic. These images were purchased from Shutterstock, a web-based graphics library[1]. 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; an example is shown in Figure 3.1. 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 found in Appendix A.

To factor out the influence of previous linguistic context, images are intentionally devoid of any text. In a few cases, symbols are used: two images have music notes, one displays a legible analog clock, one uses numerals in an arithmetic problem, and one shows a question mark. The symbols were intended to elicit abstract concepts that are otherwise

---

[1]https://www.shutterstock.com/vectors

difficult to portray visually, like *teaching math* and *asking a question*.

| Original image | Simplified PDT image |
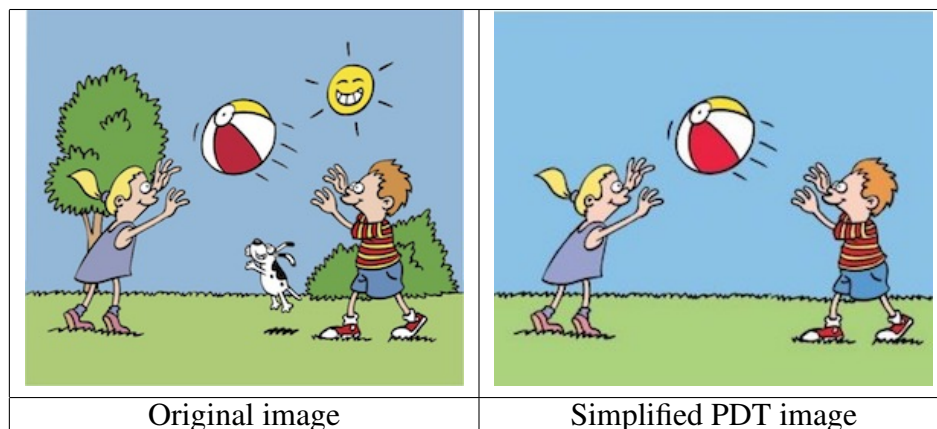
Figure 3.1: All non-essential details were removed from the PDT images in order to focus participants' attention on the main action.

Each image was chosen for its depiction of an ongoing or imminent action (as opposed to a static image) performed by a person or an animal. The images are divided evenly into actions that are canonically intransitive, transitive or 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 girl*, *the boy*; see Table 3.1). For all **untargeted** items, the question is *What is happening?* 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).

Different participants performed different versions of the PDT, and multiple versions 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, native speakers (NSs) were asked to provide two non-identical responses to each item, but non-native speakers (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[2], and all participants submitted their responses through this platform.

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. Because the PDT was presented as an online survey, 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.

The main task instructions are presented in (1). Additional instructions provided to NSs are presented in (2). The full set of PDT versions is available for download with the SAILS Corpus.[3]

(1) **Instructions:** 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) **Additional Instructions for NSs:** 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.

---

[2]https://www.surveymonkey.com
[3]https://github.com/sailscorpus/sails
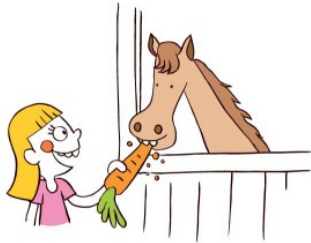
| What is the girl doing? | What is the boy doing? | What is the girl doing? |

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 in the English Language Improvement Program at Indiana University. These participants performed the task in a computer lab with a researcher 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.

Responses from 329 of the NSs were purchased via Survey Monkey, where survey takers can earn credits that they can redeem for prizes or convert to donations to charities. Due to survey length restrictions for purchased responses, these NSs each completed only half of the task, so their data is equivalent to that of 164.5 full participants.

The remaining 29 NS participants were people known to me personally. Due to this relatively small number of participants, their data was not used for modeling or evaluating

NNS responses, but it was annotated and is included in the SAILS Corpus. The NS data discussed throughout this dissertation is the crowdsourced data. Where relevant, however, I refer to these two groups as the **Familiar Native Speakers (FNSs)** and the **Crowdsourced Native Speakers (CNSs)**. With considerably more FNS data, future work should include comparisons of the two groups to better understand the differences in quality, as CNSs might be less likely to perform the task in good faith.

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, and the background information provided by participants is included in the SAILS Corpus files. A summary of some of the demographic information is shown in Table 3.2.

	NNS		CNS		FNS	
Mean age		18.7		45.0		39.1
Median age		18.0		44.0		35.0
Male	56	(39.7%)	138	(41.9%)	17	(58.8%)
Female	76	(53.2%)	172	(52.3%)	11	(37.9%)
Unknown	9	(6.4%)	19	(5.8%)	1	(3.4%)

Table 3.2: Age and gender information for the three participant groups.

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. Among other things, this involves the matching of words or syntactic dependencies and thus benefits from a broad set of acceptable responses in the gold standard. 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 3.3. Including the second responses collected from NSs, roughly two thirds of the corpus come from the NS groups. 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, as shown in Table 3.4. 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 5).

	Response Counts		
Group	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 3.3:   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.

Type-to-token ratios (TTR) are commonly used as an indication of how varied or homogenous a set of data is. This number ranges between 0 and 1. In a set of data where most instances or *tokens* are unique (*types*), the number of types per tokens approaches 1. In a set of data where most tokens are identical, the number of types per tokens approaches 0. With regard to language data, TTRs are often used on the word level, to calculate the lexical density of a document, for example (Granger et al., 2002). In this study, however, type-to-token ratios (TTR) were calculated on the response level for the entire set of items. For this

Exemplar: *The girl is laughing.*
Girl ate a 2x4 and is vomiting toothpicks.
I have to poop so bad.
Exemplar: *The boy is eating pizza.*
How is the pizza staying perfectly horizontal when the boy is holding it so close to the tip?
see my last statement
Exemplar: *The girl is feeding a carrot to a horse.*
Creepy clown child grinding her carrot down on poor Ed's beaver teeth
Hoj

Table 3.4: Crowdsourced responses for the items shown in Table 3.1, showing one exemplar response and two examples of problematic or bad faith responses for each item.

calculation, final punctuation was ignored, and all responses were converted to lowercase. To illustrate, the first three response *tokens* in Table 3.5 would constitute a single response *type*.

Types	Tokens	Response
1	1	The woman is holding a dog
1	2	the woman is holding a dog!
1	3	The Woman is holding a Dog.
2	4	The woman is hugging a puppy.
3	5	The woman squeezed a dog.

Table 3.5: This toy dataset shows how TTR is calculated on the response level. Ignoring punctuation and capitalization, the first three response tokens here constitute a single response type. The TTR for this set would be 3:5, or 0.6.

The TTRs for the corpus are presented in Table 3.6. For each cell in this 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 between NS and NNS 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 for each item. These item TTRs were then averaged as intransitives, transitives and ditransitives. 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

16

with this complexity. 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, as expected.

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

Additionally, from Table 3.6 we can see that the NS set contains a greater degree of response variation than does the NNS set. Note that the TTRs here are calculated on *all* responses, and the NS participants each provided two responses per item, whereas the NNS participants were only asked to provide one response per item. This suggests that asking for two responses is an effective way of collecting a broader range of NS responses. This variability can be more closely examined in Table 3.7, which presents separate TTRs for all NS participants' first responses and second responses. The numbers show that in general, first responses are far less varied than second responses. As we can see, among first responses, variability increases along with item complexity. The pattern holds for targeted second responses, although it is not as pronounced. For untargeted second responses, this monotonic increase in variability is not present, but all three TTRs vary by less than three percent, suggesting that a ceiling effect may be at work.

Finally, for ease of comparison, Table 3.8 presents the (NS only) first response TTRs from Table 3.7 alongside the NNS first (and only) response TTRs from Table 3.6. These comparisons should be made with caution, however, as they cannot account for the possibility of task effects arising from the different instructions given to NS and NNS participants. In other words, it is possible that the anticipation of providing a second response influences a NS participant's choice of first response, and any such effect would be absent for

NNS participants. A future study in which NSs are asked to provide only one response per item could be useful in examining the possibility of such a task effect. As it stands, the table suggests that NNSs generally do exhibit greater response variability than NSs; the only exception to this trend appears among the intransitive untargeted items. This trend is in keeping with the observations from previous work (King and Dickinson, 2013), which found that NSs tend toward canonical forms, while NNSs use whatever language may be available to them, resulting in greater variation. As described above, this was the motivation for asking NSs for two responses.

	Targeted		Untargeted	
Set	R1	R2	R1	R2
Intrans	0.343	0.819	0.549	0.939
Trans	0.509	0.895	0.682	0.926
Ditrans	0.641	0.948	0.864	0.955

Table 3.7: TTRs for complete responses, separated by first responses (R1) and second responses (R2). The ratios here are calculated from all NS responses; NNS responses are not included.

	Targeted		Untargeted	
Set	NS	NNS	NS	NNS
Intrans	0.343	0.381	0.549	0.492
Trans	0.509	0.655	0.682	0.779
Ditrans	0.641	0.817	0.864	0.936

Table 3.8: TTRs for complete responses, comparing first responses only.

Having examined response variation in a rather abstract sense here, Chapter 4 will focus on annotating response features to obtain a more fine-grained view of the ways in which responses can vary.

CHAPTER 4

ANNOTATION & WEIGHTING

I begin this chapter with a discussion of the development and implementation of an annotation scheme that captures aspects of native-likeness and accuracy that are appropriate for content analysis. In the second section of this chapter, I examine inter-annotator agreement for the individual annotation features on a sample of the responses. In the final section of this chapter, I discuss how weights were assigned to these features.

4.1   Annotation scheme

The goal of the annotation is to provide information that would be useful for the automatic content assessment of NNS responses via comparison with NS responses. The idea here is that annotations of relevant features can be used to score and then rank responses. Because my automatic assessment system relies only on surface level features (not annotations), the system's performance can be tuned and evaluated by comparing its ranked output to the annotation based rankings.

The annotation scheme was developed through an iterative process of annotation, discussion and revision, with input from two annotators and other language professionals. The annotation was developed on and applied to both NS and NNS responses. To avoid any potential bias, annotators received the responses in random order and without any demographic information.

For NS responses, the annotation allows for weighting or filtering by features to establish a set of gold standard (GS) responses, which can be used to evaluate the NNS responses. The annotation also allows for evaluation of the overall appropriateness of crowdsourcing a GS for content assessment. By having annotated responses for various groups of NS respondents (crowdsourced strangers versus personally recruited friends and

colleagues, for example), I can examine which respondents lead to the best GS.

For NNS responses, the annotation would be used in a testing scenario to evaluate responses; in an ICALL scenario, it could be used to gauge a participant's understanding and influence the next steps in the activity. In my current work, the annotation functions as a score which can be compared to scores provided by my automatic system, allowing for evaluation of the system itself. Furthermore, the annotation lends insights into which aspects of a response are the most difficult to account for in my approach to content assessment.

The scheme was initially planned as a single three-point scale, ranging from *accurate and native-like* to *accurate but not native-like* to *not accurate*. This proved problematic, however, as *accuracy* and *native-likeness* could not be adequately defined and applied to the data as a single score. For example, in Table 4.1, it is not clear how native-like *She is happy with the dog* is. Grammatically, it is native-like, but it does not seem like an appropriate answer to the question, *What is the woman doing?* Moreover, *The dog is so happy!* may be native-like in terms of language use, but does not seem appropriate in the context of the question. Thus, for the purpose of analyzing content in PDT responses, native-likeness seems to encompass considerations beyond language use and grammar.

Likewise, accuracy could not be satisfactorily defined as a simple *yes* or *no* construct. To illustrate, consider the ramifications of the response *hugging her dog Fluffy that she missed while on vacation* (Table 4.1) as either a NS or NNS response. The response does capture the main action of the item, but embellishes with unknowable details like the dog's name and the subject's motivation. This kind of response is undesirable in its own right, but could also lead to real problems during the automatic scoring. If included in a gold standard (GS) set of NS responses which serve as the basis for scoring new NNS responses, this kind of embellishment would dilute the most salient and desirable information in the GS. Furthermore, if such a NNS response is annotated as accurate, this additional information is unlikely to be readily mapped to information found in the GS, which would lead to

lower scores for the response. Accuracy, it seems, is an inadequate construct for the approach to content assessment envisioned for this work. Clearly, *verifiability* is an important consideration as well.

| 1. Holding a puppy and looks happy |
| 2. She is happy with the dog. |
| 3. She is wear a blue dress. |
| 4. hugging her dog Fluffy that she missed while on vacation. |
| 5. The dog is so happy! |
| 6. She loves her pet |

Figure 4.1: Sample responses for the targeted item, *What is the woman doing?*

In order to handle the kinds of meaningful variation observed in the responses, five binary features were eventually settled on, with each feature having some relation to the original concepts of accuracy and native-likeness. As with most annotation schemes, the final SAILS scheme is a compromise. This scheme represents the minimal set of features that I believe necessary to accomplish two major goals of this work: investigating the use of NS responses as a GS, and examining the factors that lead a NNS response to be rated highly or lowly. Besides the features explained below, others were explored but rejected.

For example, a *good faith* feature was considered to identify responses that were not given in good faith, such as gibberish, profanity and irrelevant responses. Such a discrimination was applicable to less than three percent of responses in the development set, however, so this feature was deemed too costly for the value it would provide.

A set of annotation guidelines were produced with definitions, rules and examples for each feature. For most features, the rules for targeted and untargeted items vary slightly; the untargeted rules are generally less strict to accommodate the less restrictive prompt question. The complete annotation guide is included in Appendix XYZ. The features and brief descriptions are listed here and discussed further in the discussion of inter-annotator agreement in Section 4.2.

1. **Core Event**: Does the response capture the core event depicted in the image? Core events are not pre-defined for annotators but should be clear given the stripped down nature of the images. Crucially, the response should link an appropriate subject to the event. In Table 4.1, *[The woman is] holding a puppy and looks happy* clearly captures the core event, while *She is wear a blue dress* is irrelevant to the event happening.

2. **Verifiability**: Does the response contain only information that is true and verifiable based on the image? Inferences should not be speculations and are allowed only when necessary and highly probable, as when describing a familial or professional relationship between persons depicted in the image. For example, in Table 4.1, *She is wear a blue dress* conveys information that is irrelevant to the core event but is nonetheless recoverable from the image (core event=0, verifiability=1), while *hugging her dog Fluffy that she missed while on vacation* fulfills the core event but also has information that cannot be verified from the picture (core event=1, verifiability=0).

3. **Answerhood**: Does the response make a clear attempt to answer the prompt question? This generally requires a progressive verb. For targeted items, the subject of

22

the question or an appropriate pronoun must be used as the subject of the response. For example, *The dog is so happy!* (Table 4.1) is answering a question other than *What is the woman doing?*.

4. **Interpretability**: Does the response evoke a clear mental image (even if different from the actual item image)? Any required verb arguments must be present and unambiguous. For example, *She loves her pet* (Table 4.1) is too vague to generate a clear mental image. No action is specified (unless we force an unlikely reading of *loves* as a dynamic, simple present verb), and we cannot know if the *pet* is a dog, a horse, etc.

5. **Grammaticality**: Is the response free from errors of spelling and grammar? This is a relatively straightforward feature to annotate. For example, from Table 4.1, *She is wear a blue dress* contains an ungrammatical verb form.

Annotation process The annotation was performed one feature at a time, so that annotators did not have to remember the criteria for multiple features while working through the responses. To facilitate this workflow, I created a simple interface that displays the PDT image and question, along with the current feature name and prompt for the annotator, shown in Figure 4.2. The annotations are written out to a spreadsheet.

Example annotations In Table 4.1, we see example responses with all five features annotated, illustrating each feature's distinctiveness from the others. For example, for *He is eating food* one can generate a mental picture, e.g., of someone chewing (`interpretability`=1), but the pizza is important to the item image (`core event`=0). As another example, *He may get fat eating pizza* seems to be addressing a question about the consequences of the eating action rather than the actual prompt question (`answerhood`=0). Moreover, the response talks about hypotheticals not in the picture (`verifiability`=0). Teasing apart these annotations is the focus of the next section.

Figure 4.2: Interface used for feature annotations. Note that "Not sure" is not a final annotation value; it merely puts the response aside for a later decision.

4.2 Agreement

Two annotators participated in the annotation. Both are native speakers of (US) English, and each has several years of language teaching experience with both children and adult learners. Annotator 1 (A1) annotated the complete corpus. Annotator 2 (A2) annotated only the development set and the test set, data subsets described next.

Three items were used as a development set for creating and revising the annotation scheme. These items were also used as examples in the guidelines. They represent one

What is the boy doing?	C	V	A	I	G
He is eating food.	0	1	1	1	1
eatting.	0	1	1	1	0
The child is about to eat pizza.	1	1	0	1	1
He may get fat eating pizza.	1	0	0	1	1
What is happening?	C	V	A	I	G
Child is eating pizza.	1	1	1	1	0
Tommy is eating pizza.	1	0	1	1	1
The boy's eating his favorite food.	0	0	1	0	1
Pizza is this boy's favorite food.	0	0	0	0	1

Table 4.1: Targeted and untargeted sample responses from the development set transitive item, shown with adjudicated annotations for the five features: core event (*C*), verifiability (*V*), answerhood (*A*), interpretability (*I*) and grammaticality (*G*).

intransitive, one transitive and one ditransitive event. Both annotators annotated portions of the development set multiple times throughout the process, discussing and adjudicating disagreeing annotations before moving on to the test set, which was completed without consultation between the annotators.

The test set parallels the development set and consists of one intransitive, one transitive and one ditransitive item; it is shown in Table 4.2. Agreement and Cohen's kappa scores are

25

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

Table 4.2: The annotation test set items 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.

given in Table 4.3, broken down by different criteria. The following sections will examine the results, comparing verbs types (transitivity), targeted and untargeted items, the five features, and NS and NNS participants.

Set	Total	A1Yes	A2Yes	AvgYes	Chance	Observed	Kappa
Intransitive	2155	0.863	0.855	0.859	0.758	0.978	0.910
Transitive	2155	0.780	0.774	0.777	0.653	0.949	0.853
Ditransitive	2155	0.812	0.786	0.799	0.678	0.924	0.764
Targeted	3390	0.829	0.818	0.824	0.709	0.949	0.823
Untargeted	3075	0.806	0.790	0.798	0.678	0.952	0.872
Core Event	1293	0.733	0.717	0.725	0.601	0.923	0.808
Answerhood	1293	0.834	0.831	0.833	0.721	0.982	0.936
Grammaticality	1293	0.861	0.872	0.866	0.768	0.960	0.827
Interpretability	1293	0.818	0.787	0.802	0.682	0.919	0.744
Verifiability	1293	0.845	0.817	0.831	0.719	0.968	0.884

Table 4.3: Agreement scores broken down by different properties of the test set: total annotations (*Total*), *yes* annotations for Annotator 1 and 2 (*A1Yes*, *A2Yes*), average *yes* annotations (*AvgYes*), total expected chance agreement for *yes*es and *no*s (*Chance*), actual observed agreement (*Observed*) and Cohen's kappa (*Kappa*).

4.2.1 Transitivity

Comparing the intransitive, transitive and ditransitive items reveals an association between agreement and item complexity. The highest raw agreement and Cohen's kappa scores are found with the intransitive item (97.8%, $\kappa = 0.910$) and the lowest with the ditransitive (92.4%, $\kappa = 0.764$).

This is as expected, as ditransitive sentences are longer and have more verbal arguments, making for more opportunities for responses to vary (see Table 3.6), and thus more opportunities for annotators to disagree on a response. This trend also matches annotator feedback: in a follow-up questionnaire, both noted the ditransitive item as the most difficult to annotate overall, and the intransitive as the easiest.

4.2.2 Targeting

Grouping the annotations into targeted and untargeted sets, the raw agreement scores are comparable (94.9% vs. 95.2%). However, despite a greater degree of response variation, the untargeted group has a higher kappa score (0.872 vs. 0.823). When asked to compare the annotation process for targeted and untargeted items, A2 noted that targeted responses require more concentration and closer consultation of the guidelines. For example, `answerhood` does not allow for targeted responses to modify the subject provided in the question in any way, whereas in answering *What is happening?*, the respondent is free to speak of characters in the pictures in many different ways. Both A1 and A2 thus describe the annotation of untargeted items as less restrictive and less time-consuming.

4.2.3 Features

Grouped by feature, the annotations all show raw agreement scores above 91% and Cohen's kappa scores above 0.74 (Table 4.3). For future use of this corpus in content assessment, these kappa scores are comfortably above the 0.67 suggested as a threshhold for meaningful, reliable agreement (Landis and Koch, 1977; Artstein and Poesio, 2008). I discuss each

27

feature in turn here, highlighting difficulties in coming to an agreement, as such disagreements illustrate some of the impactful ways in which responses vary.

Core event Isolating whether the main content of the picture is being described or not, the `core event` feature is the most relevant of the five for content assessment. All five features are skewed toward *yes* annotations, but with an average *yes* rate of 72.5%, core event is the least skewed; i.e., more responses receive a *no* annotation for `core event` than for any other feature.

`Core event` has the second lowest inter-annotator agreement kappa score, at 0.808. This is somewhat lower than expected, as the pre-adjudication development set score was 0.889. This appears to be largely attributable to the difficulty of the ditransitive item, challenging for both participants and annotators (section 4.2.1).

The main issue in this case has to do with the amount of specificity required to be the core event. The development set item depicts a man delivering a package to a woman, and most responses describe this as such a transaction, using *give*, *deliver* or *receive*. The test set item shows a man giving directions to a woman (Table 4.2), and this resulted in a greater degree of variation. Many (particularly NNS) responses portray this not as a canonical *giving directions* event but as *pointing*, *helping a lost person* or *reading a map*, with A2 more likely to accept these less specific descriptions.

Similarly, but to a lesser extent, the transitive item, which shows a woman hugging a dog (Table 4.2), resulted in disagreements where A2 accepts the word *pet* as the object, but A1 rejects such responses as too vague. Despite the acceptable scores for `core event` agreement, the fact that many disagreements hinge on particular word choice or annotators having minor differences in interpretation of the event suggest that greater agreement could be achieved by providing annotators with suggestions about the acceptable content for each response. In other words: by more clearly determining the desired level of specificity of a response—for the verb or its arguments—agreement could be higher. The desired speci-

ficity may vary in accordance with the intended use of the annotations; in the current annotations, the standard discussed between annotators and in the guidelines included pragmatic considerations like naturalness, native-likeness and effort.

Verifiability   On the flipside of the question of whether the core semantic content is expressed is the question of whether any extraneous content is added, or any content used in a way which cannot be verified from the picture. The average *yes* rate for `verifiability` is 83.1%, making it the third most skewed feature.

The raw agreement score is 96.8%, and the kappa score is 0.884. By both measures, this is the second highest agreement score, after `answerhood`. Of 42 disagreements for `verifiability`, annotators agree that at least eight are avoidable. Of these, five involve the incorrect use of plurals. For example, A1 accepted *A man is pointing the way for the women*, when the image shows only one woman, but the guidelines reject such responses. Two other errors stem from inaccuracy, with respondents referring to a dog in the illustration as a cat. Each annotator incorrectly accepted one such response. One disagreement involved the misspelling of a crucial object: *The woman is holding the pat*. It is unclear whether *pet* or *cat* was intended. This should render the response unverifiable, but A1 accepted it.

The remaining disagreements are attributable to different opinions about inferences, with A2 being, in general, more strict. For the ditransitive item, for example, both annotators accept responses that refer to the woman as a *hiker*, but only A1 accepts responses where the man and woman are collectively referred to as hikers. For the intransitive item depicting a woman running, A1 accepts multiple responses that refer to this as *a race*, as well as responses that infer the runner's motivation (fitness, leisure, etc.). I believe such differences are unavoidable in this annotation task. Adding more detail to the guidelines might help reduce disagreements about inferences, but the guidelines are nearly 40 pages and expanding them to cover various contingencies would certainly add to annotator de-

mand and fatigue.

Answerhood Capturing the semantic content of the picture isn't the only criterion for determining the quality of a response; the `answerhood` feature was added largely as a way to identify responses that simply do not follow the instructions. Such responses tend to fall into one of the following categories:

1. Responses that do not directly answer the given question, perhaps by reframing the perspective so that it seems like a different question was asked, e.g., *He may get fat eating pizza*, in response to *What is the boy doing?* (Table 4.1);

2. Responses that are gibberish or very low-effort and entered only so the participant can proceed to the next item, e.g., *Hey man*;

3. "Troll" responses that attempt to be clever or obscene at the cost of attempting a direct answer, e.g., *How is the pizza staying perfectly horizontal when the boy is holding it so close to the tip?*, in response to *What is happening?* (Table 4.1).

The majority of participants do attempt to follow the instructions and answer the question, however, and it is unsurprising that this feature skews strongly toward *yes* annotations and results in the highest raw agreement (98.2%) and kappa (0.936) scores among the five features.

Of 23 disagreements, seven stem from one annotator failing to enforce the requirement that a targeted response subject be either an appropriate pronoun or the exact subject given in the question, without adjectives, relative clauses or other modifiers. Given the question *What is **the woman** doing?*, for example, the responses *The **lady** is running* and *The woman **who in pink** is running* were incorrectly accepted by one annotator each. While this criterion may seem strict, this subject-identity rule separates the task of identifying an attempt to answer the question from the task of verifying information (see `verifiability` above).

Another ten disagreements involve responses lacking a progressive verb, generally required as an indication that the response refers to the specific action in the image and does not merely describe a state or a general truth (cf., e.g., *The woman is running* vs. *The woman runs*). Annotator fatigue thus appears to account for the majority of `answerhood` disagreements.

Interpretability   The average *yes* rate for `interpretability` is 0.802; only `core event` is less skewed: responses were thus also more likely to be unacceptable. The raw agreement score is 91.9% and kappa is 0.744, the lowest scores among the five features. This was anticipated, because `interpretability` is perhaps the most difficult to define, leaving room for annotators' personal judgments. Annotators must decide whether a given response evokes a clear mental image, regardless of how well that mental image matches the PDT image. In this way, responses such as *The man is working* which may be completely `verifiable` may still fall short, in that the man could be picking fruit, building a bridge, and so forth.

The guidelines place some restrictions on what it means to be a clear mental image. To begin with, if one were to illustrate the response, the result would be a complete, representational, canonical image. It would not be necessary to guess at major elements, like subjects or objects. All necessary semantic arguments would be identifiable from the sentence and thus not obscured or out of the frame in the mental image. Vague language should be avoided, but human gender does not need to be specified, especially when a non-gendered word like *doctor* or *teacher* is natural.

Consider a response like *A woman is receiving a package*. By these criteria, the response is annotated as 0 because the person or entity delivering the package is not specified, and an illustrator would need to either guess or compose the image with the deliverer conspicuously out of the frame. *A man is delivering a package*, on the other hand, would be accepted. An illustrator could simply show a delivery person carrying a package, as an

indirect object is not necessary for the verb *deliver*.

Among the 105 annotator disagreements, fatigue accounts for roughly 30; this is diffi-cult to determine precisely because annotators expressed difficulty in identifying a single root cause for many disagreements. Those that are clearly attributable to annotator error tend to involve responses with some internal inconsistency, as with subject-verb disagree-ments, where the number of the subject is uninterpretable. Among true disagreements, the level of specificity is often the point of contention, as with `core event`. For example, A1 accepted several transitive item responses with the verb *love*, as in *The woman loves her dog* (Table 4.2). A2 argued that these are too vague to illustrate as an action, but A1 disagreed. This disagreement may also hinge on differing judgments regarding the use of *love* as a dynamic verb, and such idiolectal differences are an unavoidable source of noise in annotation work (citation?). As mentioned above (see `Verifiability`), expanding the guidelines might help cover some such situations, but likely at the cost of increased annotator fatigue.

LK: citation?

Grammaticality   The `grammaticality` feature is the most heavily skewed one, with an average *yes* rate of 86.6%. As the only non-semantic annotation, this is perhaps not surprising.

Grammaticality has a raw agreement score of 96.0% and a kappa of 0.827. Among 52 disagreements, annotators concurred in discussion that 19 involve an avoidable annotator error. These are primarily responses with typos, misspellings, subject-verb disagreement and bare nouns, all rejected by the annotation rules. Such cases are likely attributable to annotator fatigue.

The remainder reflect an unavoidable level of disagreement. Many of these stem from differing interpretations of bare nouns as either errors or as acceptable mass nouns, as in *The man is giving **direction** to the tourist*. In several cases, annotators disagree over prepositions, which are known to be a common source of disagreement and pose special

challenges in the context of learner language (Tetreault and Chodorow, 2008a,b). For example, annotators could not agree on the grammaticality of the prepositions in *The girl is asking for help **to** the man* and *The girl is hugging **with** her cat*.

4.2.4 NS & NNS responses

Response quality and annotation agreement were also calculated separately for NS and NNS responses, as shown in Table 4.4. The average rate of *yes* annotations is used here as an indication of response quality. Comparing this *yes* rate shows that the NNSs outperform the NSs by between roughly 8% and 12% on all features except `grammaticality`. It is not surprising that NSs outperform NNSs on this feature (90.2% to 79.3%), but to account for their superior performance on the other features, one must consider the fact that the NNSs were recruited from ESL courses and performed the task with peers and researchers present. The NNSs were more likely to make a good faith effort than the NSs, the majority of whom performed the task anonymously and remotely. Furthermore, with twice as many responses to provide for each item for NSs, fatigue and boredom may have been a contributing factor.

	Average Yes		Chance Agree		Observed Agree		Kappa	
Set	NS	NNS	NS	NNS	NS	NNS	NS	NNS
Core	0.686	0.805	0.569	0.686	0.922	0.927	0.819	0.767
Answer	0.800	0.899	0.680	0.819	0.977	0.993	0.928	0.961
Gramm	0.902	0.793	0.823	0.671	0.962	0.955	0.786	0.863
Interp	0.764	0.881	0.638	0.789	0.910	0.936	0.752	0.697
Verif	0.807	0.882	0.687	0.791	0.970	0.962	0.904	0.819

Table 4.4: Comparing feature annotation agreement scores for NSs and NNSs: average *yes* annotations (*Average Yes*), total expected chance agreement (for *yes*es and *no*s) (*Chance Agree*), actual observed agreement (*Observed Agree*) and Cohen's kappa (*Kappa*).

Turning to the question of annotation quality, raw agreement scores are high among both groups, ranging from 91% to 99.3%. Notably, for `core event`, `verifiability` and `interpretability`, kappa scores are higher for NS responses than for NNS ones;

i.e., annotators agree more on NS responses for these features. It may be no coincidence that these three features are the most closely tied to meaning, while `answerhood` gets at pragmatics and `grammaticality` focuses on form correctness.

The lower kappa score for NS `answerhood` is also attributable to task effects, as a second response (as required of NSs) is more likely to be off topic or in bad faith. For `grammaticality`, kappas for annotator agreement are higher for NNS responses. A relatively low rate of expected (chance) agreement contributes to this fact. Additionally, annotators note that many grammar problems with NNS responses are obvious (e.g., *The man who in yellow is showing the way to a girl*, see Table 4.2), but the few grammar problems in NS data are mostly typos and more easily overlooked (e.g., *The man is giving ditections*).

4.3 Establishing Feature Weights

The five annotation features were chosen for their relevance to the construct of response "goodness" for the picture description task (PDT). However, we cannot assume that these binary features bear equal weight in determining the quality of a response. Certainly `core event` is more important than `grammaticality`, for example. Thus the annotations alone cannot be used to assign scores to responses, a crucial necessity in order to rank responses and evaluate my approach to content analysis.

Clearly, weights must be assigned to each feature. One way to derive weights would be to manually rank responses from best to worst, then use the distribution of annotations across this ranking to determine some coefficient that represents the importance (weight) of each feature in the rankings. For each task item, the corpus contains roughly 150 NS responses and 70 NNS responses, so producing a manual ranking of the full set of responses is highly impractical. Manually ranking even a subset of 10 or 20 responses is frustrating and unreliable. Ranking a single pair of responses is a much more practical task, so, I decided to have annotators perform an A/B test with pairs of responses. With enough of

these A/B decisions, it becomes possible to derive annotation weights.

The full corpus consists of 13,533 responses across 60 items (30 images presented with two prompts each; see Chapter 3.3). For the A/B test to determine feature weights, a sample of 1200 response pairs was used – 20 response pairs per item. Among the response annotations ([`core event`, `answerhood`, `grammaticality`, `interpretability`, `verifiability`]), some vectors are more common than others; *perfect* annotations ([1, 1, 1, 1, 1]) and those with grammar problems only ([1, 1, 0, 1, 1]), for example, are frequent, while responses annotated positively only for `interpretability` and `verifiability` ([0, 0, 0, 1, 1]) are far less frequent. Thus, to maximize the informativeness of the A/B tests, for each item, no annotation vector was represented multiple times in the sample until every unique vector in the item responses was included once. Moreover, no A/B pair contained responses with identical vectors; we learn nothing by comparing two *perfect* responses, for example.

Annotator 1 (A1) performed the A/B test for all 1,200 of the sampled response pairs. Annotator 2 (A2) performed the A/B test for a subset of 300 response pairs, for the purpose of measuring inter-annotator agreement. These are the same annotators from the feature annotation task, discussed in Section 4.2.

Annotators were given the following instructions for the A/B test:

> You will be presented with picture description task items and pairs of sample responses. Your task is to decide which of the two responses in each pair is the best response for the accompanying image and question. For our purposes, a good response is relevant and reasonable given the prompt. While you should consider form, please prioritize communicativeness and content. Naturally, you may consider what you know about the previously annotated features, but do not overthink them. These features are not of equal importance. A quick decision based on your own experience and intuition about communication is the goal here. If you feel that the responses are equally appropriate to the task,

or if you cannot decide which is better, you may choose the "same/unsure" option, but please do so sparingly.

Figure 4.3: Annotation interface used for the A/B response quality task.

The A/B task interface (Figure 4.3) was similar to that used for annotating the features. For each A/B decision, a pair of responses along with the item image and question were

presented to the annotator. The annotations for the responses were not included, but given their familiarity with the feature annotation, the annotators could probably determine the value for each feature.

Example response pairs and A/B decisions are shown in Table 4.5. For the first pair, both annotators preferred *The boy is carrying groceries* over *The boy carries the bag*. While the annotation features were not directly used during the A/B task, we can infer here that the present progressive *is carrying* is preferable to the simple present *carries*, and indeed it more directly answers the question *What is happening?* and thus better satisfies the `answerhood` feature. The use of the more descriptive *groceries* over *bag* also likely contributes to the preference, and arguably this better satisfies the `core event` feature.

Response	A1	A2	Agree
A: The boy carries the bag. B: The boy is carrying groceries.	B	B	yes
A: The boy is holding a box of fruits. B: A boy carries a bag.	B	Same	no
A: Little boy Towing the grocery to the car B: The boy is excited about his bag of groceries.	A	B	no

Table 4.5: A/B response quality task sample responses pairs, annotator decisions (*A1 & A2*) and agreement for the item shown in Figure 4.3.

LK: add annotation vectors to table

The disagreements shown in the table are representative of the majority of the 35/300 disagreements in the sample in that one could make a reasonable argument for preferring either response (or marking them *same* in quality). Disagreement over *The boy is holding a box of fruits* and *A boy carries a bag* (Example 2) seems to involve the weighing of issues related to `answerhood` (*is hold**ing*** versus *carries*), `core event` (i.e., the descriptiveness of *a box of fruits* versus *a bag*) and `verifiability` (with *box* being quite clearly inaccurate). The disagreement in Example 3 involves similar `answerhood` issues as well as potential concerns related to `grammaticality` (e.g., Response A is a sentence fragment and contains a bare noun). Moreover, *is excited about* in Response A would likely not satisfy the `core event` feature, while in Response B, *towing* is a questionable

verb choice, and *to the car* would arguably violate `verifiability` because the image contains no car.

Agreement was calculated for the 300 response pairs judged by both annotators, presented in Table 4.6. The agreement rate of 0.883 with a Cohen's kappa of 0.692 confirms that high agreement on this task is both possible and reliable (Landis and Koch, 1977; Artstein and Poesio, 2008). With these scores, I am confident in using the full set of Annotator 1's 1,200 A/B decisions to derive the feature weights.

Chance Agree	Observed Agree	Kappa
0.621	0.883 (265/300)	0.692

Table 4.6: Response quality A/B task agreement scores for two annotators on a sample of 300 responses pairs; chance agreement, observed agreement and Cohen's Kappa.

To calculate the weights, the total number of times a feature occurred with the dispreferred response in an A/B pair was subtracted from the total number of times that feature occurred with the preferred response to yield the net count for that feature. A/B pairs ruled *same* (no preference) were omitted. The net counts of all five features were summed. The net count for each feature is then divided by this total net sum to yield the weight—this represents the degree to which each feature contributes to a response's quality. The sum of the weights is 1.0. The counts and weights are shown in Table 4.7.

	Core	Answer	Grammar	Interpret	Verif	Total
Tot. Pref.	944	807	910	1021	1026	4708
Tot. Dispref.	367	660	822	667	611	3127
Net Pref.	577	147	88	354	415	1581
Weight	0.365	0.093	0.056	0.224	0.263	1.0

Table 4.7: Annotation counts and weights for each feature, based on a sample of 1,200 response pairs (of which 87 pairs were marked "same" and thus omitted). *Tot. Pref. & Tot. Dispref.* are the number of times the feature occurred with the preferred or dispreferred response. Each weight is the feature's net preferred count divided by the total net preferred count (for all five features) of 1581.

The weights yielded from the A/B task are well aligned with my intuitions about the

features and their importance in the PDT and seem to support this work's ethos of content and communication over form. The features that relate closely to meaning carry the most weight. `Core event`, which directly addresses the focus of the image, ranks well above the other features in terms of weight. `Verifiability`, which limits the scope of response content, and `interpretability`, which addresses a response's ability to communicate content, have similar weights that indicate a medium degree of importance. Finally, `answerhood`, which deals with discourse and pragmatics, and `grammaticality`, which only addresses surface forms, carry much lesser weights, as expected. As this work is rooted in an ethos of steering ICALL toward content and communication over form, the

LK: HUH?
LK: refer back to the AB task instructions

4.4 Annotation Conclusion

The SAILS corpus presented here was developed with specific research in mind, but also in the hopes that it may be used to address a broad range of questions. I have demonstrated here a set of binary features that were successfully implemented with reliable levels of inter-annotator agreement. These features were defined with an eye toward content analysis and ICALL, but I believe the annotations and raw responses could find uses in question answering, dialog systems, pragmatic modeling, visual references and other challenges in natural language processing. The feature set could also be expanded to better suit other purposes, and the task could easily be extended to include new items. It could also be improved by adding participants from a wider variety of L1s. Guidelines, task materials and annotation tools are included with the corpus.[1]

A number of lessons have been learned in this process, and as I intend this work to be extendable, a few suggestions are in order. The inclusion of any symbols or numerals in items should be avoided as they resulted in response complications; some participants gave clever "meta" responses (*She's breathing in music notes*, rather than *She's singing*), and others focused on the symbols rather than the abstract concepts they represent (*The teacher*

---

[1]https://github.com/sailscorpus/sails

is teaching '2 + 2 = 4', rather than *The teacher is teaching math*). The comparison of crowdsourced NS data with the data of known NS participants and the NNS student data makes it clear that motivations and task environment can affect the quality of responses.

Additionally, more clearly defining acceptable `core events` could lessen the ambiguity for annotators. While I intend the NS responses collected here to be useful for comparing with NNS responses and addressing related research questions, for specific applications like language testing, the use of expert annotators and constructed reference materials or gold standards may be more appropriate or cost effective (see, for example, Somasundaran and Chodorow (2014)).

CHAPTER 5

METHOD

5.1 Introduction

The method used to analyze picture description task (PDT) responses throughout this dissertation represents an evolution from my own earlier attempts. In this chapter, I summarize this work and explain the current method.

In short, my earlier approach assessed each non-native speaker (NNS) response by extracting a *verb(subject,object)* triple and looking for a match among triples from the native speaker (NS) responses. This involved dependency parsing the sentence then applying custom rules based on the labels, relationships and parts of speech in order to find each element of the triple. This process found moderate success, correctly assessing roughly half of NNS responses with a very small number of NS responses. Considerable weaknesses emerged, however; the rule based approach meant that it was limited in its ability to handle variation, and the use of simple triples was a hacky simplification of meaning. This lead me to the current approach, which uses a more robust representation of meaning and replaces the rule based matching with measures of semantic similarity. These changes make for a more generalizable approach.

5.2 First approaches: Rule based semantic triple matching

This section summarizes relevant work first presented in King and Dickinson (2013) and King and Dickinson (2014); please see those papers for deeper discussion. Like the current research, my previous work focused on analyzing NNS responses to a PDT by comparison with NS responses. I did not know a priori if such a task would be within reach for a single researcher using off the shelf tools, so my initial work sought to uncover challenges and

determine whether variation in the form and content of responses could be manageable.

I knew this would require some constraint of the responses, and without a suitable dataset available, I decided to develop my own. This research is motivated by a desire to see intelligent computer-assisted language learning (ICALL) applications move toward natural communication in game-like visual contexts, so I determined that a picture description task (PDT) would produce a fitting dataset. Research in second language acquisition (SLA) often relies on the ability of task design to induce particular linguistic behavior (Skehan et al., 1998), and the PDT should induce context-focused communicative behavior. PDT data allows one to investigate pure interlanguage without the influence of verbal prompts and shows learner language being used to convey meaning and not just manipulate forms. I knew that I would be relying on some kind of rule based approach for extracting *verb(subject,object)* triples from syntactic dependency parses. Thus, for this experiment, I chose or developed each of the visual stimuli because it presents an event that I believe to be transitive in nature and likely to elicit responses with an unambiguous subject, verb and object, thereby imposing some restrictions on form and content.

The PDT consists of 10 items (8 line drawings and 2 photographs) intended to elicit a single sentence each; an example is given in Figure 5.1. Participants were asked to view the image and describe the action in either past or present tense. Responses were typed by the participants themselves in a computer lab with spell checking disabled. I collected responses from 53 participants for a total of 530 sentences. There were 14 NSs (non-linguist university and graduate students) and 39 NNSs (university students enrolled in English as a Second Language courses).

5.2.1 Method

My process in this initial work was to parse a NNS sentence into a dependency representation (section 5.2.2) and then extract a simple semantic form from this parse (section 5.2.3) to compare to a set of gold standard (GS) semantic forms similarly derived from the NS

Response (L1)
He is droning his wife pitcher. (Arabic)
The artist is drawing a pretty women. (Chinese)
The artist is painting a portrait of a lady. (English)
The painter is painting a woman's paint. (Spanish)

Figure 5.1: Example item and NNS responses

responses.

5.2.2 Obtaining a syntactic representation

Because dependency parsing focuses on identifying dependency relations, rather than constituents or phrase structure, it clearly labels the subject, verb and object of a sentence, which can then map to a semantic form (Kübler et al., 2009). In these experiments, I took a naïve approach in which subject, verb and object were considered sufficient for deciding

whether or not a response accurately describes the visual prompt.

As in my current research, I used the Stanford Parser for this task, trained on the Penn Treebank (de Marneffe et al., 2006; Klein and Manning, 2003).[1] Using the parser's options, I set the output to be Stanford typed dependencies, a set of labels for dependency relations. The Stanford parser has a variety of options for the specific ouput, e.g., how one wishes to treat prepositions (de Marneffe and Manning, 2012). I used only two non-default parser options (`CCPropagatedDependencies` and `CCprocessed`[2]) in order to: 1) omit prepositions and conjunctions from the sentence text and instead add the word to the dependency label between content words; and 2) propagate relations across conjunctions. These decisions are important to consider for any semantically-informed processing of learner language.

To see the impetus for removing prepositions, consider the learner example in Figure 5.2, where the preposition *with* is relatively unimportant to collecting the meaning. Additionally, learners often omit, insert, or otherwise use the wrong preposition (Chodorow et al., 2007). The default parser would present a `prep` relation between *played* and *with*, obscuring what the object is; with the options set as above, however, the dependency representation folds the preposition into the label (`prep_with`), instead of keeping it in the parsed string, as shown in Figure 5.2.
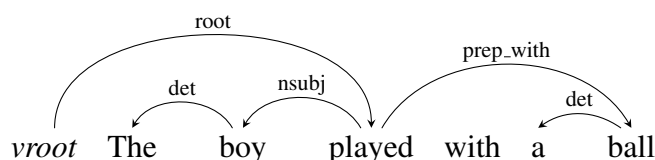
Figure 5.2: Dependency parse showing collapsed preposition dependencies.

This is a lenient approach to prepositions, as prepositions are not without semantic meaning—e.g., *the boy played in a ball* means something quite different from the *with*

---

[1]http://nlp.stanford.edu/software/lex-parser.shtml
[2]http://nlp.stanford.edu/software/dependencies_manual.pdf

example. However, this option makes it moderately easier to compare the meaning to an expected semantic form (e.g., *play(boy,ball)*).

As for propagating relations across conjunctions, this also simplifies the representation somewhat and makes it easier to connect verbs and their arguments, as needed for the semantic form used in comparisons. For a conjunction like *cats and dogs*, for example, the default settings would produce `cc(cats, and)` and `conj(cats, dogs)`, but the chosen settings would collapse this into `conj_and(cats, dogs)`, omitting the dependency that merely labels a conjunction relation between the first conjunct and the conjunction.

Given the rule based approach to matching *verb(subject,object)* triples, many dependency relations are irrelevant for the next step of obtaining a semantic form. For example, in this work I ignored determiner (`det`) relations between a noun and its determiner, allowing for variability in how a learner produces noun phrases.

5.2.3 Obtaining a semantic representation

Sentence types

I categorized the sentences in the corpus into 12 types, shown in Table 5.1. I established these types because each type corresponds to a basic sentence structure and thus has consistent syntactic features, leading to predictable patterns in the dependency parses.

Rules for sentence types

A sentence type indicates that the subject, verb, and object can be found in a particular place in the parse, e.g., under a particular dependency label. For example, for simple transitive sentences of type A, the words labeled `nsubj`, `root`, and `dobj` exactly pinpoint the necessary information. Thus, the patterns for extracting semantic information—in the form of *verb(subj,obj)* triples—reference particular Stanford typed dependency labels, part-of-speech (POS) tags, and interactions with word indices.

Type	Description	Example	NS	NNS
A	Simple declarative transitive	The boy is kicking the ball.	117	286
B	Simple + preposition	The boy played with a ball.	5	23
C	Missing tensed verb	Girl driving bicycle.	10	44
D	Missing tensed verb + preposition	Boy playing with a ball.	0	1
E	Intransitive (No object)	A woman is cycling.	2	21
F1	Passive	An apple is being cut.	4	2
F2	Passive with agent	A bird is shot by a man.	0	6
Ax	Existential version of A or C	There is a boy kicking a ball.	0	0
Bx	Existential version of B or D	There was a boy playing with a ball.	0	0
Ex	Existential version of E	There is a woman cycling.	0	0
F1x	Existential version of F1	There is an apple being cut.	0	1
F2x	Existential version of F2	There is a bird being shot by a man.	0	0
Z	All other forms	The man is trying to hunt a bird.	2	6

Table 5.1: Sentence type examples, with distributions of types for native speakers (NS) and non-native speakers (NNS)

More complicated sentences or those containing common learner errors (e.g., omission of the copula *be*) required slightly more complicated extraction rules, but, since this work examined only transitive verbs, these still boiled down to identifying the sentence type and extracting the appropriate triple. This was accomplished by arranging a small set of binary features into a decision tree to determine the sentence type, as shown in Figure 5.3.

To illustrate, consider the process for the example in Figure 5.4. The sentence is passed through the parser to obtain the dependency parse shown. The parsed sentence then moves to the decision tree shown in Figure 5.3. At the top of the tree, the sentence is checked for an `expl` (expletive) label; having none, it moves rightward to the `nsubjpass` (noun subject, passive) node. Because a `nsubjpass` label is found, the sentence moves leftward to the `agent` node. This label is also found, and because the sentence has reached a terminal node, it is labeled as a type F2 sentence.

With the sentence now typed as F2, specific F2 extraction rules are applied. The logical subject is taken from under the `agent` label, the verb from `root`, and the logical object from `nsubjpass`, to obtain *shot(man,bird)*, which can be lemmatized to *shoot(man,bird)*.
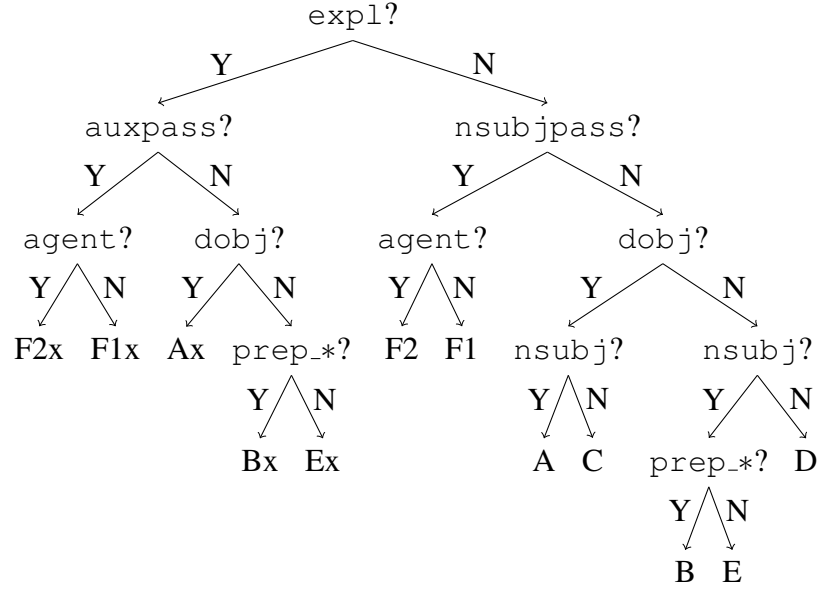
Figure 5.3: Decision tree for determining sentence type and extracting semantic information
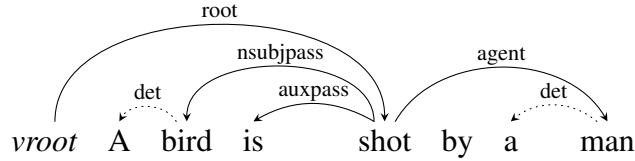
Figure 5.4: The dependency parse of an example NNS response.

This much is possible with relatively little effort in part due to the constraints in the pictures. For figure 5.1, for example, *the artist*, *the man in the beret*, and *the man* are all acceptable subjects, whereas if there were multiple men in the picture, *the man* would not be specific enough.

5.2.4 Evaluation

Evaluating this work required addressing two major questions. First, how accurately does this approach extract semantic information from potentially innovative sentences (section 5.2.6)? Due to the simple structures of the sentences (section 5.2.5), this simple system performs moderately well. Secondly, how many semantic forms does one need in order to

47

capture the variability in meaning in learner sentences (section 5.2.7)? I operationalized this second question by asking how well the set of native speaker semantic forms models a gold standard, with the intuition that a language is largely defined by native speaker usage, so their answers can serve as targets. As we will see, this is a naïve view.

5.2.5 Basic distribution of sentences

		Err. type (Tot. resp.)	Example Sentence	Triple	Count (%)
Triple error	NNS	Speaker	A man swipped leaves.	leaves(swipped,man)	16 (4.1%)
		Parser	Two boys boat.	NONE(boys,NONE)	5 (1.3%)
		Extract	A man is gathering lots of leafs.	gathering(man,lots)	9 (2.3%)
		(390)			**30 (7.7%)**
	NS	Speaker	(*None*)		0 (0%)
		Parser	An old man raking leaves on a path.	leaves(man,path)	2 (1.4%)
		Extract	A man has shot a bird that is falling from the sky.	shot(bird,sky)	8 (5.7%)
		(140)			**10 (7.1%)**
Content error	NNS	Spelling	The artiest is drawing a portret.	drawing(artiest,portret)	35 (9.0%)
		Meaning	The woman is making her laundry.	making(woman,laundry)	23 (5.9%)
		(390)			**58 (14.9%)**
	NS	Spelling	(*None*)		0 (0%)
		Meaning	A picture is being taken of a girl on a bike.	taken(NONE,picture)	3 (2.1%)
		(140)			**3 (2.1%)**

Table 5.2: Triple errors and content errors by subcategory, with error rates reported (e.g., 7.7% error = 92.3% accuracy)

The distribution of sentence types is shown in Table 5.1, broken down between native speakers (NSs) and non-native speakers (NNSs). A few sentence types clearly dominate here: simple declaratives with or without a main verb (types A and C) account for 90.7% of the NS forms and 84.6% of the NNS ones. Adding prepositional forms (types B and D) brings the total to 94.3% and 90.8%, respectively. This suggests in a constrained dataset, a

custom set of rules can provide a high degree of coverage.

5.2.6 Semantic extraction

For the purpose of evaluating this extraction system, I defined two major classes of errors. The first are *triple errors*, responses for which the system fails to extract one or more of the desired subject, verb, or object, based on the sentence at hand and without regard to the target content. Second are *content errors*, responses for which the system extracts the desired subject, verb and object, but the resulting triple does not accurately describe the image (i.e., is an error of the participant's). This work was concerned with reducing the triple errors, as such content errors are expected in NNS data, and in fact, identifying them is an important part of the overall task. Examples are in Table 5.2.

Triple errors are subcategorized as *speaker*, *parser*, or *extraction* errors, based on the earliest part of the process that led to the error. Speaker errors (addressed in detail in Section **??**) typically involve misspellings in the original sentence, leading to an incorrect POS tag and parse. Parser errors involve a correct sentence parsed incorrectly or in such a way as to indicate a different meaning from the one intended; an example is given in Figure 5.5. Extraction errors involve a failure of the extraction script to find one or more of the desired subject, verb or object in a correct sentence. These typically involve more complex sentence structures such as conjoined or embedded clauses.

As shown in table 5.2, we obtain 92.3% accuracy on extraction for NNS data and roughly the same for NS data, 92.9%. However, many of the errors for NNSs involve misspellings, while for NSs a higher percentage of the extraction errors stem only from our hand-written extractor, due to native speakers using more complex structures. For a system interacting with learners, spelling errors are thus more of a priority (cf. Hovermale, 2008).

Content errors are subcategorized as *spelling* or *meaning* errors. Spelling errors involve one or more of the extracted subject, verb or object being misspelled severely enough that the intended spelling cannot be discerned. A spelling error here is unlike those included in
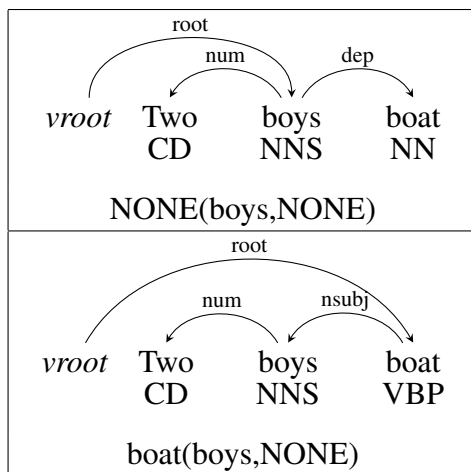
Figure 5.5: A parser error leading to a triple error (top), and the desired parse and triple (bottom).

speaker errors above in that it does not result in downstream errors and is a well-formed triple except for a misspelled target word. Meaning errors involve an inaccurate word within the triple. This includes misspellings that result in a real but unintended word (e.g., *shout(man,bird)* instead of *shoot(man,bird)*).

The goal of a system is to identify the 14.9% of NNS sentences which are content errors; in turn, this could be used to provide explicit feedback or, for example, change the direction of the story in a language tutoring application. Currently, without manual arbitration, the 7.7% triple errors would also be grouped into this set of triple errors, showing the need for further extraction improvements. Also notable is that three content errors were encountered among the NS responses. All three were meaning errors involving some meta-description of the image prompt rather than a direct description of the image contents, e.g., *A picture is being taken of a girl on a bike* vs. *A girl is riding a bike*.

5.2.7 Semantic coverage

Given a fairly accurate extraction system, as reported above, we now turn to evaluating how well a gold standard represents unseen data, in terms of semantic matching. To measure coverage, we take the intuition that a language is defined by native speaker usage, so

Item	NS	NNS	TP	TN	FN	Coverage		Accuracy	
						Ty.	Tok.	Ty.	Tok.
1	5	13	3	1	9	3/12	23/38	4/13	24/39
2	6	13	3	4	6	3/9	15/28	7/13	19/32
3	6	17	5	5	7	5/12	23/30	10/17	28/35
4	4	6	2	0	4	2/6	32/37	2/6	32/37
5	4	24	1	8	15	1/16	3/25	9/24	11/33
6	8	22	3	5	14	3/17	16/32	8/22	21/37
7	7	23	5	4	14	5/19	14/35	9/23	18/39
8	6	23	5	7	11	5/16	10/30	12/23	17/37
9	7	33	3	12	18	3/21	3/23	15/33	15/35
10	5	20	2	12	6	2/8	15/24	14/20	27/36
Total	58	194	32	58	104	32/136	154/302	90/194	212/360
						23.5%	51.0%	46.4%	58.9%

Table 5.3: Matching of semantic triples: *NS/NNS*: number of unique triples for NSs/NNSs. Comparing NNS triples to NS triples, *TP*: number of true positives (types); *TN*: number of true negatives; *FN*: number of false negatives. *Coverage* for *Ty*pes and *Tok*ens = $\frac{TP}{TP+FN}$; *Accuracy* for *Ty*pes and *Tok*ens = $\frac{TP+TN}{TP+TN+FN}$

their answers can serve as targets, and use NS triples as our gold standard. The result is more a measure of nativelikeness than accuracy, which admittedly was a limitation to this approach.

The set of NS responses was manually arbitrated to remove any unacceptable triples (both *triple* and *content* errors), and the remaining set of lemmatized triples was taken as a gold standard set for each item.

Similarly, with the focus on coverage, the NNS triples were amended to remove any triple errors. From the remaining NNS triples, I called an appropriate NNS triple found in the gold standard set a **true positive (TP)** (i.e., a correct match), and an appropriate NNS triple *not found* in the gold standard set a **false negative (FN)** (i.e., an incorrect non-match), as shown in Table 5.4. I used standard terminology here (TP, FN), but because this was an investigation of what *should be* in the gold standard, these were considered false negatives and not false positives. To address the question of how many (NS) sentences are needed to obtain good coverage, I defined **coverage** (=recall) as *TP/(TP+FN)*. As shown in Table 5.3,

I reported 23.5% coverage for unique triple types and 51.0% coverage for triple tokens.

		NNS	
		+	−
NS	Y	TP	FP
	N	FN	TN

Table 5.4: Contingency table comparing presence of NS forms (Y/N) with correctness (+/−) of NNS forms

I defined an inappropriate NNS triple (i.e., a content error) *not found* in the gold standard set as a **true negative (TN)** (i.e., a correct non-match). **Accuracy** based on this gold standard—assuming perfect extraction—is defined as *(TP+TN)/(TP+TN+FN)*.[3] I reported 46.4% accuracy for types and 58.9% accuracy for tokens.

The immediate lesson taken from this was: given a strict matching approach, NS data alone may not make a sufficient gold standard, in that many correct NNS answers are not counted as correct. However, there are a couple of issues to consider here.

First, the approach required exact matching of triples. To maximize coverage, I considered extracting individual subjects, verbs and objects from NS triples and recombining them into the various possible *verb(subj,obj)* combinations. An example of triples distribution and coverage for a single item, along with this recombination approach is presented in Table 5.5.

This modification was not pursued, however, because automating this recombination would require more sophisticated lexical knowledge to avoid adding unwanted triples in the gold standard set. Consider, for example, *do(woman,shirt)*—an incorrect triple derived from the correct NS triples, *wash(woman,shirt)* and *do(woman,laundry)*. Instead, my subsequent work has attempted to improve coverage by prompting NSs to give an initial PDT response, followed by a second alternative.

A second issue that emerged was that, even when only examining cases where the mean-

---

[3]Accuracy is typically defined as (TP+TN)/(TP+TN+FN+FP), but false positives (FPs) are cases where an incorrect learner response was in the gold standard; by removing errors from the NS responses, we prevent this scenario (i.e., FP=0).

Type	NNS	NS	Coverage
cut(woman,apple)	*5*	*0*	*(5)*
cut(someone,apple)	4	2	4
cut(somebody,apple)	3	0	
cut(she,apple)	3	0	
slice(someone,apple)	2	5	2
cut(person,apple)	2	1	2
cut(NONE,apple)	*2*	*0*	*(2)*
slice(woman,apple)	1	1	1
slice(person,apple)	1	1	1
slice(man,apple)	1	0	
cut(person,fruit)	1	0	
cut(people,apple)	1	0	
cut(man,apple)	1	0	
cut(knife,apple)	1	0	
chop(woman,apple)	1	0	
chop(person,apple)	1	0	
slice(NONE,apple)	0	2	
Total	30	12	10 *(17)*

Table 5.5: Distribution of valid tokens across types for a single PDT item. Types in italics did not occur in the NS sample, but could be inferred to expand coverage by recombining elements of NS types that do occur.

ing is literally correct, NNSs produced a wider range of forms to describe the prompts than NSs. For example, for a picture showing what NSs overwhelmingly described as a *raking* action, many NNSs referred to a man *cleaning* an area. Literally, this may be true, but it does not align with a NS based GS. This behavior was somewhat expected, given that learners are encouraged to use words they know to compensate for gaps in their vocabularies (Agustín Llach, 2010). This also parallels the observation in SLA research that while second language learners may attain native-like grammar, their ability to use pragmatically native-like language is often much lower (Bardovi-Harlig and Dörnyei, 1998). These findings highlighted the need for a more flexible approach that considers how native-like a sentence is as well as how appropriate its meaning is.

LK: Work a D Stringer citation in here?

I followed up this work with an expanded approach that relied on language models and spell checking tools to attempt to identify and fix misspellings that lead to downstream

Triple	Example sentence
shoot(man, bird)	A man just shot a bird.
shoot(man, fowl)	The man shoots the fowl.
shoot(man, duck)	A man just shot a duck.
shoot(hunter, bird)	The hunter has shot a bird.
shoot(he, bird)	He shot the bird down!

Table 5.6: The NS gold standard for Item 10.

problems (King and Dickinson, 2014). I omit this discussion because it is not applicable to the current work; I now take a simpler approach — respondents use spell checking during the task. This is because in most contexts where my system would be used, like a language tutoring application or game, spelling instruction is not really the goal, and a built in spell checker would likely be available. Moreover, omitting this step removes a layer of analysis (a potential source of errors), and allows the current study to focus more directly on meaning.

5.3 Recent work: More general, distance-based approaches

In subsequent work (King and Dickinson, 2016), I moved towards finding a "sweet spot" of semantic analysis (cf. Bailey and Meurers, 2008) for image-based learner productions. In particular, using available NLP tools, I moved away from specific correct semantic representations and an exact definition of correctness, to more abstract data representations and more gradable notions of correctness (section 5.3.1). A benefit of more abstract representations is to allow correct and relevant information to be derived from a relatively small set of native speaker responses, as opposed to deriving them by hand, in addition to allowing for a range of sentence types.

I should note, in this context, that I am discussing semantic analysis given a GS of native sentences. Image processing tasks can often rely on breaking images into semantic primitives (see, e.g., Gilberto Mateos Ortiz et al., 2015, and references therein), but for learner data, I want to ensure that I can account not just for correct semantics (the *what* of

a picture), but natural expressions of the semantics (the *how* of expressing the content). In other words, the goal is to reason about meaning based on specific linguistic forms.

A second issue regarding semantic analysis, beyond correctness, stems from using an incomplete GS. The productive nature of language means that a sentence can be expressed in countless ways, and thus a GS can never really be "complete". Examining the degree of this variability both for NSs and NNSs is necessary to determine whether a crowd-sourced GS can account for a sizable portion of test responses. Analyzing variability can also help determine the most effective parameters for an NLP system for image based responses. Additionally, it can offer insights into theoretical research on variability in learner language (cf. Ellis (1987), Kanno (1998)).

5.3.1 Generalizing the Methods

The previous work assumed that the assessment of NNS responses involves determining whether the GS contains the same semantic triple that the NNS produced, i.e., whether a *triple* is *covered* or *non-covered*. In such a situation the GS need only be comprised of *types* (not *tokens*) of semantic triples. But the GS is comprised of the small set of NS responses and is thus incomplete—meaning that exact matching is going to miss many cases, and indeed in King and Dickinson (2013), coverage was only at 50.8%. Additionally, relying on matching of triples limits the utility of the method to specific semantic requirements, namely transitive sentences. By changing to "bags of dependencies" and tallying the counts of dependencies in the NS responses comprising the GS, I moved into a gradable, or ranking, approach to NNS responses.

My goal is to emphasize the degree to which a response conveys the same meaning as the GS, necessitating an approach which can automatically determine the importance of a piece of information in the GS. In Section 5.3.2, I detail how I represented the information and in Section 5.3.3, I discuss comparing NNS information to GS information, which

Response (L1)
The man killing the beard. (Arabic)
A man is shutting a bird. (Chinese)
A man is shooting a bird. (English)
The man shouted the bird. (Spanish)
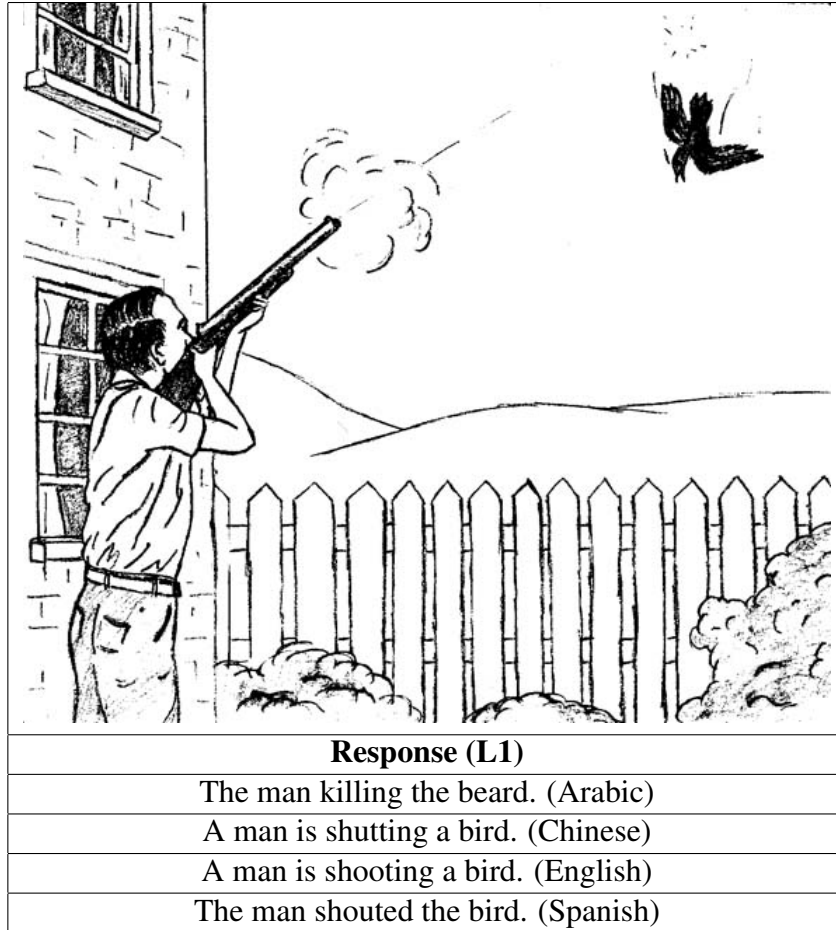
Figure 5.6: Example item and responses

allowed me to rank responses from least to most similar to the GS.[4]

This work used the same 10 item PDT dataset described in section 5.2. Another example is shown in Figure 5.6.

5.3.2 Representation

To overcome the limitations of an incomplete GS, I represented each response as a list of *terms* taken from the dependency parse (de Marneffe et al., 2006), the terms referring to individual dependencies (i.e., relations between words). This eliminates the complications of extracting semantic triples from dependency parses, which only handled a very restricted

---

[4]Although rankings often go from highest to lowest, I prioritize identifying problematic cases, so I rank accordingly.

set of sentence patterns and resulted in errors in 7–8% of cases (King and Dickinson, 2013). Operating directly on individual dependencies from the overall tree also means the system can allow for "partial credit"; it distributes the matching over smaller, overlapping pieces of information rather than a single, highly specific triple.

Specifically, representations took one of five forms. I first tokenized and lemmatized a response to a list of dependencies that represents the response. The five term representations are then variations on dependencies. The full form concatenates the label, head and dependent, as in `subj#boy#kick`. I call this **ldh** (label, dependent, head). The remaining four forms abstract over either the label, head and/or dependent, as in `X#boy#kick`. I refer to these forms as **xdh**, **lxh**, **ldx**, and **xdx**. I tested the system performance using each of these term representations separately.

The `xdx` model is on a par with treating the sentence as a "bag of words" (or more accurately, a bag of lemmas), except that some function words not receiving parses (e.g., prepositions) are not included (see section 5.2.2).

5.3.3 Scoring Responses

Taking the term representations from the previous section, my next step was to combine them in a way which ranks responses from least to most appropriate. I scored the responses with one of four approaches, each using one of two methods to **weight** response terms combined with one of two methods to **compare** the weighted NNS terms with the GS.

For weighting, I used either a simple frequency measure or one based on term frequency-inverse document frequency (**tf-idf**) (Manning et al., 2008, ch. 6). I used tf-idf as a measure of a term's importance with the hope that it could reduce the impact of semantically unimportant terms—e.g., determiners like *the*, frequent in the GS, but unimportant for evaluating the semantic contents of NNS responses—and to upweight terms which may be salient but infrequent, e.g., only used in a handful of GS sentences. For example, for an item depicting a man shooting a bird (see Table 5.7 and Figure 5.1), of 14 GS responses, 12 described

the subject as *man*, one as *he* and one as *hunter*. Since *hunter* is relatively infrequent in English, even one instance in the GS should get upweighted via tf-idf, and indeed it that was the effect. This is valuable, as numerous NNS responses used *hunter*.

Calculating tf-idf relies on both *term frequency* (tf) and *inverse document frequency* (idf). Term frequency is simply the raw count of a term, and for tf-idf of terms in the GS, I take this as the frequency within the GS. Inverse document frequency is derived from some reference corpus, and it is based on the notion that appearing in more documents makes a term less informative with respect to distinguishing between documents. The formula is in (3) for a term t, where N is the number of documents in the reference corpus, and df_t is the number of documents featuring the term ($idf_t = \log \frac{N}{df_t}$). A term appearing in fewer documents will thus obtain a higher idf weight, and this should readjust frequencies based on semantic importance.

(3) $tfidf(t) = tf_{GS} \log \frac{N}{df_t}$

After this counting or weighting, the scores are then either **averaged** to yield a response score, or NNS term weights and GS term weights are treated as vectors and the response score is the **cosine distance** between them. This yields:

Frequency Average (FA).   Within the GS, the frequency of each term is calculated. Each term in the NNS response is then given a score equal to its frequency in the GS; terms missing from the GS are scored zero. The response score is the average of the term scores, with higher scores closer to the GS.

Tf-idf Average (TA).   This involves the exact same averaging as with model FA, but now the terms in the GS are assigned tf-idf weights before averaging.

Frequency Cosine (FC).   The frequency of each term is calculated within the GS and (separately) within the NNS response. The term scores are then treated as vectors, and the

response score is the cosine distance between them, with lower scores being closer to the GS.

Tf-idf Cosine (TC).   This involves the exact same distance comparison as with model FC, but now the terms of both the GS and NNS responses are assigned tf-idf weights before comparison.

The two cosine approaches are effectively primitive versions of sentence encoders like the currently popular BERT (Devlin et al., 2018) and Universal Sentence Encoder (Cer et al., 2018). Sentence encoders are a form of language model that learns mathematical representations of words by observing them in context, accounting for things like average distance from a given word type to another given word type. Sentence encodings are thus vectors representing these word values for a full sentence. These approaches result in very high dimensional spaces—imagine a sentence representation that consists of a vector for each word in the sentence, where each vector is a list of average distances from that word type to *every other word type in the language*. Thus sentence encoders typically rely on methods of dimensionality reduction to compress these representations into vectors of manageable length. I say my cosine approaches constitute "primitive" encoders because they omit this step. In the case of my PDT constrained corpus, the number of word types (and in turn dependency types) observed for a given item remains small enough that the raw vectors representing dependencies' tf-idf scores can still be processed easily with a laptop computer. Not only does this simplify the process, it means that the process remains transparent. There are no opaque machine learning processes that derive compressed representations, and each sentence vector could be examined value by value, where each number represents a real dependency. This is important because it leaves the door open for meaningful feedback on each response. For example, one could identify the most salient dependencies in the GS of a given item and somehow present those to a user who gives a low scoring response in order to suggest a more appropriate response. That is to say, a full

encoder could determine how similar a response is to a GS, but it could not tell us *why* it made its determination.

5.3.4 System Parameters

In addition to the four approaches, the five term representations and two sets of parameters, listed below, were varied, resulting in a total of 60 settings for processing responses (see also Table 5.8).

Term form. As discussed in section 5.3.2, the terms can take one of five representations: `ldh`, `xdh`, `lxh`, `ldx`, or `xdx`.

Scoring approach. As discussed in section 5.3.3, the NNS responses can be compared with the GS via models `FA`, `TA`, `FC`, or `TC`.

Reference corpus. The reference corpus for deriving tf-idf scores can be either the Brown Corpus (Kucera and Francis, 1967) or the Wall Street Journal (WSJ) Corpus (Marcus et al., 1993). These are abbreviated as `B` and `W` in the results below; `na` indicates the lack of a reference corpus, as this is only relevant to approaches `TA` and `TC`. The corpora are divided into as many documents as originally distributed (`W`: 1640, `B`: 499). The WSJ is larger, but Brown has the benefit of containing more balance in its genres (vs. newstext only). Considering the narrative nature of PDT responses, a reference corpus of narrative texts would be ideal, but I chose manually parsed reference corpora as they are more reliable than automatically parsed data.

NNS source. Each response has an original version (`NNSO`) and the output of a language model spelling corrector (`NNSLM`). (The current dissertation relies on a corpus for which participants used spell checking at the time of the task, so this offline spelling correction is no longer applicable. In short, it used a spelling tool to find candidate spellings for each

word in a NNS sentence, pruned the lists of candidate words by comparing against words in NS responses, formed new candidate sentences by combining candidate words, and finally chose the most likely sentence by rating each candidate with a trigram word model. I omit the exact details here for brevity, but more can be found in (King and Dickinson, 2014)).

5.3.5 Results

Evaluation metrics

I ran 60 response experiments, each with different system settings (section 5.3.4). Within each experiment, I ranked the 39 scored NNS responses from least to most similar to the GS.

For assessing these settings themselves, I relied on the past annotation, which counted unacceptable responses as errors (see section 5.2.6). As the lowest rank indicates the greatest distance from the GS, a good system setting should ideally position the unacceptable responses among those with the lowest rankings. Thus, I assigned each error-containing response a score equal to its rank, or, if necessary, the average rank of responses sharing the same score.

In Table 5.7, an excerpt of sentence responses is shown for one item, ranked from lowest to highest. To take one example, the third-ranked sentence, *the man is hurting duck*, has a score of 0.996, and it is annotated as an error (1 in the *E* column). Thus, the evaluation metric adds a score of 3 to the overall sum. The sentence ranked 18, by contrast, is not an error, and so nothing is added. In the case of the top rank, two responses with errors are tied, covering rank 1 and 2, so each adds a score of 1.5.

The sum of these scores is taken as the **Raw** metric for that experimental setting. In many cases, one version of a response (NNSO or NNSLM) contains an error, but the other version does not. Thus, for example, an NNSO experiment may result in a higher error count than the NNSLM equivalent, and in turn a higher Raw score. In this sense, Raw scores emphasize error reduction and incorporate item difficulty.

R	S	Sentence	E	V
1	1.000	she is hurting.	1	1.5
	1.000	man mull bird	1	1.5
3	0.996	the man is hurting duck.	1	3.0
4	0.990	he is hurting the bird.	1	3.0
11	0.865	the man is trying to hurt a bird	1	11.0
12	0.856	a man hunted a bird.	0	0.0
17	0.775	the bird not shot dead.	1	17.0
18	0.706	he shot at the bird	0	0.0
19	0.669	a bird is shot by a un	1	19.0
20	0.646	the old man shooting the birds	0	0.0
37	0.086	the old man shot a bird.	0	0.0
38	0.084	a old man shot a bird.	0	0.0
39	0.058	a man shot a bird	0	0.0
		Total (Raw)	17	169
		Average Precision		0.75084

Table 5.7: Rankings for Item 10 from the best system setting (tf-idf cosine scoring, Brown Corpus for tf-idf reference, the language model spelling corrected NNS sentence, and the full label, dependent and head representation; TC_B_NNSLM_ldh) based on average precision scores. *R*: rank; *S*: sentence score; *E*: error; *V*: rank value. Note that not all responses are shown.

However, it is possible that the `NNSO` experiment, even with its higher error count and Raw score, does a better job ranking the responses in a way that separates good and erroneous ones. To account for this, I also used **(mean) average precision ((M)AP)** (Manning et al., 2008, ch. 8), which emphasizes discriminatory power.

For average precision (AP), one calculates the precision of error detection at every point in the ranking, lowest to highest. In Table 5.7, for example, the precision for the first cutoff (1.000) is 1.0, as two responses have been identified, and both are errors ($\frac{2}{2}$). At the 11th- and 12-ranked response, precision is 1.0 ($\frac{11}{11}$) and 0.917 ($=\frac{11}{12}$), respectively, precision dropping when the item is not an error. AP averages over the precisions for all m responses ($m = 39$ for our NNS data), as shown in (4), with each response notated as R_k. Averaging over all 10 items results in the Mean AP (MAP).

(4) $\quad AP(item) = \frac{1}{m} \sum\limits_{k=1}^{m} Precision(R_k)$

Approach		Term Form		Ref. Corpus (TA/TC)		NNS Source	
0.51577	TC	xdh	0.51810	Brown	0.51534	NNSLM	0.51937
0.50780	FC	ldh	0.51677	WSJ	0.50798	NNSO	0.49699
0.50755	TA	lxh	0.51350				
0.49464	FA	xdx	0.49901				
		ldx	0.49352				

Table 5.8: Approaches and parameters ranked by mean average precision for all 10 PDT items.

As mentioned, the Raw metric emphasizes error reduction, as it reflects not just performance on identifying errors, but also the effect of the overall number of errors. MAP, on the other hand, emphasizes finding the optimal separation between errors and non-errors and is thus more of the focus in the evaluation of the best system parameters next.

Best system parameters

To start the search for the best system parameters, it may help to continue with the example in Table 5.7. The best setting, as determined by the MAP metric, uses the tf-idf cosine (`TC`) approach with the Brown Corpus (`B`), the spelling corrected response (`NNSLM`), and the full form of the dependencies (`ldh`). It ranks highest because errors are well separated from non-errors; the highest ranked of 17 total errors is at rank 19. Digging a bit deeper, one can see in this example how the verb *shoot* is common in all the highest-ranked cases shown (#37–39), but absent from all the lowest, showing both the effect of the GS (as all NSs used *shoot* to describe the action) and the potential importance of even simple representations like lemmas. In this case, the `ldh` representation is best, likely because the word *shoot* is not only important by itself, but also in terms of which words it relates to, and how it relates (e.g., `dobj#bird#shoot`).

Table 5.9 shows the five best and five worst system settings averaged across all 10 PDT items, as ranked by MAP. Among the trends that pop out is a favoritism towards `NNSLM` models (i.e., spelling correction). This is due to the fact that higher numbers of errors inflate the MAP scores, and somewhat counterintuitively, the spelling correction module

introduces more errors than it corrects, meaning there are more errors present overall in the `NNSLM` responses than in the `NNSO` responses.[5]

Another feature among the best settings is the inclusion of heads in the dependency representations. In fact, the top 17 ranked settings all include heads (`lxh`, `xdh`, `ldh`); `xdx` first enters the rankings at 18, and `xdx` and `ldx` are common among the worst performers. This is likely due to the salience of the verbs in these transitive sentences; they constitute the heads of the subjects and objects, in relatively short sentences with few dependencies. Furthermore, the tf-idf weighted models dominate the rankings, especially `TC`. It is also clear that for my data tf-idf works best with the Brown Corpus (`B`).

Rank	MAP	Settings
1	0.5534	TC_B_NNSLM_lxh
2	0.5445	TA_B_NNSLM_lxh
3	0.5435	TC_W_NNSLM_lxh
4	0.5422	TC_B_NNSLM_xdh
5	0.5368	TC_B_NNSLM_ldh
56	0.4816	TA_B_NNSO_xdx
57	0.4796	FA_na_NNSLM_ldx
58	0.4769	FC_na_NNSO_lxh
59	0.4721	TA_W_NNSO_xdx
60	0.4530	FA_na_NNSO_lxh

Table 5.9: Based on Mean Average Precision, the five best and five worst settings across all 10 PDT items.

I also summarize the rankings for the individual parameter classes, presented in Table 5.8, confirming the trends in Table 5.9. For a given parameter, e.g., `ldh`, I averaged the experiment scores from all settings including `ldh` across all 10 items. Notably, `TC` outperforms the other models, with `FC` and `TA` close behind (and nearly tied). Performance falls for the simplest model, `FA`, which was in fact intended as a baseline. With `TC`>`FC` and `TA`>`FA`, tf-idf weighting seems preferable to basic frequencies.

Again, the importance of including heads in dependencies is apparent here; the three

---

[5]Note that among the remaining parameter classes, variation does not effect the number of errors.

dependency representations containing heads constitute the top three, with a sizable drop in performance for the remaining two forms (xdx and ldx). Moreover, given the content and narrative style of the PDT responses, it is unsurprising that the Brown Corpus serves as a better reference corpus than the WSJ Corpus for tf-idf. Finally, the NNSLM source significantly outperforms the NNSO source.

Despite the strength of these overall trends, variability does exist among the best settings for different items, a point obscured in the averages. In Tables 5.10 and 5.11, I present the best and worst ranked settings for two of the least similar items, 1 and 5. Their dissimilarity can be seen at a glance, simply from the range of the AP scores (0.05–0.31 for item 1 vs. 0.52–0.81 for item 5), which in itself reflects a differing number of erroneous responses (2 [NNSO] or 6 [NNSLM] for item 1 vs. 23 or 24 for item 5).

For item 1, a drawing of a boy kicking a ball, there is considerable variability in the best scoring approach just within the top five settings: all four approaches (TA, TC, FA, FC) are in the top five. Contrary to the overall trends, I also found the ldx form—without any head information—in the two best settings. Note also that, even though tf-idf weighting (TA/TC) is usually among the best settings, it is occurs among the worst settings, too.

For item 5 in Table 5.11, a drawing of a man raking leaves, the most noticeable difference is that of xdx being among three of the top five settings. I believe that part of the reason for the higher performance of xdx (cf. lemmas), is that for this item, all the NSs use the verb *rake*, while none of the NNSs use this word. For item 1 (the boy kicking a ball), there is lexical variation for both NSs and NNSs.

These types of differences—for these items and others—lead me to explore the clustering of item patterns, in order to leverage these differences and automatically choose the optimal settings for new items; we turn to this next.

Rank	AP	Settings
1	0.30997	TC_B_NNSLM_ldx
2	0.30466	TA_B_NNSLM_ldx
3	0.30015	TA_B_NNSLM_xdh
4	0.29704	FC_na_NNSLM_xdh
5	0.29650	FA_na_NNSLM_ldh
56	0.06474	TC_B_NNSO_ldx
57	0.06174	TC_W_NNSO_ldx
58	0.06102	TA_W_NNSO_lxh
59	0.05603	TA_W_NNSO_xdx
60	0.05094	TA_W_NNSO_ldx

Table 5.10: Based on Average Precision, the five best and five worst settings for item 1.

Rank	AP	Settings
1	0.80965	FA_na_NNSLM_xdx
2	0.80720	TA_B_NNSLM_lxh
3	0.80473	TC_B_NNSLM_lxh
4	0.79438	TC_B_NNSLM_xdx
5	0.78108	TC_W_NNSLM_xdx
56	0.56495	FC_na_NNSO_xdh
57	0.56414	TC_B_NNSO_lxh
58	0.55890	TC_W_NNSO_lxh
59	0.54506	FC_na_NNSO_lxh
60	0.52013	FA_na_NNSO_lxh

Table 5.11: Based on Average Precision, the five best and five worst settings for item 5.

5.4 Current method

The current method relies on a much larger corpus with more detailed annotation. As discussed in Chapter 4, the annotations do not encode for errors, as before, but instead give a binary score for five different features. This means the evaluation cannot take the form of MAP, but must instead compare response rankings, i.e., how well does the system rank responses in comparison to an ideal ranking based on manual annotations? Spearman rank correlation is used to provide these scores.

Because the annotations and evaluation are much different in the current work, it does

not exactly follow that findings from the previous work will hold true. However, I believe that the previous work has highlighted some of the system parameters that are most likely to perform well, and I choose to experiment among the top contenders. For example, all of the current experiments rely on the tf-idf cosine (TC) approach, as this consistently outperformed the others. The frequency average (FA) approach is used as a baseline where LK: Is it? appropriate, as this presents a very simplistic distributional approach as opposed to the more linguistically sophisticated TC. The large increase in the size of the datasets also means that running an exhaustive search for the best parameters among all combinations is not reasonable; as discussed, the previous work used 60 different system settings in total. For that reason, I have chosen to focus on experiments that optimize for each parameter individually, then combine the best parameter settings for use on held out data to ensure they perform optimally. These experiments and their results are discussed in Chapter 6.

CHAPTER 6

EXPERIMENTS

This chapter will discuss experiments in tuning the TC encoder pipeline.

6.1   Normalizing for response length

Each experimental gold standard (XGS) is comprised of some number of native speaker (NS) responses to the picture description task (PDT). Even among native speaker (NS) data, response lengths can vary; some valid responses contain only one or two words, while the longest top out at around 15 words. These longer, well-formed responses are relatively uncommon, but understanding their impact on an XGS is an important step in optimizing the rating process. So far, the approach to each XGS has been to treat it as a "bag of dependencies" in which each dependency contributes equally, meaning longer responses can carry a greater weight in the GS. This has the potential to introduce noise. Table 6.1 illustrates this. These responses are both included in a toy XGS consisting of NS responses with "perfect" feature annotations. The first response contributes four dependencies to the GS, each of which is necessary to fulfill the feature annotations and contributes meaningfully to the GS. The second response, however, contributes 10 dependencies, some of which, like *amod(purple, dress)*, add non-critical detail, and these details could receive too much weight.

0.464012874 0.463905084

To illustrate with the examples in Table 6.1, consider *det(the, girl)*. In the non-normalized setting, this dependency appears as two out of a total 14 dependencies, or 0.143 of the total XGS. In the normalized setting, the dependency appears as one out of four (0.25) dependencies in Response A, and one out of 10 (0.1) dependencies in Response B, equating to 0.175 of the total XGS (0.35 divided by two responses).

LK: is worderased actually processed through tf-idf?

LK: combine these Ps

68

Response A	Response B	Norm. wt.	Non-norm. wt.
The girl is singing	The girl in the purple dress is singing a song		
det(the, girl)	det(the, girl)	0.175	0.143
nsubj(girl, sing)	nsubj(girl, sing)	0.175	0.143
	erased(in, WORDERASED)	0.050	0.071
	det(the, dress)	0.050	0.071
	amod(purple, dress)	0.050	0.071
	prep_in(dress, girl)	0.050	0.071
aux(be, sing)	aux(be, sing)	0.175	0.143
root(sing, VROOT)	root(sing, VROOT)	0.175	0.143
	det(a, song)	0.050	0.071
	dobj(song, sing)	0.050	0.071
4	10	1.0	1.0

Table 6.1: A "toy" XGS consisting of lemmatized syntactic dependencies from only two responses, each with perfect annotation scores. (See Chapter 5 for more on the parsing and lemmatization.)

To account for this and examine whether or not it poses a problem, I conducted experiments in which each *response*, not each *dependency*, contributes equally to the XGS. This meant simply normalizing for the length of the response by applying a weight to each dependency token as it was added to the XGS, where the weight is equal to 1 divided by the total number of dependencies in the response. The responses were then ranked by these scores, and Spearman's rank correlation coefficient ("Spearman") was used to compare the ranking against the "true" GS (TGS), which is the result of ranking the responses using the weighted annotations (discussed in Chapter 4). This experiment was conducted with the largest XGS (*all NS responses*) and the smallest XGS (*all familiar NS responses*) to ensure that the effect of normalization is consistent. This process was repeated for the same data sets without the normalization; the Spearman scores for the normalized and non-normalized GSs were compared.

LK: Get avg # responses for All NS vs Familiar NS

The results of this experiment are shown in Table 6.2. The comparisons show very little difference in the correlations, with only one of the 12 experiments showing a slightly stronger correlation for the normalized GS. The simplest explanation for this is the fact that

	All NS GS		Familiar NS GS	
Dep	Norm	Non-norm	Norm	Non-norm
ldh	-0.474	**-0.477**	-0.462	**-0.471**
xdh	-0.476	**-0.478**	-0.470	**-0.474**
xdx	**-0.441**	-0.438	-0.428	**-0.431**
Avg	-0.463	**-0.465**	-0.454	**-0.458**

Table 6.2: Spearman correlation coefficient using gold standards (GSs) that are normalized for length (number of dependencies) and GSs that are non-normalized. This was conducted for various dependency representations: *label, dependent, head (ldh)*; *dependent, head* (xdh); *dependent* only (xdx). The p-values are not indicated but range between 0.034 and 0.068 for all cases, indicating that the correlations are very unlikely to be coincidental.

longer responses with extraneous information are relatively uncommon, so we can expect this normalization to have a low impact.

NTS: For different test responses, find some examples where the non-normalized model is rated/ranked higher than in the normalized model.

LK: But what else can I say here

Because the non-normalized GSs outperform their normalized counterparts, and because normalization adds complexity to the process, this step was not used in the remaining optimization experiments.

6.2 Dependency formats

One parameter I vary in my pipeline is the format with which I represent each dependency. A dependency consists of a *head*, *dependent*, and *label*. In past work, I experimented with omitting one or more of these elements to allow for less restrictive matching; the results are shown in Table 5.8. In the current dissertation, I compare the system performance using the three formats: *label-dependent-head* (ldh), *dependent-head* only (xdh), and *dependent* only (xdx). In other words, in my "bag of terms" approach, the bags contain either labeled dependencies (ldh), unlabeled dependencies (xdh) or words (xdx). The labeled and unlabeled dependencies were the top performers in my previous work, and the xdx format is included as a kind of baseline showing a bag of words approach.

6.2.1 Results

In all cases, the `ldh` format results in a higher Spearman correlation than `xdh`. As expected, `xdx` performs significantly worse than either `ldh` or `xdh`.

6.3 Targeted vs Untargeted

Here we compare the performance of my ranking system when applied to targeted vs untargeted data.

6.3.1 Results

6.4 Intransitive vs Transitive vs Ditransitive

Here we compare the performance of my ranking system when applied to items that are (predominantly): intransitive, transitive, ditransitive.

6.4.1 Results

6.5 Familiar vs Crowdsourced response XGS

Here we compare how well the system works when using different sources of NSs. (Crowdsourced informants drastically outnumbered Familiar, so this will require some cross-validation – which in turn requires tf-idf for the XGS in each cross-val cycle; i.e., time intensive computation).

6.5.1 Results

6.6 First responses XGS vs First and second responses XGS

Here we compare how well the system works when using NSs' first responses vs a mix of first and second responses. (The latter is nearly double the former, so this will require

some cross-validation – which in turn requires tf-idf for the XGS in each cross-val cycle; i.e., time intensive computation).

6.6.1 Results

6.7 XGS filtered by annotation

Here we examine how performance changes when we filter the XGS to include only "perfect" annotation responses or "core event = yes" responses. (Response counts will vary depending on exactly how I do this, so this will require some cross-validation – which in turn requires tf-idf for the XGS in each cross-val cycle; i.e., time intensive computation).

6.7.1 Results

CHAPTER 7

CONCLUSION

This chapter is concludes the thesis, summing up the work and findings, and suggesting possible future research.
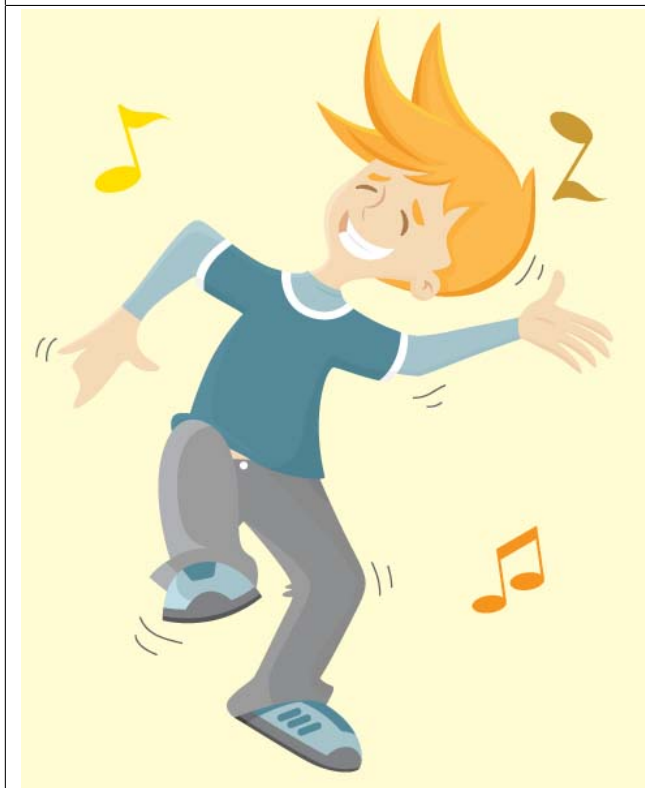
Appendices

The 30 picture description task (PDT) items are shown in this section. Each is displayed with its *targeted* prompt. For all items, the *untargeted* prompt is "What is happening?"

PDT Items

01: What is the boy doing?	02: What is the boy doing?

03: What is the man doing?	04: What is the boy doing?

PDT Items

05: What is the teacher doing?	06: What is the boy doing?

07: What is the bird doing?	08: What is the waiter doing?
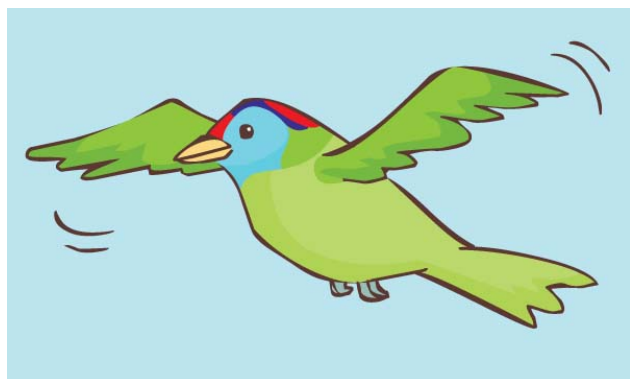
PDT Items

09: What is the girl doing?	10: What is the baby doing?

11: What is the boy doing?	12: What is the woman doing?
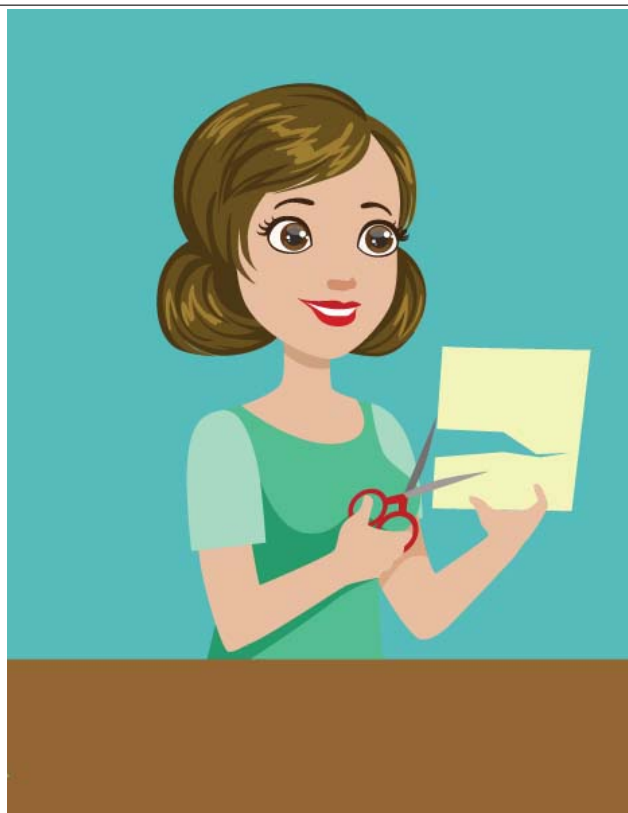
PDT Items

13: What is the man doing?	14: What is the man doing?
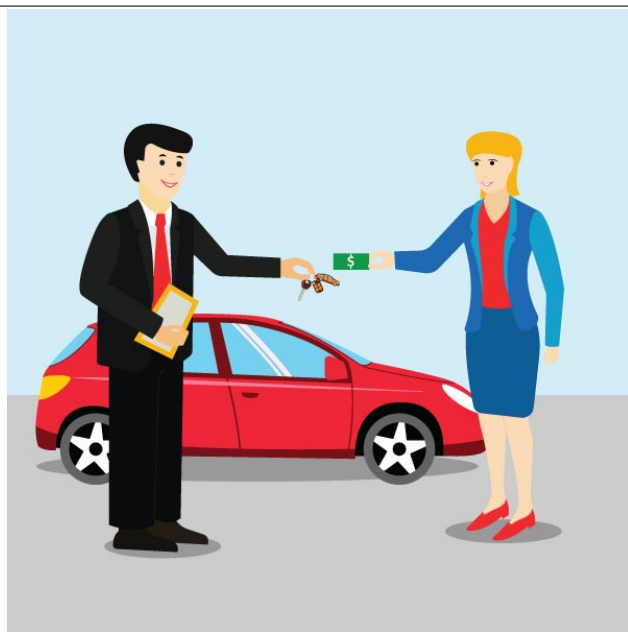
15: What is the man doing?	16: What is the frog doing?
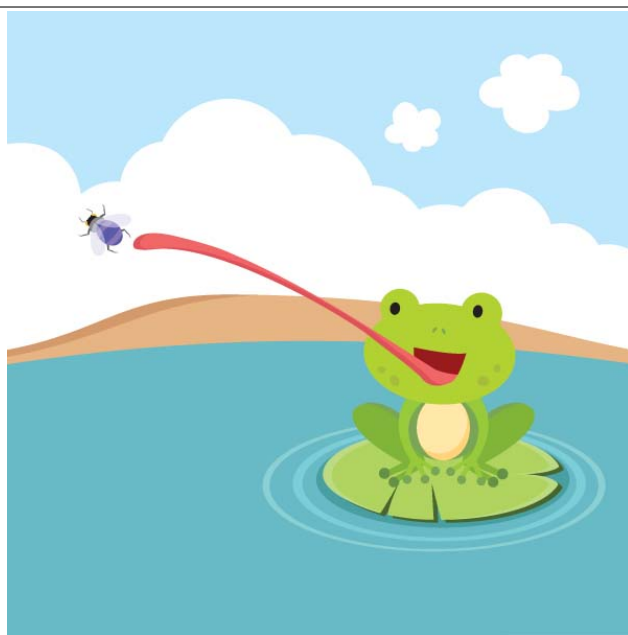
PDT Items
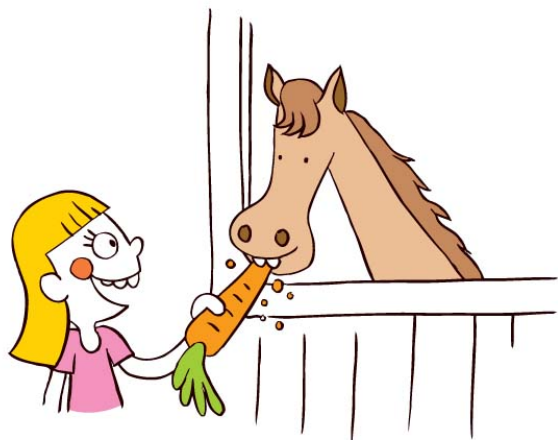
17: What is the girl doing?	18: What is the man doing?

19: What is the woman doing?	20: What is the girl doing?

PDT Items
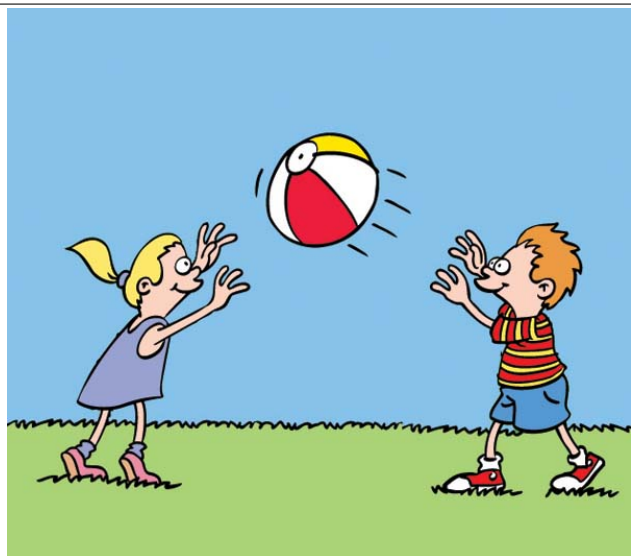
21: What is the boy doing?	22: What is the woman doing?
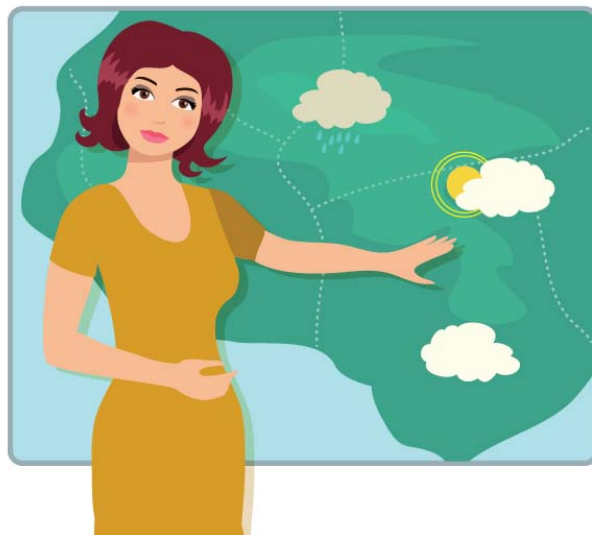
23: What is the doctor doing?	24: What is the boy doing?
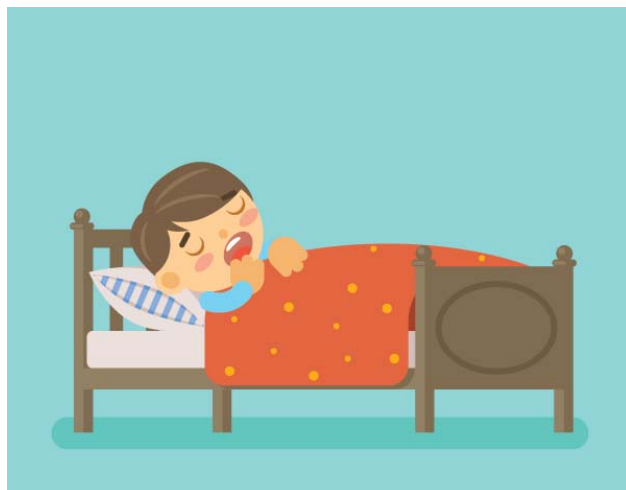
PDT Items

25: What is the dog doing?	26: What is the man doing?
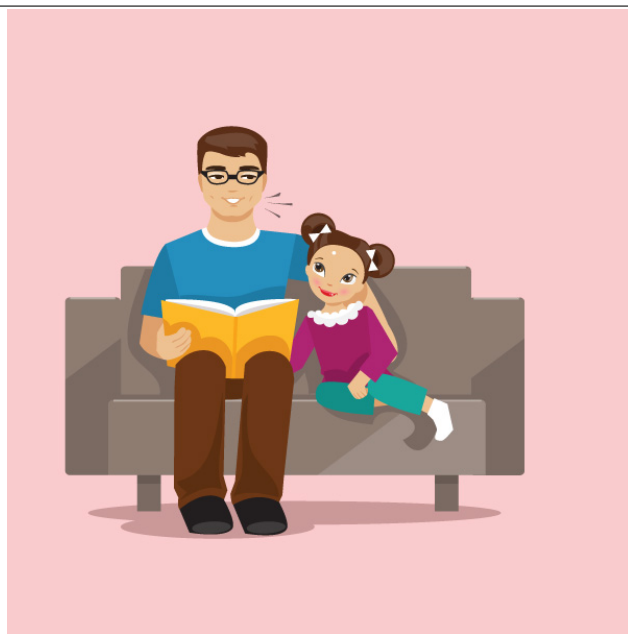
27: What is the girl doing?	28: What is the man doing?

PDT Items

29: What is the woman doing?	30: What is the woman doing?

APPENDIX B

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

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

Figure 1: An example *targeted* item, as presented in the task instructions. The "second sentence" portion is presented to native speakers only.
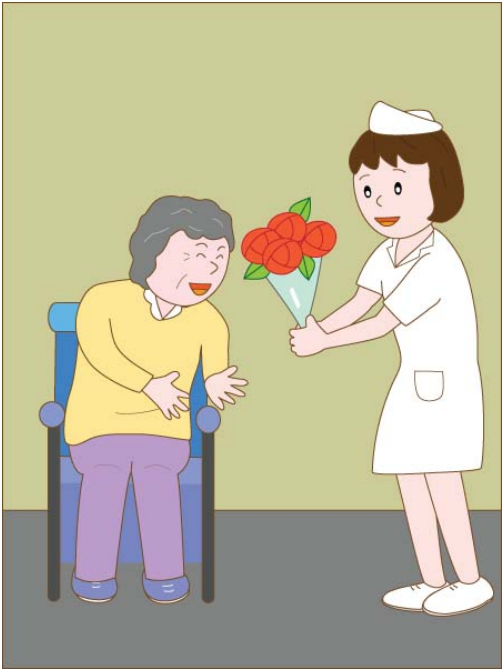
Example 2
What is happening?
Your sentence: *The nurse is giving a patient roses.*
Your second sentence: *A woman is getting flowers from a nurse.*
There is not a single correct response. Many responses may be possible. Other responses might be: *The nurse is giving a lady some red flowers.*   *A patient is receiving flowers from a nurse.*

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

2 Annotating Features

Each response is annotated according to five dimensions, or *features*. These features, explained below, are referred to as **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; annotatos 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).

| 1: He's eating a delicious slice of pizza. |
| 2: He's eating pizza, yuck. |
| 3: He will get fat eating pizza. |
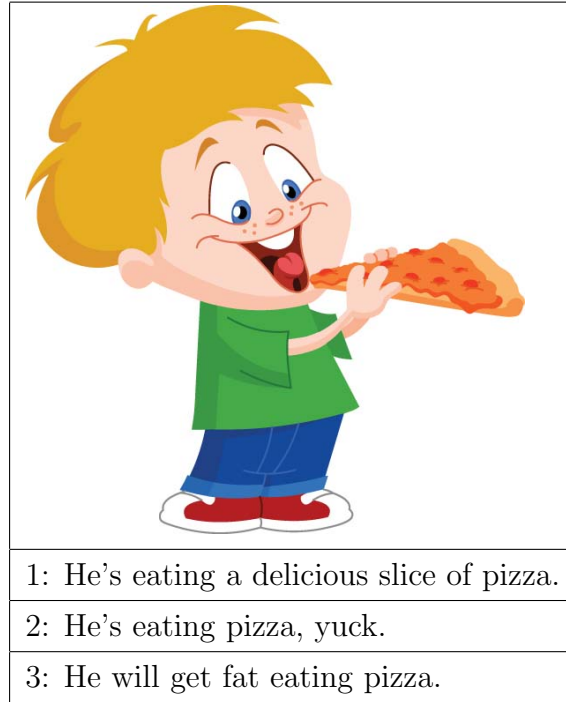
Figure 3: Item 2 (targeted: *What is the boy doing?*) and example responses.

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 in 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.

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

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.

15

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.

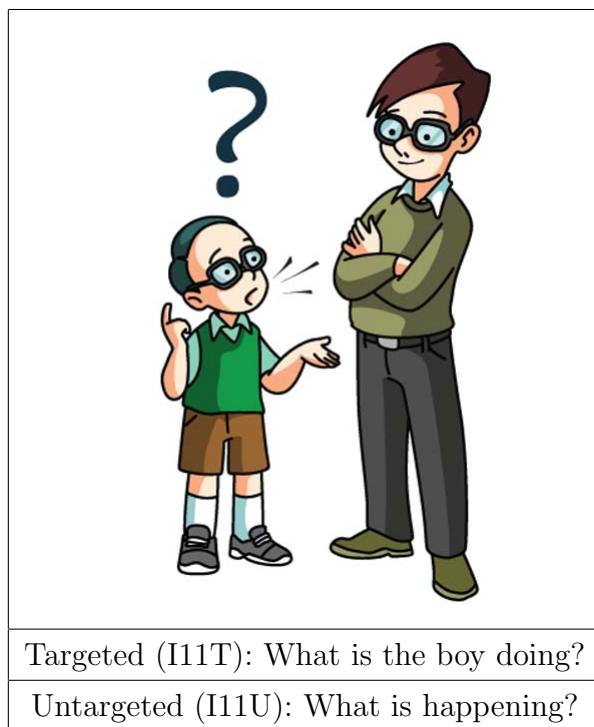| Targeted (I11T): What is the boy doing? |
| Untargeted (I11U): What is happening? |

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 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 in 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.

| Targeted (I03T): What is the man doing? |
| Untargeted (I03U): What is happening? |
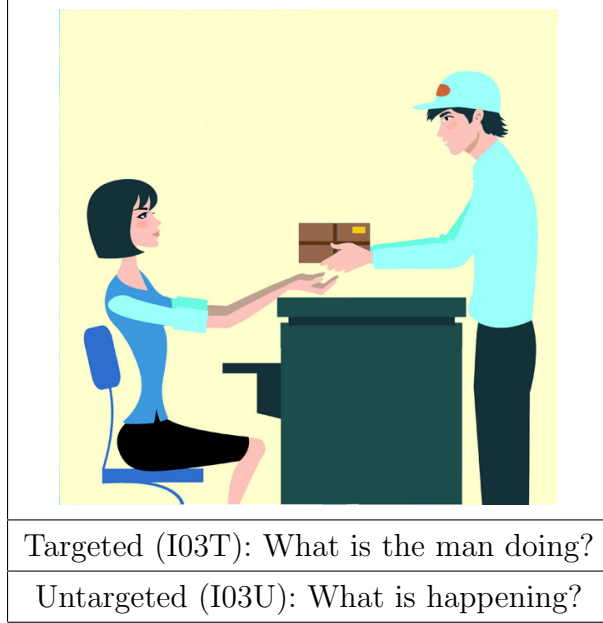
Figure 7: Item 3, in the targeted and untargeted versions.

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.
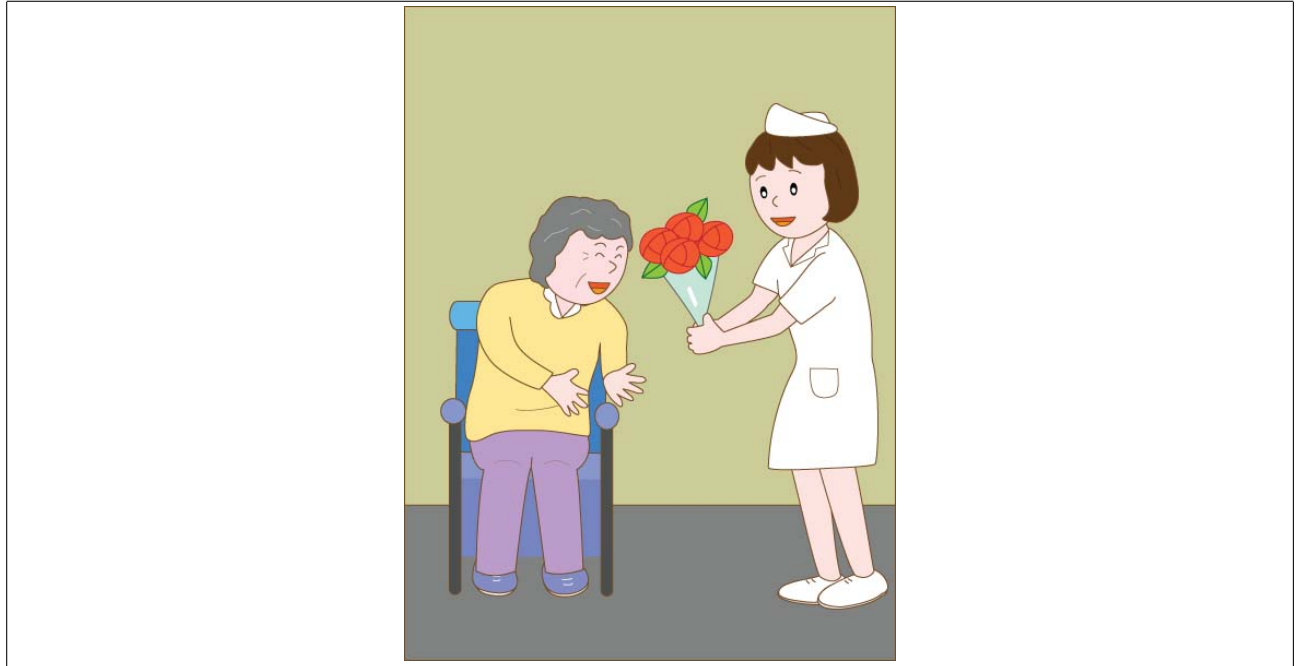
	Response	An.	Appropriate question
1	Giving a patient flowers.	yes	(prompt)
2	She's giving flowers to a patient.	yes	(prompt)
3	The nurse is giving away flowers.	yes	(prompt)
4	A nurse is giving away flowers.	no	What is happening?
5	A young nurse is giving away flowers.	no	What is happening?
6	The woman 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

24

(*(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 untargeted 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 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

Maria Pilar Agustín Llach. 2010. Lexical gap-filling mechanisms in foreign language writing. *System*, 38(4):529 – 538.

Luiz Amaral. 2007. Designing intelligent language tutoring systems: integrating natural language processing technology into foreign language teaching. *The Ohio State University, Columbus, OH*.

Luiz Amaral and Detmar Meurers. 2007. Conceptualizing student models for ICALL. In *Proceedings of the 11th International Conference on User Modeling*, Corfu, Greece.

Luiz Amaral, Detmar Meurers, and Ramon Ziai. 2011. Analyzing learner language: Towards a flexible NLP architecture for intelligent language tutors. *Computer Assisted Language Learning*, 24(1):1–16.

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.

Kathleen Bardovi-Harlig and Zoltán Dörnyei. 1998. Do language learners recognize pragmatic violations? Pragmatic versus grammatical awareness in instructed L2 learning. *TESOL Quarterly*, 32(2):233–259.

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 Morphologi-*

cally Rich Languages and Syntactic Analysis of Non-Canonical Languages, pages 66–73, Dublin, Ireland. Dublin City University.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Yeonsuk Cho, Frank Rijmen, and Jakub Novák. 2013. Investigating the effects of prompt characteristics on the comparability of toefl ibtł integrated writing tasks. *Language Testing*, 30(4):513–534.

Martin Chodorow, Joel Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30, Prague.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Rod Ellis. 1987. Interlanguage variability in narrative discourse: Style shifting in the use of the past tense. *Studies in Second Language Acquisition*, 9:1–19.

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.

Luis Gilberto Mateos Ortiz, Clemens Wolff, and Mirella Lapata. 2015. Learning to interpret and describe abstract scenes. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1505–1515, Denver, Colorado. Association for Computational Linguistics.

S. Granger, J. Hung, and S. Petch-Tyson. 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Language Learning & Language Teaching. John Benjamins Publishing Company.

Trude Heift and Mathias Schulze. 2007. *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.

DJ Hovermale. 2008. SCALE: Spelling Correction Adapted for Learners of English. Pre-CALICO Workshop on "Automatic Analysis of Learner Language: Bridging Foreign Language Teaching Needs and NLP Possibilities". March 18-19, 2008. San Francisco, CA.

Kazue Kanno. 1998. Consistency and variation in second language acquisition. *Second Language Research*, 14(4):376–388.

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.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL-03*, Sapporo, Japan.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Morgan & Claypool Publishers.

Henry Kucera and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

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.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. CUP.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*, Genoa, Italy.

Marie-Catherine de Marneffe and Christopher D. Manning. 2012. *Stanford typed dependencies manual*. Originally published in September 2008; Revised for Stanford Parser v. 2.0.4 in November 2012.

Detmar Meurers and Markus Dickinson. 2017a. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, 67(S1):66–95.

Detmar Meurers and Markus Dickinson. 2017b. 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.

Peter Skehan, Pauline Foster, and Uta Mehnert. 1998. Assessing and using tasks. In Willy Renandya and George Jacobs, editors, *Learners and language learning*, pages 227–248. Seameo Regional Language Centre.

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. The sole survivor of an ambush by the Skrull Empire, he spent his youth hunting bison, learning the songs of the direwolves and plotting his revenge.