SEMANTIC ANALYIS OF IMAGE-BASED LEARNER SENTENCES

Levi King

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Linguistics,

Indiana University

(Month) 2021

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

Doctoral Committee

Markus Dickinson, PhD

Sandra Kuebler, PhD

David Stringer, PhD

Sunyoung Shin, PhD

Date of Defense: Month/Day/2021

Copyright © 2021

Levi King

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

Acknowledgements v

Abstract vi

List of Tables xi

List of Figures xviii

Chapter 1: Introduction 1

1.1 Motivation 1

1.2 Overview 3

1.3 Research questions 7

1.4 Organization 9

1.4.1 Overview 9

1.4.2 Discussion 9

Chapter 2: Related Work 22

2.1 On interpretability for learner applications 22

2.2 Image processing 24

2.3 An overview of ICALL and content analysis 25

2.4 Learner corpora 26

2.5 Language assessment 27
2.6 NLP tools and methods 27

Chapter 3: Pilot Study 28
3.1 Introduction 28
3.2 Rule-based study 28
3.2.1 Data 29
3.2.2 Rule-based method 31
3.2.3 Semantic triple representation 33
3.2.4 Rule-based results 35
3.3 Semantic similarity study 38
3.3.1 Generalizing the methods 39
3.3.2 Term representation 40
3.3.3 Scoring responses 42
3.3.4 System Parameters 45
3.3.5 Experiments and Results 46

Chapter 4: Data Collection 51
4.1 Picture Description Task 51
4.2 Participants 55
4.3 Response Totals 56
4.4 Response Variation 57

Chapter 5: Annotation & Weighting 61

5.1 Annotation scheme 61
5.2 Agreement 66
5.2.1 Transitivity 68
5.2.2 Targeting 69
5.2.3 Features 69
5.2.4 NS & NNS responses 75
5.3 Establishing Feature Weights 76
5.4 Holistic Scoring and Ranking 81
5.5 Annotation Conclusions 85

Chapter 6: Optimization 87

6.1 Updated methodology 88
6.2 Sampling NS response models 91
6.3 SBERT as a benchmark 92
6.4 Sample statistics 94
6.5 Annotation features experiments 98
6.5.1 CORE EVENT experiments 101
6.5.2 ANSWERHOOD experiments 103
6.5.3 GRAMMATICALITY experiments 105
6.5.4 INTERPRETABILITY experiments 107
6.5.5 VERIFIABILITY experiments 109
6.6 Holistic experiments 112
6.6.1 `Term normalization` experiments 115

6.6.2 `Transitivity` experiments 117

6.6.3 `Targeting` experiments 120

6.6.4 `Familiarity` experiments 121

6.6.5 `Primacy` experiments 123

6.6.6 `Term representation` experiments 125

6.7 Optimization conclusion 127

Chapter 7: Conclusion 129

Appendix A: PDT Items 131

Appendix B: Annotation Guide 140

Curriculum Vitae

LIST OF TABLES

3.1 Sentence type examples, with distributions of types for native speakers (NS) and non-native speakers (NNS) 33

3.2 Contingency table comparing presence of NS forms (Y/N) with appropriateness (+/−) of NNS forms 36

3.3 Given the example sentence above, the updated approach represents responses in the dependency formats shown: ldh (for *label*, *dependent*, *head*; i.e., labeled dependencies)), xdh (unlabeled dependencies), lxh (label+head), ldx(label+dependent), or xdx (word, or more technically, *dependent*). 41

3.4 Example item rankings for the system setting combining these parameters: `TC`, `Brown`, and `ldh` (labeled dependencies). This was the best system setting based on average precision scores. Note that not all 39 responses are shown. 46

3.5 Based on Mean Average Precision, the five best and five worst system configurations across all 10 PDT items. 48

3.6 Individual parameters ranked by Mean Average Precision for all 10 PDT items. 48

4.1 Age and gender information for the three participant groups (Non Native Speakers, Crowdsourced Native Speakers and Familiar Native Speakers). . . 56

4.2 First and second response counts for the SAILS Corpus participant groups. Familiar (FNS) and crowdsourced (CNS) are subgroups of NS. NNS participants are not asked to provide a second response. 57

4.3 Crowdsourced responses for the items shown in Figure 4.2, showing one exemplar response and two examples of problematic or bad faith responses for each item. 57

4.4 This toy dataset shows how TTR is calculated on the response (sentence) 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. 58

4.5 NS and NNS type-to-token ratios (TTR) for complete responses (not words), for the full corpus. 59

4.6 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. 60

4.7 TTRs for complete responses, comparing first responses only. 60

5.1 Targeted and untargeted sample responses from the development set transitive item, shown with adjudicated annotations for CORE EVENT (*C*), ANSWERHOOD (*A*), GRAMMATICALITY (*G*), INTERPRETABILITY (*I*) and VERIFIABILITY (*V*). 67

5.2 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 (*Observ*) and Cohen's kappa (*Kappa*). 69

5.3 Comparing type-to-token ratios (*TTR*) for **main verbs** among the development and test set **ditransitive items**; greater variation correlates with lower CORE EVENT inter-annotator agreement, which helps explain why in Table 5.2 CORE EVENT agreement is lower than agreement for other features. . 71

5.4 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*). 76

5.5 Preference test sample responses pairs, annotator decisions (*A1 & A2*) and agreement for the item shown in Figure 5.4. 80

5.6 Preference test agreement scores for two annotators on a sample of 300 responses pairs, showing chance agreement, observed agreement and Cohen's Kappa. 80

5.7 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. 81

5.8 Example NNS responses (see Table 5.1) with feature weights applied to the binary annotations, resulting in weighted annotation scores (*WAS*) and a weighted annotation ranking (*WAR*). 82

5.9 Comparing scores for non-native speakers (*NNS*), **crowdsourced** native speakers (*CNS*) and **familiar** native speakers (*FNS*) across all items. *C+F* is the combination of CNS and FNS (i.e., *all* NS). *Total* is the response count. *Perfect* and *Zero* are the rates of responses with weighted annotation scores of 1.0 and 0.0, respectively. The *Mean*, *Median* and *Standard Deviation* values here are weighted annotation scores. 83

5.10 Examining **crowdsourced** native speaker response scores in different contexts: first and second responses (*R1* and *R2*); targeted and untargeted prompts; intransitives, transitives and ditransitives. (See Table 5.9.) 84

6.1 All parameters or variables and their settings; a system configuration combines one setting from each column. 90

6.2 Comparing average response length (in words) for the samples used throughout this chapter as NS models and NNS test sets, in total and by parameter setting. 94

6.3 Comparing average standardized type-to-token ratio (STTR) for the samples used throughout this chapter as NS models and NNS test sets, in total and by parameter setting. Tokens here are *dependencies*. 95

6.4 Comparing average NNS response scores across parameter settings and NS models, using the dependency tf-idf cosine scoring approach introduced in Section 3.3.3. The same sets of 70 NNS responses per model and configuration were scored here and throughout this chapter. Scores represent the NNS *distance* from the NS model, so lower scores are closer to NS behavior. Within each parameter, the score for the setting that minimizes distance is **bolded**, and the score for the model that minimizes distance is *italicized*. 97

6.5 Mean Average Precision (MAP) scores for the CORE EVENT annotation feature, derived from various response rankings: weighted annotation ranking (WAR), the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, primary NS models), and for the full set (Total). 102

6.6 Mean Average Precision (MAP) scores for the CORE EVENT annotation feature, comparing familiar and crowdsourced responses. MAP is derived from various response rankings: the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), weighted annotation ranking (WAR), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, targeted items), and for the full set (Total). Note that all models represented here are mixed due to the small number of familiar participants. 103

6.7 Mean Average Precision (MAP) scores for the ANSWERHOOD annotation feature, derived from various response rankings: weighted annotation ranking (WAR), the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, primary NS models), and for the full set (Total). 104

6.8 Mean Average Precision (MAP) scores for the ANSWERHOOD annotation feature, comparing familiar and crowdsourced responses. MAP is derived from various response rankings: the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), weighted annotation ranking (WAR), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, targeted items), and for the full set (Total). Note that all models represented here are mixed due to the small number of familiar participants. 104

6.9 Mean Average Precision (MAP) scores for the GRAMMATICALITY annotation feature, derived from various response rankings: weighted annotation ranking (WAR), the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, primary NS models), and for the full set (Total). 105

6.10 Mean Average Precision (MAP) scores for the GRAMMATICALITY annotation feature, comparing familiar and crowdsourced responses. MAP is derived from various response rankings: the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), weighted annotation ranking (WAR), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, targeted items), and for the full set (Total). Note that all models represented here are mixed due to the small number of familiar participants. 106

6.11 Mean Average Precision (MAP) scores for the INTERPRETABILITY annotation feature, derived from various response rankings: weighted annotation ranking (WAR), the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, primary NS models), and for the full set (Total). 107

6.12 Mean Average Precision (MAP) scores for the INTERPRETABILITY annotation feature, comparing familiar and crowdsourced responses. MAP is derived from various response rankings: the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), weighted annotation ranking (WAR), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, targeted items), and for the full set (Total). Note that all models represented here are mixed due to the small number of familiar participants. 108

6.13 Mean Average Precision (MAP) scores for the VERIFIABILITY annotation feature, derived from various response rankings: weighted annotation ranking (WAR), the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, primary NS models), and for the full set (Total). 110

6.14 Mean Average Precision (MAP) scores for the VERIFIABILITY annotation feature, comparing familiar and crowdsourced responses. MAP is derived from various response rankings: the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), weighted annotation ranking (WAR), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, targeted items), and for the full set (Total). Note that all models represented here are mixed due to the small number of familiar participants. 111

6.15 A "toy" model consisting of lemmatized syntactic dependencies from only two NS responses, each with perfect annotation scores. (Note that the version of Stanford typed dependencies used in this work collapses dependencies containing prepositions and incorporates prepositions in a label, resulting in the "prep_in" and "erased" dependencies above. See Section 3.2.2 for more on the parsing and lemmatization.) 115

6.16 Comparing Spearman rank correlation scores where all dependencies (terms) in `Non-n(ormalized)` NS models carry equal weight, and all dependencies in `Norm(alized)` NS models have their scores normalized proportionally to the length of the parent response. Results are shown using NS models of 14 responses and 50 responses. Each `Norm(alized)` and `Non-n(ormalized)` column represents 360 different rankings (12 system configurations × 30 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking. Each *SBERT* column represents 120 rankings (4 system configurations × 30 items; SBERT operates on plain text, so the `term representation` parameter does not apply). 117

6.17 Comparing Spearman rank correlation scores for `intransitive`, `transitive` and `ditransitive` PDT items, using NS models of either 14 or 50 random responses per item. Each *System* column represents 120 different rankings (12 system configurations × 10 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking. Each *SBERT* column represents 40 rankings (4 system configurations × 10 items; SBERT operates on plain text, so the `term representation` parameter does not apply). 118

6.18 Comparing Spearman rank correlation scores for `targeted` and `untargeted` versions of the PDT data, using NS models of either 14 or 50 random responses per item. Each *System* column represents 180 different rankings (6 system configurations × 30 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking. Each *SBERT* column represents 60 rankings (2 system configurations × 30 items; SBERT operates on plain text, so the `term representation` parameter does not apply). 122

6.19 Comparing Spearman rank correlation scores where `familiar` NS models contain only responses from participants *familiar* to the researcher and `Crowd` NS models contain only responses from `crowdsourced` participants. Results are shown using NS models of 14 responses; note the models used here are `mixed` (containing first and second responses; see Section 6.6.5) due to the sparsity of `familiar` data. Each *System* column represents 180 different rankings (6 system configurations $\times$ 30 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking. Each *SBERT* column represents 60 rankings (2 system configurations $\times$ 30 items; SBERT operates on plain text, so the `term representation` parameter does not apply). . 123

6.20 Comparing Spearman rank correlation scores where `primary` models contain only first responses from NSs and `mixed` models contain an equal mix of first and second responses from NSs. Results are shown using NS models of 14 responses and 50 responses. Each *System* column represents 180 different rankings (6 system configurations $\times$ 30 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking. Each *SBERT* column represents 60 rankings (2 system configurations $\times$ 30 items; SBERT operates on plain text, so the `term representation` parameter does not apply). 124

6.21 Comparing Spearman rank correlation scores where system configurations use different `term representations`: `ldh` (labeled dependencies), `xdh` (unlabeled dependencies), or `xdx` (dependents only; i.e., *words*). Results are shown using NS models of 14 responses and 50 responses. Each *System* and *SBERT* column represents 120 different rankings (4 system configurations $\times$ 30 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking. 126

LIST OF FIGURES

3.1 Example items from the previous study showing responses from native speakers of Arabic (Ar), Chinese (Ch), English (En) and Spanish (Sp). . . . 30

3.2 Dependency parse showing collapsed preposition dependencies. 32

3.3 The dependency parse of an example NNS response in a standard format (CoNLL) and the corresponding visual representation 34

3.4 Decision tree for determining sentence type and extracting semantic triple based on the presence of syntactic dependency labels 34

3.5 NNSs described this item in a variety of ways, while 100% of NSs used the verb *rake*. 37

4.1 All non-essential details were removed from the PDT images in order to focus participants' attention on the main action. 52

4.2 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. 53

5.1 Sample responses for the targeted item, *What is the woman doing?* 62

5.2 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. . . . 65

5.3 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. 68

5.4 Annotation interface used for the preference test. 79

6.1 The formula for average precision, where P_n and R_n are the precision and recall at the nth threshold. 98

CHAPTER 1
INTRODUCTION

1.1 Motivation

This work is motivated by the desire to bridge a disconnect between the fields of second language acquisition (SLA) and intelligent computer-assisted language learning (ICALL), with most SLA research supporting communicative and task-based learning methods, while many existing ICALL applications focus on explicit grammar instruction and correction. By combining existing language resources and natural language processing (NLP) tools to form an evaluation system and then investigating the results, my work aims to demonstrate that reliable, automatic, contextual, meaning-based analysis of non-native speaker (NNS) sentences (whether for testing, tutoring or otherwise) is possible without the need for developing expensive new datasets or processing tools, and thus encourage the use of such approaches in ICALL systems.

Within SLA and language pedagogy, tasks and task-based language teaching (TBLT) have been recognized for decades for their effectiveness and ability to elicit natural interlanguage (Ellis, 2003). In the context of language learning, whereas an *exercise* is an activity focused on language form, a *task* is "an activity which requires learners to use language, with an emphasis on meaning, to attain an objective" (Bygate et al., 2001). Since the inception of ICALL as a concept roughly 40 years ago, it has largely failed to incorporate task-based learning or the findings of SLA, relying on exercises instead. Nearly three decades ago, Oxford (1993) lamented that ICALL research relied on "outdated language learning and teaching references" and urged the field to "devote as much attention to its language learning/teaching principles as it does to its exciting technology". A number of researchers have worked to narrow this gap, but it remains a major obstacle today. Ziegler

et al. (2017) argue that the "current lack of a stronger connection between SLA and the NLP research underlying ICALL is indeed a missed opportunity," not only because the implications of SLA research can lead to more effective ICALL systems, but also because ICALL, with its ability to administer interesting tasks and collect large amounts of learner data in a variety of forms, can unlock new methodological opportunities in SLA.

I begin in Chapter 2 with an examination of a few ICALL projects that focus on meaning or are otherwise informed by SLA. Assessing why such projects have not gained traction, Schulze (2010) points out numerous concerns. First, these projects tend to require tremendous time and effort and are tailored for highly specific situations, meaning they do not easily transfer to new learners, languages, or activities. Moreover, they tend to rely on academic research funding, meaning development stops when the funding is exhausted. Because this dissertation does not present a full ICALL system nor a specific learning module but rather a mechanism for collecting and assessing responses to visual prompts, my hope is that it will inspire ICALL developers to adapt it to a range of visual learning situations or even new languages.

The main vision of this work is the automatic meaning-focused evaluation of the appropriateness of English NNSs' responses to visual cues by comparing the responses to those of native speakers (NSs). To overcome the many barriers to development and adoption of meaning-focused ICALL systems, a guiding principle of my work is that any system based on it should be easy to use and expand. If the ultimate goal is something like a story-based ICALL system in which users respond to images, an English language instructor should be able to easily add new stories by importing images and text prompts, then administering the task to NSs to automatically crowdsource a model, which would then be used to automatically evaluate NNS responses. Such an ICALL end product would certainly require a team of developers. What I present instead in this dissertation is a proof-of-concept using picture description task (PDT) responses from NSs and NNSs and a prototype evaluation system.

1.2 Overview

In Chapter 2, I examine related work from the fields of SLA, ICALL, language testing, NLP, corpus linguistics, and image analysis, pointing out where my work is similar to or inspired by others, and where my work diverges. I begin the chapter with a look at research and discussions from SLA that support the use of task-based and communicative approaches to second language learning, along with a look back at the methods they replaced. I follow with a brief overview of the history of computer-assisted language learning (CALL) and an examination of the shift to *intelligent* CALL, or *I*CALL, and what this shift means and entails. I give examples of past and present systems that rely on grammar and lexical drills, and contrast these with more sophisticated systems that attempt to focus on meaning over form. Much of the outside work relevant to this dissertation comes from the field of language testing, so I include an overview of testing research involving methods for automatically evaluating the sentences of language learners, focusing on systems that prioritize meaning.

I include a review of the NLP topics and technologies that my work is built on and explain why they are suited for my goals. This includes discussions of semantic textual similarity, bag-of-words analysis, topic analysis, dependency grammars and syntactic dependency parsing, lemmatization, and language modeling. I describe the tools available for these tasks, point to their use in ICALL and language testing, and present my rationale for relying on them in my work. Because my work involves handling human descriptions of images, I also look at exciting current work in automatic image captioning and discuss why I chose not to rely on such technologies.

My ambitions for this project present an unavoidable chicken-and-egg problem, which I attempt to take on in Chapter 3. In order to determine if a semantic similarity rating system can work for image-based data, one needs such data, and in order to determine if image-based data is suitable for a semantic similarity rating system, one needs such a system.

This is not a standard NLP task like sentence parsing or part-of-speech tagging for which appropriate datasets are available, so I began with a pilot study in which I collected a small set of data and developed a handful of working scoring approaches. This chapter describes the pilot data and its collection and annotation. It then looks at my initial, rule-based system for extracting *verb-subject-object* triples from dependency parses and attempting to match these NNS triples with NS triples. I examine the weaknesses of this system and explain how my initial findings lead to an improved approach. This new approach is more robust and adaptable, using term frequency-inverse document frequency (tf-idf) on dependency-parsed sentences to score NNS sentences according to their distance from a collection of NS parsed sentences. This found moderate success, but was limited by the small dataset and its inadequate *error-nonerror* annotation, leading to the expanded dataset, more detailed annotation, and refined scoring system covered in the remainder of this dissertation.

In Chapter 4, I present the data collection instrument and the resulting dataset. In this PDT, I took even greater care to strip unnecessary detail from the images and thus elicit a more constrained range of responses. More importantly, whereas the pilot PDT depicted 10 transitive events only, this version included 10 each of intransitive, transitive, and ditransitive events. Additionally, I varied between *targeted* prompts, which ask about the subject specifically, and *untargeted* prompts, which simply ask the respondent to describe the event. I also collected responses from many more participants: 499 participants versus 54 in the pilot study. This resulted in 13,533 responses, where the pilot dataset contained only 530. The participants included 141 NNSs (university level English as a Second Language students—overwhelmingly native speakers of Chinese) and 358 NSs. Of these NSs, 329 were crowdsourced and 29 were personally known to me; this distinction allowed me to compare the behavior of these groups for suitability for this project. I discuss this and other modifications in the data collection, and I describe surface level features of the dataset like the frequency of identical responses.

The development and implementation of the annotation scheme for this project is de-

tailed in Chapter 5. I explain the evolution of my scheme, which began as a three-point scale indicating *accurate and native-like*, *accurate but not native-like*, and *neither accurate nor native-like*. This was unreliable, and I describe how the constructs of *accurate* and *native-like* were ultimately split into five binary features, which I call CORE EVENT, ANSWERHOOD, GRAMMATICALITY, INTERPRETABILITY, and VERIFIABILITY. This chapter explains how these features and the annotation guidelines developed through an iterative process with annotators. I present inter-annotator reliability data, with all features showing satisfactory raw agreement scores above 91% and Cohen's kappas above 0.74. I also examine how the different participant groups compare with regard to the annotation features. The final section of this chapter discusses how I used a paired-response preference task to determine weights for each feature, allowing me to interpolate a holistic score for each response, which can in turn be used to produce a benchmark ranking of NNS responses. It is worth noting here that while the goal of my project is a system that can operate reliably on *unannotated* (or *unsupervised*) data, an important step to that end is producing a reliably annotated dataset to allow for evaluation of such systems.

In Chapter 6, I return to automatic content analysis, applying it to the new dataset. This begins with discussion of the ways in which I modified my system and expanded its settings. The general focus of this chapter is the search for correlations between the performance of particular system settings on particular items. Essentially, this is an attempt to answer questions like, *For predicting* CORE EVENT *annotations, should responses be represented internally as labeled dependencies or unlabeled dependencies?* and *For evaluating NNS intransitive sentences, do crowdsourced or familiar NS responses work best as a model?* By identifying such trends, I show that not only can my system successfully operate as intended, but it can even leverage these correlations to automatically choose the optimal settings for new items.

I offer two methods for evaluating my system. First, for the individual binary annotation features, I use mean average precision to determine how well system-produced

rankings separate positively and negatively annotated responses. For comparison, I use a state-of-the-art statistical language modeling tool, SBERT, to rank NNS sentences by their similarity to the NS sentences. These experiments show that my system generally outperforms SBERT at assessing annotation features. For evaluating my system's performance at holistically ranking responses, I use Spearman's rank correlation to compare against the benchmark rankings derived from weighted annotations. I also compare SBERT response rankings with the benchmark rankings. These experiments show that SBERT generally outperforms my system at holistic ranking. Taken together, these results show that my system is preferable for identifying custom features, which may be very useful in particular ICALL tasks, but for more general purposes, a sophisticated tool like SBERT is more effective. Importantly, as the differences in performance of my system and SBERT are generally small, my system can offer a parallel analysis that is more accessible and explainable than SBERT's, and thus better suited for developing a feedback module for end users.

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.

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

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 **??**, 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: 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*.

LK: cite Ellis, Meurers

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

CHAPTER 3
PILOT STUDY

3.1 Introduction

As discussed in Chapter 1, this dissertation grew from a desire to introduce lightweight, transparent methods of content analysis for contexts involving non-native speakers (NNSs) and second language instruction or assessment. In this chapter, I detail my initial foray into this work, where I conducted experiments to uncover the challenges involved and determine the feasibility of my goals. In that sense, the work described in this chapter can be seen as a pilot study and a proof-of-concept for the expanded study described in the chapters that follow.

In Section 3.2, I discuss a set of experiments that relied on custom rules to extract important elements from responses in order to match NNS responses to native speaker (NS) responses for assessment. In Section 3.3, I lay out a more data-driven set of experiments and the findings there that led directly to the collection and annotation of a much more robust corpus (see Chapters 4 and 5) and a deeper investigation into the impact of several variables on system performance (see Chapter 6).

3.2 Rule-based study

This study focused on analyzing English NNS responses to a picture description task (PDT) by comparison with NS responses. It was largely intended to determine if such a task would be feasible for a single researcher using off-the-shelf tools. This meant identifying problem areas and gauging whether variation in the form and content of responses could be manageable.

This section summarizes relevant work first presented in King and Dickinson (2013)

and King and Dickinson (2014); please see those papers for deeper discussions. However, please note that some experiments included in those papers are omitted here, and thus some results involving the averaging or ranking of scores differ slightly here.

3.2.1 Data

My work is inspired by various intelligent computer-assisted language learning (ICALL) projects and their reliance on visual stimuli (Nagata, 2002; Granström, 2004; Yamazaki, 2014). ICALL itself is heavily influenced by video games, which tend to be visual and interactive (Collentine, 2011). Thus, a dataset of responses to visual prompts was needed to develop the kind of content analysis pipeline I envisioned.

Without an appropriate dataset available, I opted to collect my own using a PDT. Linguistic research often relies on the ability of task design to induce particular behavior (Skehan et al., 1998), and a PDT can be used to induce behavior similar to that of an ICALL application. Moreover, the use of the PDT as a reliable language research tool is well-established in areas of study ranging from language acquisition to Alzheimer's disease (Ellis, 2000; Forbes-McKay and Venneri, 2005). Crucially, a PDT constrains the range of potential responses to contain only information depicted in the image, which greatly reduces the kind of processing needed to "understand" and assess responses with an automated system. Relatedly, particularly simple images should restrict elicited responses to a an even tighter range of expected contents.

I designed a data collection instrument consisting of a brief background questionnaire and 10 PDT items (eight line drawings and two photographs). Examples are given in Figure 3.1. Each item was intended to elicit a single sentence and consisted of an image depicting an action canonically framed as a transitive event in English, along with the question, "What is happening?" The instructions simply asked participants to "view the image and describe the action in either past or present tense" alone and without the use of any resources. The task was administered in a computer lab, where participants typed their

own responses. Automatic spelling correction was intentionally disabled on the machines.

	
Response (L1)	**Response (L1)**
He is droning his wife pitcher. (Ar)	The man killing the beard. (Ar)
The artist is drawing a pretty women. (Ch)	A man is shutting a bird. (Ch)
The artist is painting a portrait of a lady. (En)	A man is shooting a bird. (En)
The painter is painting a woman's paint. (Sp)	The man shouted the bird. (Sp)

Figure 3.1: Example items from the previous study showing responses from native speakers of Arabic (Ar), Chinese (Ch), English (En) and Spanish (Sp).

I collected responses from 53 participants for a total of 530 sentences. There were 14 NSs (non-linguistics undergraduate and graduate students) and 39 NNSs (university students enrolled in intensive English as a Second Language courses at Indiana University). The distribution of NNSs' first languages (L1s) was: 16 Arabic, 7 Chinese, 2 Japanese, 4 Korean, 1 Kurdish, 1 Polish, 2 Portuguese, and 6 Spanish.

Once the data had been collected, I removed a small number of low quality responses from the NS set and annotated the NNS responses for appropriateness, with respect to the content of the picture. My annotation guidelines for this labeling were minimal: *Given the prompt, would the response be acceptable to most English speakers?* This simplistic approach to annotation was admittedly a weak point in the study, and in subsequent work I sought to correct it (see Chapter 5). The resulting corpus was used in the experiments described next.

3.2.2 Rule-based method

The processing behind my work was inspired by research from areas such as sentiment analysis, topic modeling, and content assessment that used rule-based approaches to extract important elements from dependency-parsed text (Nastase et al., 2006; Bailey and Meurers, 2008; Di Caro and Grella, 2013). 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 this study, 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. My idea was to extract and lemmatize a *verb(subj,obj)* semantic triple from the dependency parse for each sentence. Each NNS triple could be compared against the list of NS triples for a match; a NNS response with a NS triple match would be labeled automatically as "correct," while a non-match would be labeled "incorrect."

For lemmatization, I used the pre-trained tool in the Stanford Core NLP package (Manning et al., 2014).[1] The purpose of lemmatization is to minimize data sparsity by reducing the number of word forms. For example, the main verb in the forms *kicks*, *kicked*, *has kicked* and *is kicking* was reduced to *kick* in all cases, increasing the likelihood of finding matches. For parsing, I used the Stanford Parser, trained on the Penn Treebank (de Marneffe et al., 2006; Klein and Manning, 2003).[2]

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 the parser's default settings, but added two non-default options (*CCPropagatedDependencies* and *CCprocessed*[3]) in order to: 1) omit prepositions and conjunctions as heads and dependents and instead add such words to the dependency label between content words;

[1] `https://stanfordnlp.github.io/CoreNLP/`
[2] `http://nlp.stanford.edu/software/lex-parser.shtml`
[3] `http://nlp.stanford.edu/software/dependencies_manual.pdf`

and 2) propagate relations across conjunctions. These decisions are important to consider for any semantically-informed processing of NNS language.

To see the impetus for removing prepositions, consider the learner example in Figure 3.2, 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 3.2. Importantly, this option results in a direct relationship between the verb *played* and the object *ball*.

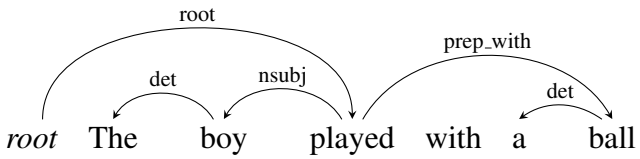


Figure 3.2: Dependency parse showing collapsed preposition dependencies.

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* 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 option is less important as conjunctions are rare in the data, but it is advantageous as it simplifies the representation 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.

3.2.3 Semantic triple representation

Type	Description	Example	NS	NNS
A	Simple declar. trans.	The boy is kicking the ball.	117	286
B	Simple + preposition	The boy played with a ball.	5	23
C	No tensed verb	Girl driving bicycle.	10	44
D	No tensed verb + prep	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 A or C	There is a boy kicking a ball.	0	0
Bx	Existential B or D	There was a boy playing with a ball.	0	0
Ex	Existential E	There is a woman cycling.	0	0
F1x	Existential F1	There is an apple being cut.	0	1
F2x	Existential 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 3.1: Sentence type examples, with distributions of types for native speakers (NS) and non-native speakers (NNS)

I manually categorized the 530 sentences in the dataset into 11 types plus one catch-all category, as shown in Table 3.1. I established these types because each one corresponds to a basic sentence structure and thus has consistent syntactic features, leading to predictable patterns in the dependency parses. A sentence type indicates that the subject, verb, and object can be found in a consistent place in the parse, such as under a particular dependency label. For simple transitive sentences (type *A* in Table 3.1), for example, the dependents labeled *nsubj*, *root*, and *dobj* 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 locations relative to word indices (see Figure 3.3).

Determining the sentence type was accomplished by arranging a small set of binary decisions into a tree, as shown in Figure 3.4. This decision tree checks for the presence of various dependency labels. The extraction rules for the particular sentence type were then applied to obtain the semantic triple. Finally, for each NNS response, the resulting

Index	Dependent	Lemma	POS	Head	Label
1	the	the	DET	2 (boy)	det
2	boy	boy	NN	4 (kicking)	nsubj
3	is	be	VBZ	4 (kicking)	aux
4	kicking	kick	VBG	0 (*root*)	root
5	the	the	DT	6 (ball)	det
6	ball	ball	NN	4 (kicking)	dobj

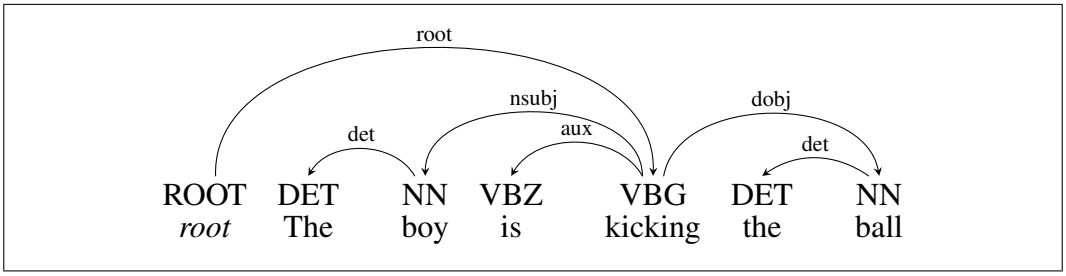


Figure 3.3: The dependency parse of an example NNS response in a standard format (CoNLL) and the corresponding visual representation

triple was lemmatized and checked against the list of lemmatized NS triples. Ideally, each acceptable response will find a match, and unacceptable responses will not.

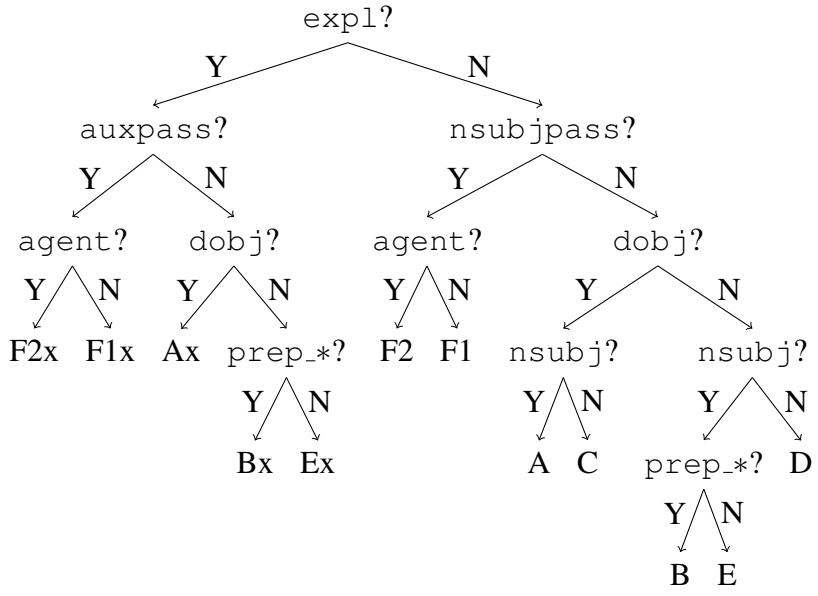


Figure 3.4: Decision tree for determining sentence type and extracting semantic triple based on the presence of syntactic dependency labels

3.2.4 Rule-based results

Evaluating this work required addressing two major questions. First, how accurately does this approach extract semantic information from potentially innovative sentences? Second, how many semantic forms does one need in order to capture the variability in meaning in NNS sentences? I operationalized this second question by asking how well a given set of NS semantic forms functions as an "answer key" or gold standard.

An accurate extraction was defined as one in which the extraction rules chose the desired subject, verb, and object given the sentence at hand and without regard to the PDT image. Accuracy was 92.3% for NNS responses and 92.9% for NS responses. I attribute the high extraction scores to the constrained nature of the task and the relatively small range of sentence types it elicits. As seen in Table 3.1, only three sentence types (A, B, and C) account for more than 90% of all responses.

Assessing the coverage of NNS forms required first manually determining which extracted triples *should* be matched given a hypothetical perfect gold standard set of triples. To separate the problem of coverage from extraction, I first removed any incorrectly extracted triples from the NNS set and the NS gold standard.

Using the annotations discussed in Section 3.2.1, I called an appropriate NNS triple found in the NS 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 3.2. 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, **coverage** was defined as recall: *TP/(TP+FN)*. I reported 23.5% coverage for unique triple *types* and 51.0% coverage for triple *tokens*.

I defined an inappropriate NNS triple (i.e., a content error) *not found* in the gold standard as a **true negative (TN)** (i.e., a correct non-match). **Accuracy** based on this gold

		NNS	
		+	−
NS	Y	TP	FP
	N	FN	TN

Table 3.2: Contingency table comparing presence of NS forms (Y/N) with appropriateness (+/−) of NNS forms

standard—assuming perfect extraction—is defined as *(TP+TN)/(TP+TN+FN)*.[4] I reported 46.4% accuracy for types and 58.9% accuracy for tokens.

The immediate lesson taken from this work was this: given a strict matching approach, NS data alone does not make for a sufficient gold standard, in that many appropriate NNS answers are not counted as correct. I explored expanding the set of NS triples by separating individual subjects, verbs and objects from NS triples and recombining them into the various possible combinations. However, this recombination generates a lot of nonsensical triples and degrades the gold standard. Consider, for example, *do(woman,shirt)*—an incorrect triple derived from the correct NS triples, *wash(woman,shirt)* and *do(woman,laundry)*. This could be improved somewhat by evaluating new combinations with a language model, but this would both complicate the approach and diverge from my vision of content analysis driven by real speaker behavior. Instead, my current work has attempted to improve coverage by prompting NSs to give an initial PDT response, followed by a second alternative, as discussed in Chapters 4 and 6.

A related concern was that, even when only examining cases where the meaning is literally correct, NNSs produced a wider range of forms to describe the prompts than NSs. For example, for a picture showing what 100% of NSs described as a *raking* action, many NNSs referred to a man *cleaning* an area (see Figure 3.5). Literally, this may be true, but it does not align with a NS gold standard. This behavior was expected, given that learners are encouraged to use words they know to compensate for gaps in their vocabularies

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

Response (L1)
The gardener is cleaning the street. (Portuguese)
a man removing the tree leafs. (Arabic)
The man is sweeping the floor. (Chinese)
A man is gathering lots of leafs. (Korean)

Figure 3.5: NNSs described this item in a variety of ways, while 100% of NSs used the verb *rake*.

(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.

Moreover, evaluating this strict matching approach required an annotator to decide whether a given response is correct or incorrect. Partial matching is not allowed; this is an inherent weakness of the approach, because while a complete triple gives some indication of the meaning of the sentence, any single element of the triple taken alone does not provide enough context to indicate meaning. This inflexibility means that using this approach would effectively require the manual curation of a robust gold standard set of acceptable responses, which is counter to my goal of producing an approach that can be expanded to new PDT items simply by crowdsourcing responses from NSs.

I followed up this work with a modification that included language models and spelling

correction tools to attempt to identify and fix misspellings that lead to downstream problems (King and Dickinson, 2014). I omit this discussion because it is not applicable or comparable to my expanded corpus; I now take a much simpler approach—respondents used their browser's spell checker as necessary during the task (see Section 4.1). This is because in most contexts where my system would be used, like a language tutoring application or game, spelling instruction is not the objective, and a built-in spell checker would likely be available. Moreover, omitting this step removes a layer of analysis—and importantly, a potential source of errors—and allows the research to focus more directly on meaning.

Ultimately, I was not satisfied with the effectiveness of the rule-based approach. While the roughly 92% accuracy rate for extraction may seem high, triple extraction only represents an intermediate step of the processing and effectively sets a ceiling for downstream performance. With regard to matching semantic triples, the coverage score of 51% is somewhat less informative because it is at least partially a function of the small gold standard (n=14). However, the homogenous nature of NS responses seen for several items here (such as the *raking* example discussed above) suggests that expanding NNS coverage is not simply a question of collecting responses from more NSs. These findings led me to experiment with new approaches, beginning with the work described in Section 3.3.

3.3 Semantic similarity study

In subsequent work (King and Dickinson, 2016), I began looking for a "sweet spot" of semantic analysis (cf. Bailey and Meurers, 2008) for image-based learner productions. I should note, in this context, that I am discussing semantic analysis given a reliable set of NS sentences, as opposed to other areas of research tying images and language. Image processing tasks like automatic captioning often rely on breaking images into semantic primitives (see, e.g., Gilberto Mateos Ortiz et al., 2015, and references therein), but for NNS 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 content analysis, beyond correctness, stems from using an incomplete gold standard. The productive nature of language means that a sentence can be expressed in countless ways, and thus a gold standard can never really be "complete." For this reason, it is necessary to assume that novel NNS productions will occur, and the methods of analysis should be robust enough to differentiate between responses that do not appear in the gold standard because they are bad responses, and those that do not appear in the gold standard simply because they are infrequent. In particular, using available NLP tools, I moved away from discrete representations of correctness in the form of a NS gold standard set of semantic triples to a more continuous notion of correctness using a NS model comprised of smaller, overlapping pieces of information. This obviates the need for a rule-based extraction of semantic triples, which is a source of errors and must be customized for a limited range of expected sentence types. It also allows for graded scoring of results, meaning that a response is not outright rejected because only one argument of a triple is not found. On the other hand, it means that the system does not provide a discrete "thumbs up" or "thumbs down" decision for each response, which may be desirable in some use cases.

3.3.1 Generalizing the methods

The previous work assumed that the assessment of NNS responses involves determining whether a NS gold standard set contains the same semantic triple that the NNS produced, i.e., whether a triple is *covered* or *non-covered*. In such a situation the gold standard need only be comprised of *types* (not *tokens*) of semantic triples. Here the gold standard is comprised of the small set of NS responses—only 14. This means that exact matching is going to miss many cases, and indeed as discussed in Section 3.2.4, coverage was only 51%. Even with a much larger gold standard, we can expect responses to follow Zipf's

law; a sample of language data will always be incomplete because it will not contain all "long tail," low-frequency phenomena.

Additionally, relying on matching of triples limits the utility of the method to specific semantic constraints, namely transitive sentences. By dropping the exact matching approach and instead comparing the frequencies of elements in the NNS response with those of the gold standard, I moved into a gradable, or ranking, approach to the analysis, which is agnostic with regard to an item's transitivity or sentence type (see Table 3.1). For this reason, I shift terminology here from NS ***gold standard*** to NS ***model***. A gold standard is roughly akin to an answer key, which was appropriate for my strict triple-matching approach. A model is typically a richer data structure containing statistics from observations of known correct data. This distinction and its relevance will be more apparent with the discussion of response representations in Section 3.3.2.

My goal is to emphasize the degree to which a NNS response conveys the same meaning as the set of NS responses, necessitating an approach which can automatically determine the importance of a piece of information among the set of NS responses. This required two major decisions: 1) how to **represent** each response as a set of sub-elements, and 2) how exactly to **score** these sub-elements via comparison with the NS data. In Section 3.3.2, I detail how I represented the information and in Section 3.3.3, I discuss comparing NNS information to NS information, which allowed me to rank responses from most to least similar to the NS model. In Section 3.3.4, I describe the system parameters that I combined to generate scores, and in Section 3.3.5 I present and interpret the results of the various settings.

3.3.2 Term representation

I first represented each NNS response as a list of dependencies taken directly from the parse. This eliminates the complications of extracting semantic triples from dependency parses, which only handled a very restricted set of sentence patterns and resulted in errors in

7–8% of cases, as discussed in Section 3.2.4. 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. As before, the responses were lemmatized to minimize data sparsity.

The boy is kicking the ball.				
ldh (labeled)	**xdh** (unlab.)	**lxh**	**ldx**	**xdx** (word)
det(the,boy)	*x*(the,boy)	det(*x*,boy)	det(the,*x*)	*x*(the,*x*)
nsubj(boy,kick)	*x*(boy,kick)	nsubj(*x*,kick)	nsubj(boy,*x*)	*x*(boy,*x*)
aux(is,kick)	*x*(is,kick)	aux(*x*,kick)	aux(is,*x*)	*x*(is,*x*)
root(kick,root)	*x*(kick,root)	root(*x*,root)	root(kick,*x*)	*x*(kick,*x*)
det(the,ball)	*x*(the,ball)	det(*x*,ball)	det(the,*x*)	*x*(the,*x*)
dobj(ball,kick)	*x*(ball,kick)	dobj(*x*,kick)	dobj(ball,*x*)	*x*(ball,*x*)

Table 3.3: Given the example sentence above, the updated approach represents responses in the dependency formats shown: ldh (for *label*, *dependent*, *head*; i.e., labeled dependencies)), xdh (unlabeled dependencies), lxh (label+head), ldx(label+dependent), or xdx (word, or more technically, *dependent*).

Next, I obtained five different representations from the lemmatized dependencies, as shown in Table 3.3. I refer to this variable as **term representation**, in keeping with *term* frequency-inverse document frequency, discussed in Section 3.3.3. The five term representations are then variations on dependencies. The full form is the **labeled dependency** and includes the **l**abel, **d**ependent and **h**ead, so I refer to it in shorthand as **ldh**. The remaining four forms abstract over either the label, dependent and/or head. I refer to these forms as **xdh** (i.e., unlabeled dependency), **lxh** (label+head), **ldx** (label+dependent), and **xdx** (dependent only, roughly equivalent to *word* in a bag-of-words approach).

The goal in choosing these five representations was to find the optimal combination of dependency features and the right level of detail to obtain the best system performance. The bag-of-words representation was implemented largely to provide a baseline by which to compare the others.

This processing was applied to the collection of NS responses as well. For each item, the dependencies from all NS responses was pooled into a single flat list—a "bag-of-

dependencies." From this list, a copy in each of the five term representations was produced, allowing for comparison with the corresponding NNS data.

3.3.3 Scoring responses

Taking the five term representations, my next step was to automatically score them in a way which ranks responses from most to least appropriate. I devised four scoring 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 NS data.

For weighting, the simplest method used the relative frequency of each term (i.e., dependency). This is the token count of a given term in a document normalized by the total count of tokens in the document.

The other method of weighting was based on *term frequency-inverse document frequency* (*tf-idf*), which scores the importance of terms in a test document according to their frequencies in the language generally (Manning et al., 2008, ch. 6). It approximates this by taking a large reference corpus comprised of many smaller documents, and counting the number of those documents in which the terms in question occur. Most commonly, a *term* for tf-idf purposes would be a *word*. However, *term* here refers to a single syntactic dependency, and this is a central conceit of this dissertation: the dependency, which captures aspects of semantic and syntactic relationships, is an ideal atomic unit for evaluating meaning by comparing distributions in crowdsourced data. As discussed in Section 3.3.2, the dependencies were represented in one of the five term representations—some combination of label, dependent and head. For the NNS data, each response was treated as a tf-idf test document. For the NS data, the entire collection of NS responses was treated as one tf-idf test document. The idea here is to obtain a similarity measure between a single NNS response and the full collection of NS responses, so each is handled as a single document.

I used tf-idf as a measure of a term's importance with the expectation that it would reduce the impact of semantically less important terms—e.g., determiners like *the*, fre-

quent in the responses, but mostly unnecessary for evaluating the semantic contents—and to upweight terms which may be salient but infrequent, e.g., only used in a handful of NS sentences. For example, for an item depicting a man shooting a bird (see Figure 3.1), of 14 NS 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 NS responses should get upweighted via tf-idf, and indeed that was the effect; in the bag-of-words approach, the term *hunter* is weighted among the highest, and the same is true among the other term representations for dependencies containing the word *hunter*. 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 within a document. Inverse document frequency is derived from a reference corpus of documents, 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 shown in (1) 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.

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

After this frequency counting or tf-idf weighting, the scores were then either **averaged** to yield a response score, or NNS term weights and NS term weights were treated as vectors and the response score was the **cosine distance** between them. This yields the four approaches:

Frequency Average (FA). Within the set of NS responses, the relative frequency of each term is calculated. The NS *model* here is simply each NS term and its relative frequency. Each term in the NNS response is then given a score equal to its frequency in the NS model; terms missing from the NS model are scored zero. The response score is the average of the

term scores, with higher scores closer to the NS model.

Tf-idf Average (TA). This involves the same averaging of term scores as with approach FA, but here the term scores are the tf-idf scores. The NS model here is thus each NS term and its tf-idf score. These NS tf-idf term scores are applied to the terms in the NNS response and then averaged for the NNS response score.

Frequency Cosine (FC). Each term score is taken as its relative frequency calculated *within* its document: either the NS response set or the single NNS response. In other words, the NS term scores are not applied to the NNS terms. The NS model is then the set of all NS terms and their scores. The term scores are then treated as vectors—one vector of the NS term scores (i.e., the NS model here)—and one vector of the NNS term scores. Each vector is an ordered list of term scores for each term observed in either the NS document or the NNS document. In other words, each vector represents term scores for the sorted union set of NS and NNS terms. Naturally, many of the term scores are zero for the much shorter NNS document. The response score is the cosine distance between the vectors, with lower scores being closer to the NS model.

Tf-idf Cosine (TC). This involves the same distance comparison as with approach FC, but now the term scores in the vectors are tf-idf scores. The NS model here is thus the vector representing the union set of terms for the NS and NNS document, populated with the tf-idf scores from the NS document.

In many natural language processing scenarios where high dimensional vectors are involved, such as sentence encoders or word embeddings, methods for dimensionality reduction are employed (Devlin et al., 2018; Mikolov et al., 2013). This improves efficiency by reducing the storage and computing power needed. In my FC and TC approaches, however, the number of term types remained small enough that the raw vectors representing dependencies' tf-idf scores can be processed easily with an ordinary PC. Not only does this

simplify the task, it means that the process remains transparent. There are no transformers or attention mechanisms to produce compressed and unexplainable representations. If desired, each sentence vector can be examined value by value, where each number maps to a real syntactic dependency. This is important because it leaves the door open for meaningful feedback on each response. For example, one might choose to identify the most salient dependencies in the NS model and use them to guide an ICALL user from a low scoring response to a better response.

3.3.4 System Parameters

Each experiment was run with a unique system **configuration**, which is a combination of the following parameter settings:

Term representation As discussed in section 3.3.2, the terms can take one of five representations: **`ldh`**, **`xdh`**, **`lxh`**, **`ldx`**, or **`xdx`**.

Scoring approach. As discussed in section 3.3.3, the NNS responses can be compared with the NS models via approaches **`FA`**, **`TA`**, **`FC`**, or **`TC`**.

Reference corpus. The reference corpus for deriving tf-idf scores can be either the Brown Corpus (**`Brown`**) (Kucera and Francis, 1967) or the Wall Street Journal Corpus (**`WSJ`**) (Marcus et al., 1993). The Brown Corpus is just over one million words across 500 documents intended to cover a broad range of genres, registers, and contents and to serve as a sample of the written English language at large. The WSJ Corpus used here consists of 1,640 documents; the documents are newspaper articles from 1989 and total one million words. Considering the narrative nature of PDT responses, a reference corpus of narrative texts would be ideal, but as no such reliably parsed corpus is available, I chose the widely used, pre-parsed `Brown` and `WSJ` corpora. The corpora were converted from their standard dependency parse format to each of the five term representations used in order to be compat-

ible with the NS and NNS data for calculating frequencies and tf-idf.

3.3.5 Experiments and Results

Evaluation metrics

Combining the various parameter settings results in 30 different configurations, and an experiment was run with each (section 3.3.4). For example, one such configuration is [`ldh`, `TC`, `WSJ`]. Within each experiment, I ranked the 39 scored NNS responses from least to most similar to the NS model.

For assessing these configurations, I relied on the past annotation, which counted unacceptable responses as errors (see section 3.2.4). As the lowest numerical rank indicates the greatest distance from the NS model, a good system configuration should position the unacceptable responses among those with the lowest rankings. To evaluate this discriminatory power, I used **(mean) average precision ((M)AP)** (Manning et al., 2008, ch. 8).

Rank	Score	Sentence	Error
1	1.000	she is hurting.	1
	1.000	man mull bird	1
3	0.996	the man is hurting duck.	1
4	0.990	he is hurting the bird.	1
11	0.865	the man is trying to hurt a bird	1
12	0.856	a man hunted a bird.	0
17	0.775	the bird not shot dead.	1
18	0.706	he shot at the bird	0
19	0.669	a bird is shot by a un	1
20	0.646	the old man shooting the birds	0
37	0.086	the old man shot a bird.	0
38	0.084	a old man shot a bird.	0
39	0.058	a man shot a bird	0
Average Precision: 0.75084			

Table 3.4: Example item rankings for the system setting combining these parameters: `TC`, `Brown`, and `ldh` (labeled dependencies). This was the best system setting based on average precision scores. Note that not all 39 responses are shown.

For average precision (AP), one calculates the precision of error detection at every point

in the ranking, lowest to highest. Consider Table 3.4, which presents an excerpt of ranked sentence responses for one PDT item. The precision for the first cut-off (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 the NNS data), as shown in (2), with each response notated as R_k. Averaging over all 10 items results in the Mean AP (MAP).

(2) $$AP(item) = \frac{1}{m} \sum_{k=1}^{m} Precision(R_k)$$

Best system parameters

To start the search for the best system parameters, it may help to continue with the example in Table 3.4. The best configuration, as determined by the MAP metric, uses the tf-idf cosine (`TC`) approach with the Brown Corpus (`Brown`), and the full form of the labeled 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 NS model (as all NSs used *shoot* to describe the action) and the potential importance of even simple representations like lemmas. In this case, the labeled dependency (`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 3.5 shows the five best and five worst configurations averaged across all 10 PDT items, as ranked by MAP. The table clearly indicates that the `TC` approach is superior, occurring in all of the top five combinations. `Brown` appears among top scoring configurations more often than does `WSJ`. Finally, for the term representation, labeled (`ldh`) and unlabeled (`xdh`) dependencies are used most among the top scoring configurations.

I also summarize the rankings for each isolated parameter, presented in Table 3.6. For

Rank	MAP	Approach	Reference	Term rep.
1	0.5168	TC	Brown	ldh
2	0.5128	TC	WSJ	ldh
3	0.5124	TC	Brown	xdh
4	0.5109	TC	Brown	lxh
5	0.5102	TC	WSJ	xdh
26	0.4826	FA	*na*	ldx
27	0.4816	TA	Brown	xdx
28	0.4769	FC	*na*	lxh
29	0.4721	TA	WSJ	xdx
30	0.4530	FA	*na*	lxh

Table 3.5: Based on Mean Average Precision, the five best and five worst system configurations across all 10 PDT items.

Approach		Term rep.		Ref. corpus	
0.50630	TC	0.50499	ldh	0.50461	Brown
0.49609	TA	0.50405	xdh	0.49777	WSJ
0.49471	FC	0.49287	ldx		
0.48247	FA	0.49190	xdx		
		0.49115	lxh		

Table 3.6: Individual parameters ranked by Mean Average Precision for all 10 PDT items.

a given parameter, e.g., ldh, I averaged the scores from all configurations including ldh across all 10 items. Generally, the same trends appear salient. Notably, TC outperformed the other approaches, with FC and TA close behind (and nearly tied). Performance fell for the simplest approach, FA, which was in fact intended as a kind of baseline. With **T**C>**F**C and **T**A>**F**A, tf-idf weighting seems preferable to basic frequencies. Likewise, with T**C**>T**A** and F**C**>F**A**, for my term based scoring, taking the cosine of score vectors outperformed simply comparing score averages.

These findings largely confirmed my expectations. The TC approach was intended to evaluate responses by focusing comparison on the most salient content. The scores here show that to be successful. Regarding the top term representations, labeled and unlabeled dependencies both capture the relationship between dependents and their heads, making

them an ideal unit for analyzing "who did what to whom" in the context of a PDT. Finally, given the subject matter and narrative style of the task, it is unsurprising that `Brown` serves as a better tf-idf reference than `WSJ`.

The trends noted in these averages were strong overall, but a closer look at individual PDT items revealed some exceptions. For example, for the item depicting a man raking leaves, the `ldx` and `xdx` term representations were the top performers. As discussed in Section 3.2.4, 100% of NSs used the verb *rake*, but only 3/39 NNSs used *rake*. Consider the response, "The man rakes leaves." We can expect the main verb of a transitive sentence to appear in at least three dependencies: root(rakes, ROOT), nsubj(man, rakes), and dobj(leaves, rakes). Note that in two of these three, the verb is the syntactic head. Thus, by omitting heads, the `ldx` and `xdx` representations increase the likelihood of overlap between the NS response and the NNS model. For example, "The man rakes leaves" and "The man sweeps leaves" both result in the `ldx` terms nsubj(man, *x*) and dobj(leaves, *x*).

A similar pattern emerged for an item that 12 out of 14 NSs described as some version of "a woman is riding a bike/bicycle," using the same subject, verb and object. Seven of 39 NNSs simply framed this as an intransitive, e.g., "The woman is biking." Two NNSs chose another verb—*pedal* or *drive*. Finally, while 30 NNSs used the verb *ride*, six of these misspelled it—*rid/ridding* or *ridging*. It is also worth noting that seven of 37 instances of *bike* or *bicycle* were misspelled. For this item, the top system performance came from combining the `FC` approach with the `xdx` (bag-of-words) representation. The combination of noisy (misspelled) NNS data and the small, homogenous set of NS data meant that the least granular term representation (`xdx`) worked best. The more sophisticated tf-idf based approaches suffer here due to the level of noise, allowing `FC` to win out.

Overall, the results of the dependency-based similarity approaches to content analysis were sufficiently promising to warrant deeper investigation. The biggest challenge to the validity of these results was the size of the dataset and the nature of its annotation, so I next sought to expand on this work by building a bigger and better dataset and conducting similar

experiments. In Chapter 4, I discuss the collection of a much larger dataset that includes intransitive and ditransitive items in addition to transitives. This also involved a new task variable, where half of the prompts focus the participant's attention on the desired subject of the response, and the other half do not. Additionally, NS participants were instructed to provide two responses per item. These modifications allowed for deeper investigation into task effects and the variability of responses under different conditions. In Chapter 5, I discuss the development of a new non-binary annotation and weighting system intended to improve the utility of the dataset and the reliability of its annotation. In Chapter 6, I detail the revisions to my methods that sprang from the findings here. These include building NS models based on responses from personally known NSs and crowdsourced NSs and the move to rank correlation metrics to accommodate the updated annotation scheme. The chapter includes the results of several experiments aimed at finding optimal system settings.

CHAPTER 4
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 and the lessons learned from the dataset described in Chapter 3 in mind, I decided to collect a corpus of picture description task responses for use in my experiments. I discuss the data collection instrument in Section 4.1 and the participants in Section 4.2. In Sections 4.3 and 4.4, I present the total response counts for various categories and examine the variability of responses.

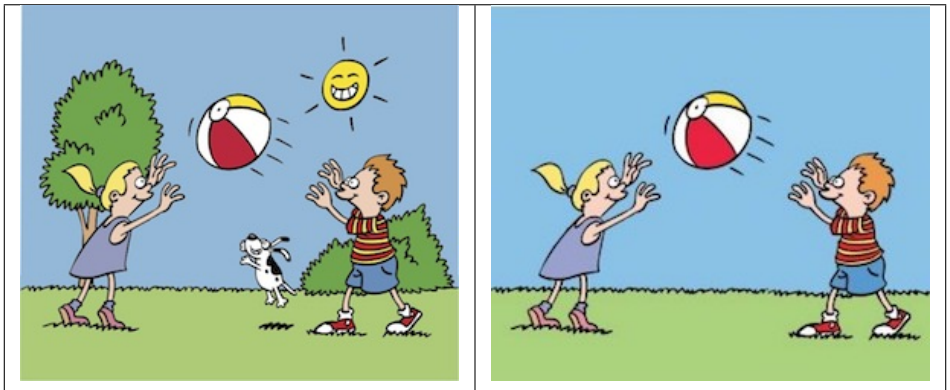
4.1 Picture Description Task

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 4.1. 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. More example images are presented in Figure 4.2 and the full set is found in Appendix A.

To factor out the influence of previous linguistic context, images are intentionally de-

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

void 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 difficult to portray visually, like TEACHING MATH and ASKING A QUESTION.



Original image	Simplified PDT image

Figure 4.1: 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 or "still life") performed by a person or an animal. The images are divided evenly into actions that are canonically intransitive, transitive or ditransitive in English. I chose these three categories because they indicate the number of actors and objects in a given event, and my approach to scoring responses should be able to handle this range of complexity. It should be noted that this categorization is imperfect, however, as some events in the PDT can be expressed in multiple ways, like *The girl is riding a horse* (transitive construction) versus *The girl is horseback riding* (intransitive construction). I attempted to minimize ambiguity (especially between intransitives and transitives) by avoiding images with possible light constructions, like *He is taking a shower* versus *He is showering*.

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 Figure 4.2).

		
What is the girl doing?	What is the boy doing?	What is the girl doing?

Figure 4.2: 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.

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. To elaborate, for targeted items, I expect less variation among responses; defining the subject in the prompt means all responses should reuse this subject and only vary in how they express the predicate. For untargeted items, some image prompts might allow for variation of the subject, however. For the image in Figure 4.1, for example, valid responses could include *The boy is throwing a ball to the girl* as well as *The girl is catching a ball from the boy*. Understanding the effect of the subject constraint 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 (see Section 4.2), 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 (3). Additional instructions provided to NSs are presented in (4). The full set of PDT versions is available for download with the SAILS Corpus.[3]

(3) **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.

(4) **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

4.2 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. Asking NSs to provide two responses doubles the length of the survey, exceeding the platform's limits on survey length for purchases responses, so the task was divided into two separate surveys for NSs. Thus while each NS and NNS provided 30 responses, each NNS responded to all 30 PDT items while each NS responded to only 15.

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. Unless specifically noted otherwise, 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)**. Future work should include collecting much more FNS data and comparison of the two groups to better understand the differences in quality, as CNSs are almost certainly 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 4.1.

	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 4.1: Age and gender information for the three participant groups (Non Native Speakers, Crowdsourced Native Speakers and Familiar Native Speakers).

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.

4.3 Response Totals

A total of 13,533 responses were collected. The response counts for each participant group are presented in Table 4.2. 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 4.3. 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 5). Indeed, automatically assigning low scores to inappropriate responses is a central challenge and goal in this project.

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

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 4.3: Crowdsourced responses for the items shown in Figure 4.2, showing one exemplar response and two examples of problematic or bad faith responses for each item.

4.4 Response Variation

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 case, however, I calculated type-to-token ratios (TTRs) on the response level for the entire set of items. For this calculation, final punctuation was ignored, and all responses were converted to lowercase. To illustrate, the first three response *tokens* in Table 4.4 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 4.4: This toy dataset shows how TTR is calculated on the response (sentence) 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 4.5. 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. TTR is highly sensitive to text length, so to control for the imbalance between NS and NNS responses, the TTR was calculated for each item and each group (NS and NNS) based on a random sample of 50 responses (Grieve, 2007). 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 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.

Additionally, from Table 4.5 we can see that the NS set contains a greater degree of

	Targeted		Untargeted	
Set	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 4.5: NS and NNS type-to-token ratios (TTR) for complete responses (not words), for the full corpus.

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 4.6, 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, indicating that a ceiling effect is at work. In other words, untargeted second responses are unconstrained by the task to such an extent that even the least complex responses—the intransitives—approach a level of variation roughly equal to the more complex transitive and ditransitive responses.

Finally, for ease of comparison, Table 4.7 presents the (NS only) first response TTRs from Table 4.6 alongside the NNS first (and only) response TTRs from Table 4.5. 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 4.6: 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 4.7: TTRs for complete responses, comparing first responses only.

Having examined response variation in a rather abstract sense here, Chapter 5 will focus on annotating response features to obtain a more fine-grained view of the ways in which responses can vary.

CHAPTER 5
ANNOTATION & WEIGHTING

Using the dataset introduced in Chapter 4, this chapter focuses on adding annotation to allow for content analysis. I begin with a discussion of the development and implementation of an annotation scheme that captures aspects of native-likeness and accuracy in the picture description task (PDT) responses. 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 are assigned to these binary features in order to determine a holistic score for each response.

5.1 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, responses were presented to annotators in random order and without any demographic information.

An ICALL system following my approach would crowdsource NS responses and use those to evaluate NNS responses, but of course such responses would not be annotated. Thus, by annotating the NS data collected in the current work, I can assess the quality of

crowdsourced NS responses for the task of evaluating NNS responses.

For NNS responses, such annotation could 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 annotations function as benchmarks which can be compared to scores provided by my automatic system, allowing for evaluation of the system itself (See Section 5.3). Furthermore, the annotation lends insights into which aspects of a response are the most difficult to account for in my approach to content assessment.

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

The scheme was initially envisioned 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 Figure 5.1, 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* (Figure 5.1) 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 would also lead to problems during the automatic scoring. If included in a 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 NS set. Furthermore, if such a NNS response is annotated as accurate, this additional information is unlikely to be readily mapped to information found in the NS set, which would lead to lower scores for the response. Accuracy, then, is an inadequate construct for the approach to content assessment envisioned for this work. Clearly, *verifiability* is an important consideration, as well.

In order to handle the issues discussed above, five binary features were developed, with each feature having some relation to the original concepts of accuracy and native-likeness. As with most annotation schemes in linguistics, the final SAILS scheme is a compromise. This scheme represents the minimal set of features necessary to accomplish two major goals of this work: investigating the use of NS responses as an evaluation model for NNS, 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. Moreover, this feature is largely subsumed by the others, as bad faith responses tend to score poorly across the board.

A set of annotation guidelines was produced with definitions, rules and examples for each feature. For most features, the rules for targeted and untargeted items (see Section 4.1) vary slightly; the untargeted rules are generally less strict to accommodate the less restrictive prompt question. The complete annotation guide is included in Appendix B. The features and brief descriptions are listed here and discussed further in the discussion of inter-annotator agreement in Section 5.2.

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 Figure 5.1, *[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. **ANSWERHOOD**: Does the response make a clear attempt to answer the prompt question? This generally requires a progressive verb, because the PDT questions are in the present progressive. For targeted items, the subject of the question or an appropriate pronoun must be used as the subject of the response. For example, *The dog is so happy!* (Figure 5.1) is answering a question other than *What is the woman doing?*.

3. **GRAMMATICALITY**: Is the response free from errors of spelling and grammar? This is a relatively straightforward feature to annotate. For example, from Figure 5.1, *She is wear a blue dress* contains an ungrammatical verb form.

4. **INTERPRETABILITY**: Does the response evoke a clear mental image (even if differ-

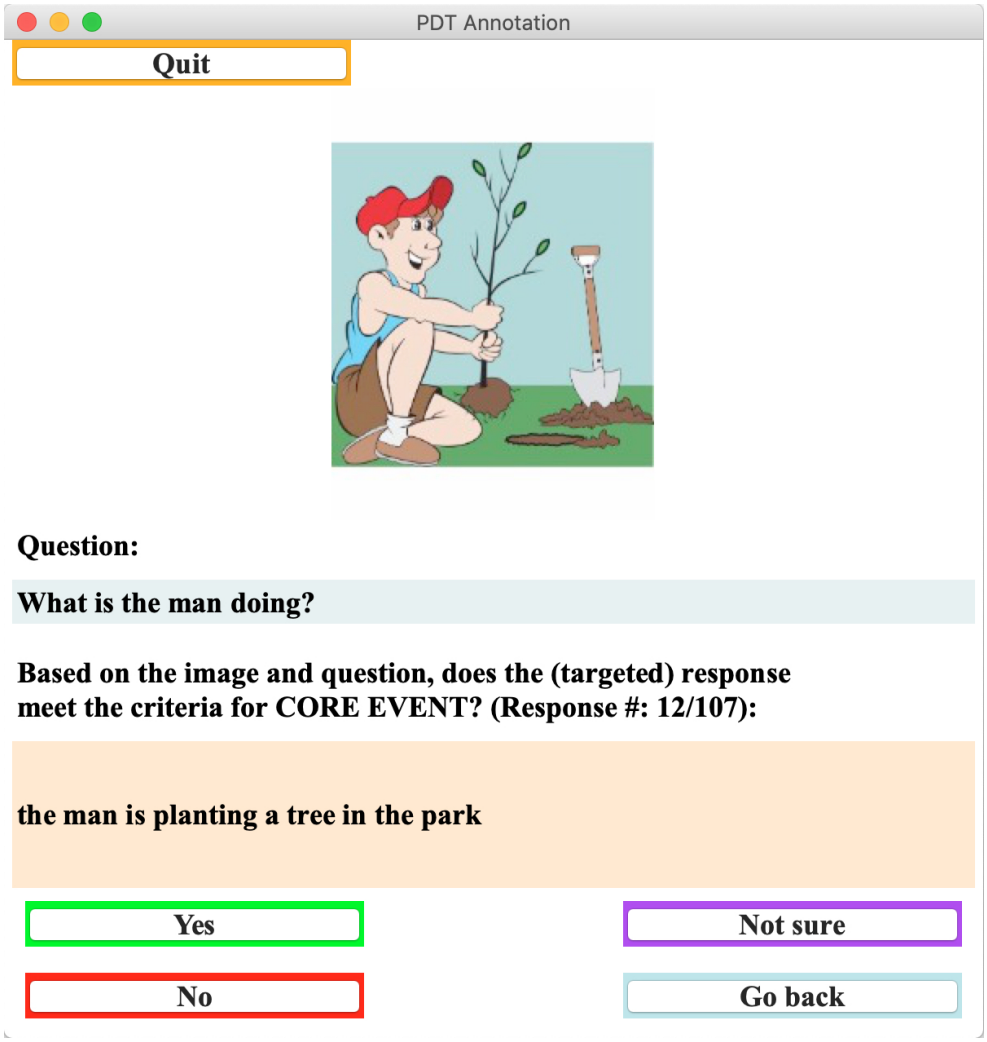


Figure 5.2: Interface used for feature annotations. Note that "Not sure" is not a final annotation value; it merely puts the response aside for a later decision.

ent from the actual item image)? Any required verb arguments must be present and unambiguous. For example, *She loves her pet* (Figure 5.1) 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 goldfish, etc.

5. **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 Figure 5.1, *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).

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 5.2. The annotations are written out to a spreadsheet.

Example annotations In Table 5.1, 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.

5.2 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.

What is the boy doing?	C	A	G	I	V
He is eating food.	0	1	1	1	1
eatting.	0	1	0	1	1
The child is about to eat pizza.	1	0	1	1	1
He may get fat eating pizza.	1	0	1	1	0

What is happening?	C	A	G	I	V
Child is eating pizza.	1	1	0	1	1
Tommy is eating pizza.	1	1	1	1	0
The boy's eating his favorite food.	0	1	1	0	0
Pizza is this boy's favorite food.	0	0	1	0	0

Table 5.1: Targeted and untargeted sample responses from the development set transitive item, shown with adjudicated annotations for CORE EVENT (*C*), ANSWERHOOD (*A*), GRAMMATICALITY (*G*), INTERPRETABILITY (*I*) and VERIFIABILITY (*V*).

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 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.

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

Figure 5.3: 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.

The test set parallels the development set and consists of one intransitive, one transitive and one ditransitive item; it is shown in Figure 5.3. Agreement and Cohen's kappa scores are given in Table 5.2, 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.

5.2.1 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 4.5), 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.

Set	Total	A1Yes	A2Yes	AvgYes	Chance	Observ	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 5.2: 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 (*Observ*) and Cohen's kappa (*Kappa*).

5.2.2 Targeting

Grouping the annotations into targeted and untargeted sets, the raw agreement scores are comparable (94.9% and 95.2%, respectively). However, despite a greater degree of response variation, the untargeted group has a higher kappa score (0.872 versus 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.

5.2.3 Features

Grouped by feature, the annotations all show raw agreement scores above 91% and Cohen's kappa scores above 0.74 (Table 5.2). 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 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 described in the response, 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 5.2.1).

The main issue in this case has to do with the amount of specificity required to capture the core event. The development set ditransitive 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 (Figure 5.3), and this resulted in a greater degree of variation. This is confirmed by the lower type-to-token ratio (TTR) of main verbs among development set responses versus test set responses (0.189 versus 0.247), as presented in Table 5.3. 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 (Figure 5.3), 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

	Development Set		Test Set	
Version	Types/Tokens	TTR	Types/Tokens	TTR
NNS Target	12/71	0.169	16/70	0.229
NS Target	32/157	0.204	37/156	0.237
NNS Untarg	14/70	0.200	18/71	0.254
NS Untarg	33/180	0.183	36/134	0.269
Average	22.8/119.5	0.189	26.8/107.8	0.247

Table 5.3: Comparing type-to-token ratios (*TTR*) for **main verbs** among the development and test set **ditransitive items**; greater variation correlates with lower CORE EVENT inter-annotator agreement, which helps explain why in Table 5.2 CORE EVENT agreement is lower than agreement for other features.

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

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 5.1);

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 sometimes obscene) at the cost of at-

tempting 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 5.1).

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 below).

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.

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*.

Interpretability The average *yes* rate for INTERPRETABILITY is 0.802; only CORE EVENT is less skewed. 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 verb 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 or placing it in a mailbox or on a doorstep, as an indirect object is not necessary to convey the meaning of the verb *deliver*.

Among the 105 annotator disagreements, fatigue accounts for roughly 30; this is difficult 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 disagreements, 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* (Figure 5.3). 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 annotating this feature. As mentioned above (see VERIFIABILITY below), expanding the guidelines might help cover some such situations, but likely at the cost of increased annotator fatigue.

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 illustra-

tion 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. 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 demand and fatigue.

5.2.4 NS & NNS responses

Response quality and annotation agreement were also calculated separately for NS and NNS responses, as shown in Table 5.4. 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 English 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.

Turning to the question of annotation quality, raw agreement scores are high among

	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 5.4: 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*).

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 responses. 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.

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 Figure 5.3), but the few grammar problems in the NS data are mostly typos and more easily overlooked (e.g., *The man is giving* ***ditections***).

5.3 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. These could simply be intuitively chosen, but a data-driven approach would be both more justifiable and more reliable. One might consider starting by manually ranking the responses. With responses ranked from best to worst, the distribution of annotations across this ranking could be used to determine some coefficient that represents the importance (weight) of each feature in the rankings. However, 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 a holistic preference test with pairs of responses. With enough of these decisions, it becomes possible to derive annotation weights.

The full corpus consists of 13,533 responses across 30 items (30 images each presented with a targeted and untargeted prompt; see Section 4.3). For the preference test to determine feature weights, a sample of 1200 response pairs was used: 20 targeted and 20 untargeted response pairs from each of the 30 PDT items. 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 preference 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 pair contained responses with identical vectors, as nothing is learned by comparing two *perfect* responses, for example.

Annotator 1 (A1) performed the preference test for all 1,200 of the sampled response pairs. Annotator 2 (A2) performed the preference 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 5.2.

Annotators were given the following instructions for the preference 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 a better 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.

The preference test interface (Figure 5.4) was similar to that used for annotating the features. For each preference 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 if they tried.

Example response pairs and decisions are shown in Table 5.5. 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 preference test, 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.

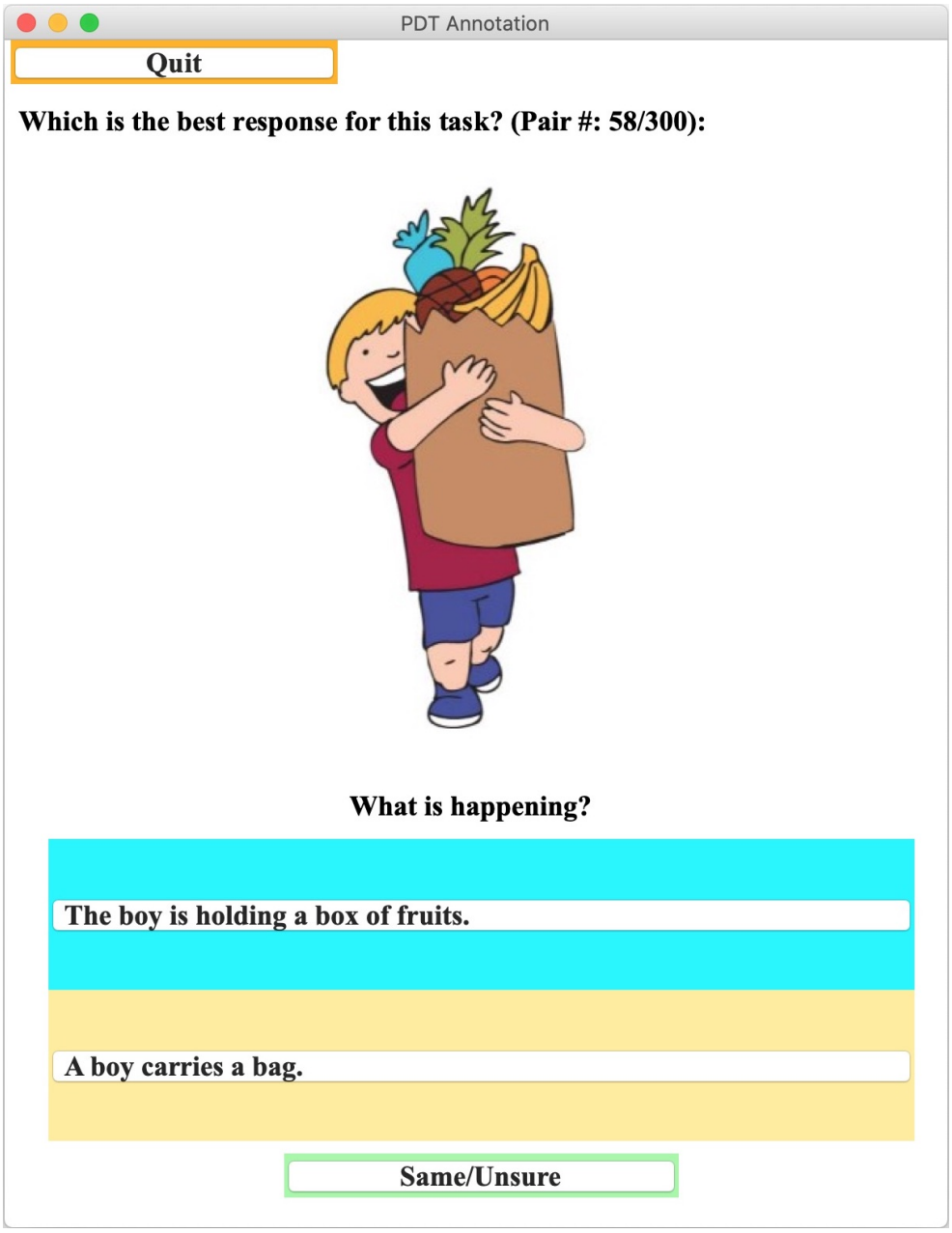


Figure 5.4: Annotation interface used for the preference test.

For the two disagreements shown in the table, one could make a reasonable argument for preferring either response (or marking them *same* in quality); this is true for most of the 35/300 disagreements in the sample. Disagreement over *The boy is holding a box of fruits* and *A boy carries a bag* 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

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 5.5: Preference test sample responses pairs, annotator decisions (*A1 & A2*) and agreement for the item shown in Figure 5.4.

a bag) and VERIFIABILITY (with *box* being quite clearly inaccurate). The disagreement in the third pair 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 B would likely not satisfy the CORE EVENT feature, while in response A, *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 5.6. 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). Moreover, the disagreements appear to be noise spread among all features, rather than an indication of difficulty with a particular feature. 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 5.6: Preference test agreement scores for two annotators on a sample of 300 responses pairs, showing 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 a test pair was subtracted from the total number of times that feature

occurred with the preferred response to yield the net count for that feature. Pairs ruled *same* (no preference) were omitted. The net counts of all five features were summed. The net count for each feature was 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 5.7.

	CORE	ANSWER	GRAMM	INTERP	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 5.7: 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 preference test 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.

5.4 Holistic Scoring and Ranking

The feature weights established by the preference tests can now be applied to the binary annotations to produce a holistic score for each response, which I call the **weighted annotation score**.

What is happening?	C	A	G	I	V	WAS	WAR
Child is eating pizza.	0.365	0.093	0.000	0.224	0.263	0.945	1
Tommy is eating pizza.	0.365	0.093	0.056	0.224	0.000	0.738	2
The boy's eating his fav...	0.000	0.093	0.056	0.000	0.000	0.514	3
Pizza is this boy's fav...	0.000	0.000	0.056	0.000	0.000	0.056	4

Table 5.8: Example NNS responses (see Table 5.1) with feature weights applied to the binary annotations, resulting in weighted annotation scores (*WAS*) and a weighted annotation ranking (*WAR*).

For each item, I rank the NNS responses according to this score. I call the resulting ranking a **weighted annotation ranking**. Because these rankings are based on human annotator decisions, they can serve as a benchmark for comparing the output of an automatic scoring system, as discussed in Chapter 6.

The NS responses were also annotated for the binary features, but in the current work there is no practical use for a weighted annotation ranking of the NS responses. However, by applying the weights and obtaining weighted annotation scores for the NS responses, I can compare the holistic performance of NS and NNS participants. As shown in Table 5.9, in terms of weighted annotation scores, *familiar native speakers* (*FNSs*; see Section 4.2) outperform NNSs, who outperform *crowdsourced native speakers* (*CNSs*). The FNSs have the highest rate of perfect responses, the lowest rate of zero-scoring responses and the highest mean weighted annotation scores. Moreover, the standard deviation shows that FNSs have the least varied weighted annotation scores, which makes sense as scores for this group are heavily skewed toward the upper limit. In each of these measures, the FNSs are followed by the NNSs, then the CNSs. It should be noted that these scores cover the entire set of all responses for all items, which includes the two responses from NSs per item, whereas the NNSs provided only a single response per item.

The relative performance of these three groups has important implications for how this data is used and how data for related purposes might be collected in the future. Because my work is geared toward communicative applications like intelligent computer-assisted

	NNS	CNS	FNS	C+F
Total	4230	7723	1580	9303
Perfect	0.614	0.495	0.692	0.528
Zero	0.011	0.048	0.001	0.040
Mean	0.862	0.763	0.880	0.783
Median	1.000	0.944	1.000	1.000
Std Dev	0.248	0.323	0.226	0.312

Table 5.9: Comparing scores for non-native speakers (*NNS*), **crowdsourced** native speakers (*CNS*) and **familiar** native speakers (*FNS*) across all items. *C+F* is the combination of CNS and FNS (i.e., *all* NS). *Total* is the response count. *Perfect* and *Zero* are the rates of responses with weighted annotation scores of 1.0 and 0.0, respectively. The *Mean*, *Median* and *Standard Deviation* values here are weighted annotation scores.

language learning (ICALL) rather than grammatical error correction or placement testing, the models I use must be flexible enough to process NNS responses without heavily penalizing them for minor issues in form or usage. For my purposes, models built only from FNS responses would be overfitted to the near-perfect language usage of the FNS group. Models built from the slightly noisier CNS responses are preferable for my purposes. In effect, they will allow for more of the variations in form and usage that we see among the NNS responses. In other words, an FNS-based model will accurately identify those responses that closely match a limited set of well-formed possibilities, but it will harshly penalize minor deviations, resulting in scores that tend toward the high and low ranges. A CNS-based model will be less strict; slight deviations from ideal, "native-like" possibilities will be penalized less harshly, and scores will be distributed more evenly across the range from 0.0 to 1.0.

Due to the sparsity of the FNS data, it is not possible to analyze it on a more granular level. With more data from the CNS group, however, it is possible to break down the CNS metrics seen in Table 5.9 for a closer look, as shown in Table 5.10.

The trends seen in the CNS table confirm my intuitions about the data—scores decrease as complexity and variation increase. First response scores are higher than second response scores. This makes sense given my observations from Chapter 4, where I found a higher

rate of bad faith or low-effort answers among the second responses, likely owing to fatigue and boredom, and a higher type-to-token ratio, indicating a higher rate of unique responses (see Table 4.6). This increase in variation means more creative responses, which are more likely to include unverifiable details, for example. Similarly, untargeted response scores are lower than targeted response scores. This can also be explained by the greater degree of variability among untargeted responses, as discussed in Chapter 4 (see Tables 4.5, 4.6 and 4.7). This is expected, because the subject is provided in targeted prompts but not untargeted prompts, so naturally the untargeted responses include more cases where the subject is incorrect, irrelevant or unclear. Finally, among the item (verb) types shown in Table 5.10, the scores decrease as the complexity increases; intransitive scores are highest, followed by transitives and then ditransitives. Again, this correlates with type-to-token ratios (see Tables 4.5, 4.6 and 4.7).

	R1	R2	Target	Untarg	Intran	Trans	Ditran
Total	3872	3851	3877	3846	2592	2569	2562
Perfect	0.623	0.366	0.535	0.454	0.519	0.502	0.463
Zero	0.037	0.058	0.043	0.053	0.040	0.051	0.053
Mean	0.838	0.687	0.772	0.753	0.775	0.757	0.756
Median	1.000	0.851	1.000	0.944	1.000	1.000	0.907
Std Dev	0.286	0.340	0.315	0.330	0.319	0.330	0.320

Table 5.10: Examining **crowdsourced** native speaker response scores in different contexts: first and second responses (*R1* and *R2*); targeted and untargeted prompts; intransitives, transitives and ditransitives. (See Table 5.9.)

These observations strengthen my conviction that CNS-based scoring models are preferable for my approach, while higher-quality, more uniform FNS-based scoring models would be preferable for stricter contexts like language assessment or placement testing. Naturally, the best way to score NNS responses would be to use models trained on NNS responses. However, these responses would need to be validated or classified in some way by annotators, which would require developing new annotation guidelines. The idea is also counter to the motivations behind my work, because it means every new item would require man-

ual annotation by experts in order to train a scoring model, which places a greater burden on educators or researchers who might follow my approach. Ideally, a system could train an initial model based on unannotated CNS responses; after scoring some NNS responses, it could then add examples of the highest scoring NNS responses to the training data and retrain iteratively—an approach known as self-training. I do not explore the use of NNS responses as training data, but future work to do so can build on the work here.

5.5 Annotation Conclusions

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 the annotations and raw responses would also be useful for question answering, dialog systems, pragmatics modeling, visual references and other challenges in natural language processing. The feature set can also be expanded to better suit other purposes, and the task can easily be extended to include new items. To facilitate expansion, 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 is teaching '2 + 2 = 4'*, rather than *The teacher is teaching math*). The comparison of crowdsourced NS data with the data of familiar 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

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

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 desirable or cost effective (see, for example, Somasundaran and Chodorow (2014)).

CHAPTER 6
OPTIMIZATION

In this chapter, I seek to answer questions surrounding the quality of the dataset and its annotation scheme, the feasibility of implementing a transparent, dependency and tf-idf based system for approximating manual annotations, and metrics for evaluating my approach. I see this work as a proof-of-concept, where finding consistent, exploitable trends and correlations validates the overall project. The findings here are presented not as hard and fast universal truths, but as indications of the utility of the annotations and when and where such a ranking system can find success.

This takes the form of experiments aimed at optimizing the configuration of my system for rating and ranking responses automatically, first introduced in Chapter 3. This means using my system with various settings to produce response rankings, then examining how these system rankings correlate with an annotation-based ranking to find which system settings work best. Specifically, I look for correlations between the performance of my system (and its particular settings) and known features of the native speaker (NS) and non-native speaker (NNS) responses, such as the transitivity of the PDT item event, the size of the NS model, and whether models contain only primary responses or a mix of primary and secondary responses. I also consider correlations between system performance and observable measures of the NS and NNS data, namely type-to-token ratios and mean response lengths.

My earliest attempts at ranking responses were rule-based and relied on strict matching with a pre-established set of acceptable responses, described in Section 3.2. This found moderate success, leading to the improved approach described in Section 3.3, which is data-driven and relies on more flexible methods of comparison. In Section 6.1, I give an overview of the updates to this approach used here to process the new, larger dataset described in Chapter 4.

I present some initial statistics and observations regarding the NS models and NNS test sets in Section 6.4, as these observations help explain trends seen in the experiments that follow. Ideally, for each of the five annotation features introduced in Chapter 5, my system rankings should maximize the separation of positively and negatively annotated NNS responses. In Section 6.5, I examine how well each annotation feature correlates with my dependency-based similarity results and which system settings and model sizes maximize these correlations. In Section 6.6, I step back from trying to approximate the individual feature annotations with my system, looking instead to automatically rank NNS responses according to overall quality and determine which settings produce rankings that best correlate with the desired weighted annotation rankings described in Section 5.4.

6.1 Updated methodology

The work discussed in Chapter 3 relied on a shaky implementation of "correctness" or "appropriateness" for responses, and this needed improvement, first and foremost. Developing a better and more reliable annotation scheme was a key goal for me in expanding the work to its current scale, in order to give the work more meaning and context and make my corpus useful for a broad range of uses. The annotation planning discussed in Section 5.1 was a direct result of the challenges of working with an inadequate annotation scheme.

As discussed in Sections 5.3 and 5.4, in the current corpus, annotations no longer directly encode for errors, but instead give a binary score for five different features, which can then be weighted and combined to produce a score between zero and one. As a result, the of metrics for judging system performance have changed. Mean average precision (MAP), previously used with the "good response"/"bad response" annotations (Section 3.3.5), is now used to judge system performance focused on individual annotation features, presented in Section 6.5. I also compare system produced rankings against the weighted annotation rankings for a holistic approach to evaluating responses; i.e., how well does the system rank responses in comparison to a ranking derived from all five manual feature annota-

tions? As presented in Section 6.6, I evaluate the holistic experiments with Spearman rank correlation, which is a measure of correlation between two sets of rankings (Dodge, 2008).

Because the annotations and evaluation are different in the current work, it does not exactly follow that findings from the pilot study work will hold true. However, I believe that the previous work has highlighted some of the system settings that are most likely to perform well, and I chose to focus my experiments on some of the best performing settings to allow me to also examine some previously unexplored parameters. For example, all of the current experiments rely on the tf-idf cosine (`TC`) approach, as this generally outperformed the others. As discussed in Section 3.3.5, the `TC` performance suffers for items where the non-native speaker (NNS) data is noisiest (with regard to spellings) and the native speaker (NS) data is relatively homogenous (particularly with regard to verb choice). Rather than continue experimenting with the underperforming approaches, I chose to address these issues directly instead. As discussed in Chapter 4, to address the spelling noise, I made sure that data collection participants had access to spelling correction while typing their responses. To address the uniformity of the NS responses, I surveyed a much larger group of participants and instructed each of them to provide two responses per picture description task (PDT) item. In other words, the limitations of the TC model have already been addressed, and so I expect results in this chapter to be generally applicable.

The current work retains just three of the five `term representations` previously used (see Section 3.3.2): `ldh`, `xdh`, and `xdx`. The `ldh` and `xdh` forms performed best with the older dataset. Moreover, as these represent labeled (`ldh`) and unlabeled (`xdh`) dependencies, their use in linguistics is well established. The `xdx` representation is kept here as a rough equivalent of a bag-of-words model, another well-established linguistic representation.

Finally, as the Brown Corpus overwhelmingly outperformed the Wall Street Journal corpus as a tf-idf reference, the current experiments rely exclusively on Brown. The larger PDT dataset follows a similar narrative style to that described in Section 3.2.1, so I am

transitivity	targeting	familiarity	primacy	Term Rep.
intransitive	targeted	Familar	primary	ldh
transitive	untargeted	crowdsourced	mixed	xdh
ditransitive				xdx

Table 6.1: All parameters or variables and their settings; a system configuration combines one setting from each column.

confident Brown is again the best option here.

All experiments throughout this chapter score and rank 70 NNS responses per item; this is the maximum number of NNS responses available across all PDT items. Where more than 70 responses are available, a random sample of 70 is used. The experiments in Chapter 3 used 30 different combinations of system settings (i.e., **configurations**) for each of 10 items. By comparison, the variables and parameters used in this chapter result in up to 72 different configurations (see Table 6.1). Subtracting transitivity (which cannot be varied for a given item) leaves 24 configurations which apply to all 30 PDT items. To make sense of this large number of results, I have chosen to focus my optimization efforts on each parameter individually, rather than attempt an exhaustive search through all configurations. I begin this by producing NNS response rankings for all PDT items using all 24 system configurations. Then, in order to compare performance for targeted and untargeted settings, for example, I form one pool of results from all configurations including the targeted setting and another pool from all configurations including the untargeted settings.

To help contextualize the results of these experiments, I used a state-of-the-art language modeling tool called SBERT to rank responses according to their similarity to the NS model. I discuss this tool and its use as a benchmark in Section 6.3.

I begin the optimization experiments in Section 6.6 with a new parameter that I call term normalization. In the usual non-normalized setting, all *terms* in the NS model carry equal weight, but in the normalized setting, all *responses* carry equal weight. This is achieved by normalizing the weight of each NS term in the model accord-

ing to the length of the response in which it appeared. While I expected the `normalized` setting to improve performance, the opposite was true, so I do not include it elsewhere in system configurations. I discuss the findings in Section 6.6.1, before moving on to focus on the more useful parameters that I retained in configurations throughout this chapter.

These sections are organized according to the sequence in which the parameters are relevant in my process, which begins with data collection and ends with scoring and ranking non-native speaker (NNS) responses. Thus, in Section 6.6.2, I begin with the variable I refer to as `transitivity`, which emerged during task design for the PDT described in Chapter 4; I look at the effects of applying my dependency-based tf-idf cosine pipeline to new item types, namely `intransitives` and `ditransitives`, and compare against performance on `transitive` items. Next, in Section 6.6.3, I turn to experiments regarding `targeting`, which refers to whether or not the PDT item subject was referenced in the prompt (as discussed in Section 4.1). In Section 6.6.4, I examine a variable I call `familiarity`, which refers to whether the native speakers (NSs) contributing to the model are `familiar` to me personally or are `crowdsourced`. Another new variable follows in Section 6.6.5; I call this `primacy`, which refers to whether the NS model contains only first (`primary`) responses, or an equal number of first and second (`mixed`) responses (also discussed in Section 4.1). Note that I do not investigate the use of models comprised only of secondary responses. This is deliberate, because in real use cases there would never be a scenario in which I have secondary responses but not primary responses. I return in Section 6.6.6 to the best performing dependency `term representations` from Section 3.3.2 to see how they perform with the current dataset.

6.2 Sampling NS response models

Throughout this chapter, the experiments are performed with randomly sampled NS models of two sizes in order to examine the effects of model size on system performance. The larger models contain 50 NS responses per PDT item. This is the maximum number of

NS responses that are available across all PDT items using relevant system configurations. For example, item 26 in the corpus contains 290 total NS responses, but roughly 90% are `crowdsourced`, 50% are `targeted`, and 50% are first (`primary`) responses, and selecting for this configuration leaves exactly 50 responses from which to form a model. The smaller models throughout this chapter contain 14 NS responses per item. This is the maximum number of responses available across all items from `familiar` NSs, so I chose this size in order to fairly compare `crowdsourced` and `familiar` responses (Section 6.20) and I retained it throughout this chapter to best contextualize the `familiarity` experiments.

6.3 SBERT as a benchmark

The central task of my work—ranking NNS responses using a set of NS responses—is not a standard task with established metrics in any relevant field like natural language processing (NLP) or language testing. Moreover, this work relies on a custom dataset which has not been widely adopted. These facts make it challenging to assess the performance of my ranking system and its various configurations or to compare this work against similar research. In order to give some frame of reference for this work, I chose to use a widely adopted language embedding tool known as SBERT (Reimers and Gurevych, 2019). In practice, this means swapping the sentence similarity component of my pipeline (dependency tf-idf cosine) for SBERT.

SBERT is a modified version of BERT (Bidirectional Encoder Representations from Transformers), which is a pre-trained transformer network that has set records on NLP tasks including semantic textual similarity (STS), topic modeling and question answering (Devlin et al., 2018). I choose SBERT over BERT because it is faster, more efficient and more readily usable on a typical personal computer. Specifically, I use the pre-trained version of SBERT distributed in the `sentence_transformers` Python package (Reimers and Gurevych, 2020). For semantic textual similarity, SBERT is trained on the Stanford

Natural Language Inference (SNLI) corpus (Bowman et al., 2015) and the Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018). These two corpora consist of sentence pairs manually labeled as *contradiction*, *entailment*, or *neutral*; the SNLI corpus contains 570,000 sentence pairs, and the MutliNLI corpus contains 430,000 sentence pairs.

My system scores each single NNS test response according to its similarity with the set of NS model responses. Measuring semantic textual similarity is one of SBERT's most used functions, so for each scoring and ranking experiment in this chapter, I use SBERT to generate corresponding output. The resulting mean average precision (MAP) and Spearman rank correlation scores are discussed throughout to help contextualize my system's performance.

For each experiment reported throughout this chapter, my system and SBERT are given access to the same NS responses as the basis for their similarity measures. Naturally, however, each is trained on or makes use of very different language resources. The linguistic "intelligence" of my system comes largely from the Stanford Parser and its pre-trained grammar model, as discussed in Section 3.2 (Klein and Manning, 2003). The model is trained on the standard training sections of the Penn Treebank, which contain over one million words of English text, manually part-of-speech tagged and parsed, sourced from the Wall Street Journal and the Brown Corpus (Marcus et al., 1993). My approach also uses the Brown Corpus for tf-idf, meaning a word (or more accurately, dependency) frequency model extracted from the Brown Corpus also serves as a linguistic resource (Kucera and Francis, 1967). SBERT, on the other hand, is trained on a million sentence pairs, each manually labeled to indicate similarity.

My use of SBERT here varies slightly from the way I implemented my own system's similarity measuring approach throughout most of this chapter. In my system, each NS response in the model is processed and dumped into a single "bag of dependencies," which is then used to generate a single similarity score (via tf-idf cosine). SBERT operates directly

on plain text sentences, and because the use of punctuation in the PDT responses is not consistent, concatenating all NS model responses in order to generate a single similarity score is not ideal. Instead, I use SBERT to do a pairwise comparison between the NNS test response and each NS response in the model and then average these similarity scores.

6.4 Sample statistics

Before jumping into the experiments in Sections 6.5 and 6.6, I present here some initial statistics for the samples used as NS models and NNS test sets. In general, I expect the NS models that most resemble the NNS test sets to perform best in the ranking experients, so these statistics may shed light on the experimental results that follow.

	n=14		n=50	n=70
	Fam	Crowd	Crowd	NNS
Intrans	5.5	4.9	4.9	4.9
Trans	6.9	6.3	6.2	6.7
Ditrans	7.8	7.2	7.2	8.3
Target	6.5	5.4	5.4	6.3
Untarg	6.9	6.8	6.8	6.9
primary	N/A	5.7	5.8	6.6
mixed	6.7	6.5	6.4	N/A
Total	6.7	6.1	6.1	6.6

Table 6.2: Comparing average response length (in words) for the samples used throughout this chapter as NS models and NNS test sets, in total and by parameter setting.

Table 6.2 presents average response length for the samples, in total and broken down by parameter settings. As expected, for all samples we see an increase in response length as we move from intransitives to transitives to ditransitives. The same is true in moving from targeted to untargeted settings, and from primary to mixed settings (where applicable). These numbers also show that in most cases, the familiar responses are the longest and the crowdsourced responses are the shortest, with NNS responses falling somewhere in between. The only exception is for ditransi-

tive items, where the NNS responses are longest. By comparing the 14-response and 50-response crowdsourced samples, we can see that the response length is quite stable across sample sizes.

As an indication of complexity or lexical density, in Table 6.3, I present the standardized type-to-token ratios (STTR) for the response samples. Higher complexity inherently means greater data sparsity and more low-frequency events, and this generally necessitates the use of richer models and more sophisticated methods in natural language processing (Malvern et al., 2004). I use STTR as opposed to TTR as a means to normalize for large differences in size, with the smallest samples containing only 14 responses (per item), and the largest containing 70. A *standardized* type-to-token ratio, sometimes referred to as a *mean segmental* type-to-token ratio, simply calculates the TTR for each segment of *n* words in the sample, then averages these TTRs at the end (Johnson, 1944; Richards and Malvern, 2000). Here I use a window of 40 words, as this is the maximum available for all samples.

	n=14		n=50	n=70
	Fam	Crowd	Crowd	NNS
Intrans	0.558	0.525	0.535	0.391
Trans	0.569	0.580	0.581	0.517
Ditrans	0.598	0.640	0.637	0.606
Target	0.545	0.535	0.545	0.481
Untarg	0.610	0.633	0.621	0.528
primary	N/A	0.517	0.523	0.505
mixed	0.576	0.652	0.645	N/A
ldh	0.665	0.664	0.671	0.578
xdh	0.658	0.661	0.660	0.572
xdx	0.364	0.424	0.421	0.364
Total	0.576	0.583	0.584	0.505

Table 6.3: Comparing average standardized type-to-token ratio (STTR) for the samples used throughout this chapter as NS models and NNS test sets, in total and by parameter setting. Tokens here are *dependencies*.

Within each parameter, STTR shows a similar pattern to response length. Complexity increases as we move from intransitives to transitives to ditransitives,

from targeted to untargeted settings, and from primary to mixed response models. As the representation is simplified from labeled dependencies to unlabeled dependencies to dependents only, STTRs decrease. For example, as seen in Table 6.3, for the 70 response NNS set, these values decrease from 0.578 to 0.572 to 0.364. This decrease is small when moving from labeled to unlabeled dependencies, indicating that for a given head and dependent, syntactic labels tend to be the same. A much larger decrease is observed when moving from unlabeled dependencies to dependents only, indicating that dependents occur with a wider variety of heads than labels. This is to be expected, but it may be useful in understanding differences in performance across transitivity types, for example; ditransitives mean more verb arguments than intransitives, and this means more variety in the combinations of heads and dependents.

In comparing across the models, the story is complicated. First, STTR is less stable than response length across sample sizes, as seen between the 14-response and 50-response crowdsourced models. One consistent pattern seen here is that across the board (i.e., on any given row of Table 6.3), the 50-response crowdsourced models are least like the NNS test sets. This is a pattern which plays out in the following sections. In the STTRs seen here, the NNS test sets are sometimes closest to the 14-response familiar models, and at other times closest to the 14-response crowdsourced models. As I will point out in the following sections, this aligns with some of the patterns seen in the feature and holistic experiments.

The average response length and STTR statistics presented in Tables 6.2 and 6.3 are both calculated before any similarity scoring and ranking, and do not require annotation. This means that any predictive patterns observed there could be easily applied to new items. In Table 6.4, I present similarly organized statistics for response scores. These scores are averages of response scores, which represent the *distance* between an NNS test response and an NS model. Note that there is no column for the NNS sample here because the NS columns are calculated according to how they score that same NNS 70-response sample.

	NS = 14		NS = 50
	Fam	Crowd	Crowd
Intrans	**0.410**	**0.382**	***0.342***
Trans	0.490	0.504	0.465
Ditrans	0.576	0.601	0.563
Target	**0.466**	**0.470**	***0.436***
Untarg	0.517	0.521	0.478
primary	N/A	**0.475**	***0.447***
mixed	0.492	0.516	0.466
ldh	0.553	0.557	0.513
xdh	0.541	0.546	0.503
xdx	**0.382**	**0.383**	***0.353***
SBERT	0.675	0.640	*0.641*

Table 6.4: Comparing average NNS response scores across parameter settings and NS models, using the dependency tf-idf cosine scoring approach introduced in Section 3.3.3. The same sets of 70 NNS responses per model and configuration were scored here and throughout this chapter. Scores represent the NNS *distance* from the NS model, so lower scores are closer to NS behavior. Within each parameter, the score for the setting that minimizes distance is **bolded**, and the score for the model that minimizes distance is *italicized*.

Naturally, such figures would not be available for new items, but they also do not require annotation, only similarity scoring.

These scores average over all NNS test response scores—the response scores that are used to rank responses. Because my work is more interested in the ranking of responses than the scores, a given average score in the table is not particularly meaningful. In combination, however, they give an indication of how skewed the scores are for each model and parameter setting. In all cases, the 50-response `crowdsourced` models do best at minimizing the distance between NNS test responses and the model. As seen in the experiments in Section 6.6, these average scores may be predictive of my system's performance at holistic ranking, where the larger NS models provide better response rankings.

6.5 Annotation features experiments

In this section, I revisit the five annotation features discussed in Chapter 5 to see how they correlate with the performance of my system. For a given feature, I use mean average precision (MAP) to see how well the system rankings separate the positively and negatively annotated responses. In other words, for the purposes of calculating MAP for the CORE EVENT feature, a "0" annotation for CORE EVENT is treated as an error. As discussed in Section 3.3.5, average precision is used as a measure of how well a ranking separates errors from non-errors. I used the `average_precision_score` implementation from the Scikit-learn package for Python, which explains that average precision "summarizes a precision-recall curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight," (Pedregosa et al., 2011). The formula is presented in Figure 6.1. I refer to *mean* average precision throughout this section because I calculate average precision for all 30 items, then average these values for the purpose of evaluating the system configuration that produced the rankings.

$$\text{AP} = \sum_{n} (R_n - R_{n-1}) P_n$$

Figure 6.1: The formula for average precision, where P_n and R_n are the precision and recall at the nth threshold.

I isolate these MAP scores for each setting within the `transitivity`, `targeting`, and `primacy` variables, as well as in total. I also compare across `term representations`. To contextualize these scores, I include MAP scores for the weighted annotation rankings (WAR) and SBERT rankings. Note that the high WAR MAP scores seen here are to be expected. First, the weights are based on the preference judgments of the same annotators who developed the annotation scheme. More importantly, the WAR is derived from weighting and combining the five feature annotations for each response, so

precision in assessing any one of these features is inherently correlated to the WAR. The function of combining a response's five binary annotations into a single annotation score using weights is a form of lossy compression; the WAR MAP gives a sense of just how lossy this weighted compression function is in terms of recovering the annotations for the five individual features and across different item types. In other words, the WAR MAP scores can be seen as an upper bound on performance, given this annotation scheme and its weights.

These experiments would be useful to anyone considering an approach to content analysis like mine. In an intelligent computer-assisted language learning (ICALL) game, for example, there may be times when CORE EVENT is the main concern, or others where ANSWERHOOD or another feature is most relevant. By isolating each annotation feature and examining how model sizes and `term representations` effect MAP—both overall and for individual variables, such as `intransitives` or `targeted` items—I observe a number of trends that would be helpful in designing an ICALL system that selects an optimal system configuration for handling each user response.

I present here two tables for each of the five annotation features. For each feature, this first table presents MAP scores for models comprised of `crowdsourced` responses, showing both the 14-response and 50-response models. This table serves to examine whether the task of recognizing the feature is sensitive to differences in model size, and whether parameters like `targeting` or `primacy` interact with any such differences. The second table compares MAP scores for the `familiar` and `crowdsourced` NS models, using models of 14 responses each. Note that the first table covers both `primary` and `mixed` response models, but the second table covers only `mixed` response models (for both `familiar` and `crowdsourced` responses), because for some items, 14 is the maximum number of `familiar` responses available (including first and second responses). MAP differences between the `crowdsourced` 14-response models in each first table and corresponding second table stem from this difference. In all 10 tables, I use **bold** to indicate

the highest `term representation` MAP *within* each of the two models presented and ***bold with italics*** to indicate the highest MAP *between* the two models.

Overall, a few trends stand out, which I outline here and discuss in detail in the following sections. First, both as a trend and a caveat, note that in many cases the relevant differences in MAP scores observed here are small, and may not always be statistically significant. The observations here are thus not intended as specific guidance for high stakes decisions, but in aggregate as a proof-of-concept for the overall project, from the annotated data to the content-focused, similiarity-based analysis. Individually, the trends discussed in this chapter serve as an indication of promising directions for future work.

One notable observation is the underperformance of unlabeled dependencies (`xdh`). In some cases, labeled dependencies (`ldh`) achieve the highest MAP, and in some cases dependent-only terms (`xdx`) work best, but there are no cases in which unlabeled dependencies win. This suggests that future iterations of my system could safely eliminate the use of unlabeled dependencies for the sake of simplicity. I also observed that for the features where labeled dependencies work best (CORE EVENT, INTERPRETABILITY, and VERIFIABILITY), the `crowdsourced` NS models also work best, but for the features where the dependent-only terms work best (ANSWERHOOD and GRAMMATICALITY), the `familiar` NS models work best.

The MAP scores also show that with only one exception, SBERT underperforms the system. This means that with regard to recognizing custom annotation features, a custom pipeline based on dependency parsing and tf-idf can outperform newer, more sophisticated machine learning approaches. Another notable trend seen here is that `transitive` items are often an exception; where `intransitives` and `ditransitives` see higher performance with a given model size or `term representation`, `transitives` frequently differ. Yet another observation is that with regard to NS model size, less is overwhelmingly more. For each of the five features, the total MAP (which covers all rankings for all items, provided by all available system configurations) is highest for the

smaller model. In some cases, the larger model may be better for a particular item type or parameter setting, but the total MAP is always highest for the smaller model.

The MAP scores also show that `crowdsourced` models usually outperform `familiar` models. On the surface, this may seem counterintuitive, as `familiar` participants are expected to complete the PDT most faithfully. However, as seen in this chapter and Section 5.4, `crowdsourced` responses are more like NNS responses than are `familiar` responses. `crowdsourced` participants are less motivated than `familiar` participants, and this manifests in a higher rate of lazy or bad faith responses; this noise may simply better model NNS responses. Moreover, the `familiar` participants were all hand-picked native English speakers, whereas the `crowdsourced` participants are anonymous, with no way of confirming that they are in fact native English speakers. It is possible that some "NS" responses come from non-native speakers, which could explain why `crowdsourced` models achieve higher MAP; in theory, the best model would be one of responses that are well-formed enough to have perfect annotations but nonetheless capture the range of NNS responses, which sometimes include minor errors and non-nativelike forms.

Another salient pattern seen in all 10 tables is that MAP scores are highest for `intransitives` and lowest for `ditransitives`. In other words, as sentence complexity increases, there is a monotonic decrease in system performance. Because `transitives` and `ditransitives` cannot simply be avoided or transformed to `intransitives`, this trend suggests that an ICALL tool or similar application should carefully consider how to best optimize for more complex items, e.g., by selecting the best `term representation`, as this correlates with item complexity.

6.5.1 CORE EVENT experiments

CORE EVENT, as discussed in Section 5.1, assesses whether a response captures the main action of the PDT item and requires that the event is linked to a subject (and an object or

	Crowd NS model = 14					Crowd NS model = 50				
	ldh	xdh	xdx	WAR	SBERT	ldh	xdh	xdx	WAR	SBERT
Intr	***0.859***	0.856	0.854	0.865	0.835	**0.855**	0.854	0.852	0.865	0.831
Tran	***0.737***	0.735	0.728	0.742	0.703	**0.736**	0.733	0.725	0.742	0.701
Ditr	***0.665***	0.661	0.664	0.660	0.634	0.657	0.656	**0.661**	0.660	0.629
Targ	***0.739***	0.738	0.732	0.735	0.708	**0.737**	0.735	0.729	0.735	0.704
Untg	***0.768***	0.763	0.765	0.777	0.740	0.762	0.759	**0.763**	0.777	0.736
Prim	***0.754***	0.752	0.747	0.756	0.723	**0.750**	0.748	0.745	0.756	0.719
Mix	***0.753***	0.749	0.750	0.756	0.725	**0.749**	0.746	0.746	0.756	0.721
Total	***0.753***	0.751	0.748	0.756	0.724	**0.750**	0.747	0.746	0.756	0.720

Table 6.5: Mean Average Precision (MAP) scores for the CORE EVENT annotation feature, derived from various response rankings: weighted annotation ranking (WAR), the three system `term representation` rankings (labeled dependencies (`ldh`), unlabeled dependencies (`xdh`), and dependents only (`xdx`)), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, `intransitive` items, `primary` NS models), and for the full set (Total).

objects where necessary).

Table 6.5 presents CORE EVENT MAP scores for the 14-response and 50-response `crowdsourced` models. These scores show that for assessing CORE EVENT, the smaller model, used with labeled dependencies is superior across the board. The 50-response model comes closest for `transitives`, but never outperforms the 14-response model. We can also observe here that the `term representation` setting is more relevant in the larger model, with `xdx` besting `ldh` in the case of `ditransitives` and `untargeted` items. Responses to `ditransitives` and `untargeted` items tend to be the least homogenous; they have higher standardized type-to-token ratios than their counterparts (see Table 6.3). Moving from `ldh` to `xdx` representations results in a lower standardized type-to-token ratio (Table 6.3), so it is unsurprising that doing so improves performance for these items.

Table 6.6 compares CORE EVENT MAP scores for `familiar` and `crowdsourced` models. As seen in the total MAP scores, the `crowdsourced` models outperform the `familiar` models overall, but the difference is slight. The table also shows that for CORE

	Familiar NS model = 14					Crowd NS model = 14				
	ldh	xdh	xdx	WAR	SBERT	ldh	xdh	xdx	WAR	SBERT
Intr	0.859	0.859	***0.865***	0.865	0.838	**0.857**	0.852	0.848	0.865	0.833
Tran	***0.740***	0.737	0.726	0.742	0.703	**0.738**	0.735	0.728	0.742	0.702
Ditr	0.651	0.648	**0.660**	0.660	0.625	0.663	0.659	***0.673***	0.660	0.641
Targ	**0.733**	0.732	0.732	0.735	0.707	***0.739***	0.736	0.733	0.735	0.709
Untg	0.767	0.764	***0.769***	0.777	0.737	**0.767**	0.761	**0.767**	0.777	0.742
Total	0.750	0.748	**0.751**	0.756	0.722	***0.753***	0.749	0.750	0.756	0.725

Table 6.6: Mean Average Precision (MAP) scores for the CORE EVENT annotation feature, comparing familiar and crowdsourced responses. MAP is derived from various response rankings: the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), weighted annotation ranking (WAR), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, targeted items), and for the full set (Total). Note that all models represented here are mixed due to the small number of familiar participants.

EVENT, the familiar models generally perform best with the dependent-only (xdx) term representation, whereas the crowdsourced models generally perform best with labeled dependencies. As seen in Table 6.5, the trend is unchanged for the crowdsourced 50-response models, but more data is needed to see how larger familiar models behave.

6.5.2 ANSWERHOOD experiments

ANSWERHOOD, as discussed in Section 5.1, assesses whether a response presents a direct answer to the question asked in the PDT prompt.

Table 6.7 presents ANSWERHOOD MAP scores for the 14-response and 50-response crowdsourced models. The dependent-only (xdx) models outperform the others across the board here. We can also observe that unlabeled dependencies outperform labeled dependencies in every case. This would indicate that identifying a direct answer based on my similarity metrics is a relatively simple task that works best with the simplest representations. The scores also show that model size has little bearing here; overall, the smaller

	Crowd NS model = 14					Crowd NS model = 50				
	ldh	xdh	xdx	WAR	SBERT	ldh	xdh	xdx	WAR	SBERT
Intr	0.868	0.871	***0.878***	0.881	0.869	0.866	0.868	**0.874**	0.881	0.868
Tran	0.816	0.819	**0.846**	0.845	0.838	0.818	0.823	***0.851***	0.845	0.838
Ditr	0.824	0.826	***0.841***	0.837	0.833	0.821	0.822	**0.840**	0.837	0.833
Targ	0.787	0.788	**0.810**	0.817	0.799	0.787	0.789	***0.811***	0.817	0.798
Untg	0.885	0.890	***0.900***	0.892	0.894	0.883	0.886	**0.899**	0.892	0.895
Prim	0.837	0.840	***0.854***	0.854	0.845	0.837	0.840	***0.854***	0.854	0.846
Mix	0.835	0.838	***0.857***	0.854	0.848	0.833	0.835	**0.856**	0.854	0.847
Total	0.836	0.839	***0.855***	0.854	0.847	0.835	0.838	**0.855**	0.854	0.846

Table 6.7: Mean Average Precision (MAP) scores for the ANSWERHOOD annotation feature, derived from various response rankings: weighted annotation ranking (WAR), the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, primary NS models), and for the full set (Total).

model's MAP is just 0.0005 higher.

	Familiar NS model = 14					Crowd NS model = 14				
	ldh	xdh	xdx	WAR	SBERT	ldh	xdh	xdx	WAR	SBERT
Intr	0.868	0.871	***0.882***	0.881	0.868	0.869	0.873	**0.878**	0.881	0.870
Tran	0.824	0.826	***0.852***	0.845	0.840	0.817	0.818	**0.847**	0.845	0.840
Ditr	0.820	0.822	***0.846***	0.837	0.832	0.820	0.822	**0.845**	0.837	0.835
Targ	0.786	0.787	***0.815***	0.817	0.798	0.785	0.787	**0.813**	0.817	0.802
Untg	0.889	0.892	***0.904***	0.892	0.896	0.885	0.889	**0.900**	0.892	0.894
Total	0.837	0.840	***0.860***	0.854	0.847	0.835	0.838	**0.857**	0.854	0.848

Table 6.8: Mean Average Precision (MAP) scores for the ANSWERHOOD annotation feature, comparing familiar and crowdsourced responses. MAP is derived from various response rankings: the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), weighted annotation ranking (WAR), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, targeted items), and for the full set (Total). Note that all models represented here are mixed due to the small number of familiar participants.

Table 6.8, which compares familiar and crowdsourced models, confirms the higher performance of dependent-only models. We also see a slight but consistent advan-

tage for the familiar models on this task. The scores here also confirm the relative ease of assessing this feature. The top scores seen in the two the ANSWERHOOD MAP tables are only bested by MAP scores for one other feature—INTERPRETABILITY. The fact that the highest ANSWERHOOD MAP scores come from the smaller 14-response models (both familiar and crowdsourced) again support this idea. The task is simple enough for small models, and larger models risk adding noise.

6.5.3 GRAMMATICALITY experiments

The GRAMMATICALITY feature, as discussed in Section 5.1, indicates whether a response is free from any grammatical errors.

	Crowd NS model = 14					Crowd NS model = 50				
	ldh	xdh	xdx	WAR	SBERT	ldh	xdh	xdx	WAR	SBERT
Intr	0.868	0.870	***0.872***	0.887	0.866	0.863	0.864	**0.866**	0.887	0.864
Tran	0.753	0.756	**0.757**	0.781	0.757	0.758	0.760	***0.761***	0.781	0.757
Ditr	0.682	0.685	***0.700***	0.695	0.694	0.679	0.685	**0.697**	0.695	0.693
Targ	0.777	0.778	***0.784***	0.800	0.782	0.776	0.776	**0.783**	0.800	0.781
Untg	0.758	0.763	***0.769***	0.776	0.762	0.757	0.762	**0.766**	0.776	0.761
Prim	0.769	0.773	***0.776***	0.788	0.770	0.768	0.770	**0.774**	0.788	0.770
Mix	0.766	0.768	***0.776***	0.788	0.774	0.765	0.768	**0.775**	0.788	0.772
Total	0.768	0.770	***0.776***	0.788	0.772	0.767	0.769	**0.775**	0.788	0.771

Table 6.9: Mean Average Precision (MAP) scores for the GRAMMATICALITY annotation feature, derived from various response rankings: weighted annotation ranking (WAR), the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, primary NS models), and for the full set (Total).

Table 6.9 presents GRAMMATICALITY MAP scores for the 14-response and 50-response crowdsourced models. These scores show a slight but consistent preference for dependent-only (xdx) models, with labeled dependencies consistently producing the lowest MAP and unlabeled dependencies falling in the middle. The table also shows that the smaller, 14-response models outperform the larger, 50-response models, but these differences are very

small. There is one exception to this pattern, with the larger models performing slightly better for transitives. Keeping in mind that in all cases, the heads and dependents within dependencies are lemmatized, these observations suggest that like ANSWERHOOD, GRAMMATICALITY is relatively simple to asses, requiring only a small "bag of lemmatized dependents" model.

	Familiar NS model = 14					Crowd NS model = 14				
	ldh	xdh	xdx	WAR	SBERT	ldh	xdh	xdx	WAR	SBERT
Intr	0.863	0.864	**0.873**	0.887	0.863	0.868	0.869	***0.874***	0.887	0.869
Tran	0.760	0.759	***0.762***	0.781	0.760	0.752	0.754	**0.757**	0.781	0.758
Ditr	0.678	0.685	**0.698**	0.695	0.698	0.678	0.680	***0.699***	0.695	0.696
Targ	0.776	0.776	***0.787***	0.800	0.783	0.776	0.777	**0.786**	0.800	0.786
Untg	0.757	0.762	***0.768***	0.776	0.764	0.756	0.759	**0.767**	0.776	0.763
Total	0.767	0.769	***0.778***	0.788	0.773	0.766	0.768	**0.776**	0.788	0.774

Table 6.10: Mean Average Precision (MAP) scores for the GRAMMATICALITY annotation feature, comparing familiar and crowdsourced responses. MAP is derived from various response rankings: the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), weighted annotation ranking (WAR), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, targeted items), and for the full set (Total). Note that all models represented here are mixed due to the small number of familiar participants.

Table 6.10 presents the GRAMMATICALITY MAP scores for familiar and crowdsourced models. As before, the models here show a consistent preference for the dependent-only (xdx) term representation. The total MAP scores show that familiar models perform best overall. transitives stand out again here, where the best familiar model score is approximately 0.005 higher than the best crowdsourced score, which is the largest such difference in the table;for each intransitives and ditransitives, the crowdsourced models are higher by approximately 0.001 points. This suggests that in describing transitive events, familiar NSs behave more like NNSs than do crowdsourced NSs, at least with regard to grammar and grammatical errors. Participant motivation is the mostly likely factor here.

6.5.4 INTERPRETABILITY experiments

INTERPRETABILITY, as discussed in Section 5.1, assesses whether a response evokes a clear mental image, although it does not need to resemble the actual image in the PDT item. This feature also requires a response's verb to have all necessary arguments specified.

	Crowd NS model = 14					Crowd NS model = 50				
	ldh	xdh	xdx	WAR	SBERT	ldh	xdh	xdx	WAR	SBERT
Intr	0.932	0.931	***0.933***	0.930	0.922	0.928	0.927	**0.933**	0.930	0.923
Tran	***0.823***	0.821	0.811	0.803	0.806	**0.821**	0.816	0.812	0.803	0.804
Ditr	0.789	0.784	***0.794***	0.721	0.777	0.786	0.782	**0.792**	0.721	0.772
Targ	0.835	0.832	***0.836***	0.804	0.828	0.833	0.829	**0.834**	0.804	0.826
Untg	***0.862***	0.858	0.856	0.833	0.842	**0.857**	0.855	**0.857**	0.833	0.840
Prim	***0.847***	0.845	0.846	0.818	0.837	0.845	0.842	**0.846**	0.818	0.833
Mix	***0.849***	0.846	0.846	0.818	0.833	0.844	0.841	**0.845**	0.818	0.833
Total	***0.848***	0.845	0.846	0.818	0.835	**0.845**	0.842	**0.845**	0.818	0.833

Table 6.11: Mean Average Precision (MAP) scores for the INTERPRETABILITY annotation feature, derived from various response rankings: weighted annotation ranking (WAR), the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, primary NS models), and for the full set (Total).

Table 6.11 presents INTERPRETABILITY MAP scores for the 14-response and 50-response crowdsourced models. The most prominent trend here is that the smaller, 14-response models outperform the 50-response models in all cases. (The top intransitive scores are both shown as 0.933, but these are truncated for space and the smaller model score is higher by 0.0003.) These differences across model sizes are small, at roughly 0.002 points or less each; the untargeted setting is the only exception, with a difference of approximately 0.005 points. This suggests that with regard to INTERPRETABILITY, the untargeted PDT setting elicits a greater degree of noise than does the targeted setting, and this plays a larger role in the larger models. In other words, the larger the sample size of the crowdsourced NS model, the more likely it will include responses which do not align well with those of NNSs. Thus, when attempting to infer the INTERPRETABILITY

of an NNS response based on models containing outlier NS responses, performance suffers. This is almost certainly tied to motivation, and it tracks with my own observations from the data. An `untargeted` prompt gives the participant greater freedom—or perhaps "creative license"—to describe the image. NNSs and `familiar` NSs stick to the spirit of the task, whereas off-target responses are easy to find among the `crowdsourced` NS data. For example, for the item showing a woman teaching a math class to a student, `crowdsourced` NSs in the `targeted` setting overwhelmingly refer to the action of teaching. Responses in the `untargeted` setting include several that do not address the action (and incidentally, fail at ANSWERHOOD as well), like "school" and "Time for math"; multiple others simply transcribe the math problems shown on the chalkboard or comment inappropriately on the teacher's appearance. None of these responses are helpful in capturing the more constrained behavior of the NNS participants.

	`Familiar` NS model = 14					`Crowd` NS model = 14				
	`ldh`	`xdh`	`xdx`	WAR	SBERT	`ldh`	`xdh`	`xdx`	WAR	SBERT
`Intr`	0.930	0.930	***0.934***	0.930	0.923	**0.933**	0.931	0.932	0.930	0.922
`Tran`	**0.822**	0.819	0.811	0.803	0.805	***0.826***	0.824	0.811	0.803	0.805
`Ditr`	0.787	0.786	***0.796***	0.721	0.782	0.788	0.783	**0.795**	0.721	0.772
`Targ`	0.835	0.833	***0.836***	0.804	0.830	**0.835**	0.832	**0.835**	0.804	0.825
`Untg`	**0.858**	0.857	**0.858**	0.833	0.843	***0.863***	0.859	0.857	0.833	0.841
Total	**0.847**	0.845	**0.847**	0.818	0.837	***0.849***	0.846	0.846	0.818	0.833

Table 6.12: Mean Average Precision (MAP) scores for the INTERPRETABILITY annotation feature, comparing `familiar` and `crowdsourced` responses. MAP is derived from various response rankings: the three system `term representation` rankings (labeled dependencies (`ldh`), unlabeled dependencies (`xdh`), and dependents only (`xdx`)), weighted annotation ranking (WAR), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, `intransitive` items, `targeted` items), and for the full set (Total). Note that all models represented here are `mixed` due to the small number of `familiar` participants.

Another trend seen here is that although the effect of `term representation` is small, the highest MAP scores for `transitives` come from labeled dependencies, whereas dependent-only representations work best for `intransitives` and `ditransitives`.

The dependent-only representation collapses labeled dependencies into a smaller number of terms, reducing the distance between NS models and NNS responses (as seen in Table 6.4). Given this fact, the trend here suggests that the NNS participants and `crowdsourced` NS participants exhibit more convergent behavior in response to `transitive` items (as opposed to `intransitives` and `ditransitives`), and thus a relatively granular representation (labeled dependencies) works well in similarity measures, leading to better discriminatory power for the INTERPRETABILITY feature.

Turning to comparisons of `familiar` and `crowdsourced` NS models shown in Table 6.12, this pattern is repeated, with labeled dependencies performing best for `transitives` and dependents-only performing best for `intransitives` and `ditransitives`. Overall, the `crowdsourced` models perform better than the `familiar` models, but this difference is very slight.

6.5.5 VERIFIABILITY experiments

VERIFIABILITY, as discussed in Section 5.1, requires that all information presented in a response must be clearly verifiable from the PDT image.

As seen in Table 6.13, for discriminating the VERIFIABILITY feature, the smaller 14-response models consistently outperform the 50-response models by a slim margin. This effect is greatest for `untargeted` and `mixed` settings. As compared to their counterparts (`targeted` and `primary` settings, respectively), these settings result in greater response distances from the NS models (see Table 6.4). These are relatively unconstrained settings, both eliciting a wider variety of responses and potential noise, and such outlier responses have a greater chance of appearing in larger models. This is supported by the fact that MAP differences between `targeted` and `untargeted` settings and `primary` and `mixed` settings appear greater for the 50-response models than the 14-response models. The results seen here show that the VERIFIABILITY feature is sensitive to this pattern across model sizes.

	Crowd NS model = 14					Crowd NS model = 50				
	ldh	xdh	xdx	WAR	SBERT	ldh	xdh	xdx	WAR	SBERT
Intr	0.852	0.852	***0.853***	0.866	0.840	0.849	0.849	**0.851**	0.866	0.836
Tran	***0.809***	0.808	0.803	0.798	0.787	**0.807**	0.806	0.803	0.798	0.785
Ditr	0.814	0.812	***0.815***	0.780	0.798	0.811	0.809	**0.812**	0.780	0.796
Targ	***0.825***	0.824	0.825	0.815	0.812	**0.825**	0.824	0.823	0.815	0.810
Untg	***0.825***	0.824	0.822	0.815	0.805	**0.820**	0.819	0.820	0.815	0.802
Prim	***0.826***	0.824	0.823	0.815	0.808	**0.824**	0.823	0.822	0.815	0.806
Mix	***0.825***	0.824	0.824	0.815	0.808	**0.821**	0.821	0.821	0.815	0.805
Total	***0.825***	0.824	0.824	0.815	0.808	**0.823**	0.822	0.822	0.815	0.806

Table 6.13: Mean Average Precision (MAP) scores for the VERIFIABILITY annotation feature, derived from various response rankings: weighted annotation ranking (WAR), the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, primary NS models), and for the full set (Total).

These results also show that term representation has a very small effect on discriminating for VERIFIABILITY. The effect is greatest, however, when comparing intransitive, transitive and ditransitive items. While intransitives and ditransitives show a slight preference for dependents-only term representation, transitive items work best with labeled dependencies (ldh), regardless of model size. Again, this indicates that NS and NNS behavior is most convergent in response to transitive items, allowing for a relatively granular representation. transitives also stand out here because they result in the lowest MAP scores for VERIFIABILITY. The reasons for this are not obvious, but it may be that the relatively clear and concrete nature of transitive PDT items means annotators behave more strictly when marking VERIFIABILITY for transitives than for intransitives or ditransitives.

When comparing familiar and crowdsourced models for discriminating for VERIFIABILITY, as shown in Table 6.14, a slight preference for crowdsourced models is evident. Transitives are again an exception here, where familiar models slightly outperform crowdsourced models on this task.

	Familiar NS model = 14					Crowd NS model = 14				
	ldh	xdh	xdx	WAR	SBERT	ldh	xdh	xdx	WAR	SBERT
Intr	0.847	0.847	**0.852**	0.866	0.836	0.852	0.852	***0.854***	0.866	0.843
Tran	***0.808***	0.807	0.803	0.798	0.787	**0.807**	0.807	0.802	0.798	0.786
Ditr	0.811	0.811	**0.812**	0.780	0.802	0.815	0.812	***0.817***	0.780	0.796
Targ	0.821	0.821	**0.822**	0.815	0.814	0.824	0.824	***0.826***	0.815	0.811
Untg	**0.824**	0.822	0.823	0.815	0.803	***0.825***	0.824	0.823	0.815	0.806
Total	0.822	0.822	**0.823**	0.815	0.808	***0.825***	0.824	0.824	0.815	0.808

Table 6.14: Mean Average Precision (MAP) scores for the VERIFIABILITY annotation feature, comparing familiar and crowdsourced responses. MAP is derived from various response rankings: the three system term representation rankings (labeled dependencies (ldh), unlabeled dependencies (xdh), and dependents only (xdx)), weighted annotation ranking (WAR), and SBERT rankings. MAP scores are shown for each item type or parameter setting (e.g, intransitive items, targeted items), and for the full set (Total). Note that all models represented here are mixed due to the small number of familiar participants.

The crowdsourced 14-response models covered in Table 6.14 contain *only* mixed response models and are thus not identical to the the crowdsourced 14-response models covered in Table 6.13, which also include primary models. To clarify, the crowdsourced 14-response *total* row in Table 6.14 is equivalent to the mixed row in Table 6.13. In terms of performance, the crowdsourced results in Table 6.14 differ from those in Table 6.13 in that they show dependents-only representations working best for targeted settings, but labeled dependencies (ldh) working best for untargeted settings. The familiar models also exhibit this pattern. One explanation here is related to the fact that untargeted settings elicit more undesirable responses; it is not uncommon for such responses to mention the entities and actions in the image but somehow mischaracterize the scenario. Such responses will appear more like the NS model given a dependents-only setting than they do with a labeled dependencies setting. In other words, VERIFIABILITY is challenging to capture with a bag-of-words style similarity approach because it means verifying not only the words but also their relationships.

6.6 Holistic experiments

In this section I turn from the use of similarity scoring to discriminate annotations for individual features to the use of similarity scoring to approximate an ideal holistic ranking of NNS responses. These experiments rely on the weighted annotation ranking (WAR), which is based on the weighted annotation score (WAS) for each NNS response, as described in Section 5.4. While in many use cases it may be preferable to focus on specific features, the experiments here provide insight into the feasibility of using annotation-free, surface-level similarity measures to approximate a determination of response "goodness" based on human judgment.

This section is broken into subsections for each parameter, where I compare the performance of different settings for that parameter. Ideally, this would allow me to identify which parameter settings work best and in which contexts (i.e., with which item types and in combination with which other parameter settings). Such trends could then suggest how to configure my system for new items. Clear, predictive trends are not always evident here, however, and the observations in this section are presented in the hopes that they can guide future investigations, rather than for immediate application.

As before, I compare models of two different sizes in order to see the effects of sample size on performance. The smaller of the two models contains 14 NS responses per item, and the larger model contains 50 NS responses per item. My system uses these models as the basis of its tf-idf cosine similarity measure that is used to score each response and in turn rank the full set of 70 NNS responses. As a benchmark, I also use SBERT with the same NS models to produce similarity scores and rank the NNS responses.

The metric used throughout this section is Spearman's rank correlation coefficient, as implemented in the SciPy Python package (Virtanen et al., 2020) and described in Zwillinger and Kokoska (1999). I use this metric to assess how well each system (or SBERT) ranking of the NNS responses correlates with the weighted annotation ranking (WAR). I

chose Spearman over similar measures, such as Pearson correlation, because Spearman has been found to be better suited for semantic textual similarity (STS) tasks (Reimers et al., 2016).

A caveat is in order here. A Spearman score is always accompanied by a p-value, which is a measure of statistical significance. The p-value gives some indication of the probability of achieving the correlation score under the null hypothesis; in other words, how likely is the observed correlation (or a stronger one) if there is no real relationship between the two rankings? A long-standing tenet holds that p-values should be less than 0.05 to indicate statistical significance (Zar, 1972). In recent years, the use of p-values and claims around them has been the subject of much debate in numerous fields, and the reliance on p-values in linguistic research has slipped from an unassailable orthodoxy to a point of contention (Moran et al., 2012; Tomczak and Tomczak, 2014). For the results presented here, roughly half of p-values fall below 0.05, but many exceed it by varying degrees. This is largely a function of the relatively small size of the test sets (70 NNS responses); the SciPy authors note that p-values "are not entirely reliable but are probably reasonable for datasets larger than 500 or so" (Virtanen et al., 2020). It is important to note that p-values above 0.05 do not indicate that the null hypothesis is true, but rather that the evidence is not statistically significant enough to claim a correlation exists (Vasishth and Nicenboim, 2016).

It is also important to note that higher p-values naturally occur in cases where the annotations underlying the rankings are heavily skewed. Many of the PDT items in this study exhibit a ceiling effect, where a majority of responses receive perfect annotations. For the inter-annotator response set used in developing the annotation scheme, the overall rate of positive annotations ranges from 0.717 to 0.872 for all five features (see Table 5.2). For the full corpus, the overall rate of perfectly annotated (i.e, 5/5 positive feature annotations) responses was 0.614 for NNSs and 0.528 for NSs (see Table 5.9). This results in an unevenly distributed weighted annotation ranking, with the perfect responses tied in rank. Tied ranks are also common among less-than-perfect responses. For example, it is common for NNS

responses to receive positive annotations for all features except GRAMMATICALITY, and thus these responses share a weighted annotation rank, which leads to skewed rankings and high p-values. Applying the annotation feature weights helps differentiate response scores, but with only five features, such ties are common. Moreover, skewedness is common in the system-produced rankings as well, because it is common for NNSs to provide identical responses, especially in `targeted` PDT settings.

With regard to the analysis in this section, one could opt to handle concerns about p-values by omitting any results where the Spearman correlation p-value is above the 0.05 threshold. This leaves a patchwork of results that are unevenly distributed with regard to the various item types, parameter settings, and model sizes that I am interested in comparing, effectively complicating and limiting the analysis. Another option would be to point out and attempt to explain wherever results involve p-values above 0.05, but this is not practically feasible. I opt instead to avoid presenting or interpreting p-values and the significance of Spearman correlations altogether and instead present my findings here with the caveat that they may not always be statistically significant. They should not be relied upon for any immediate decision making. These findings are likely to indicate useful trends, but analysis with much larger datasets should be explored before making any determinations about the statistical significance of these results or the reliability and predictive power of trends seen here.

To help mitigate this fact, I report Spearman correlations not as, for example, a single mean for each setting within a parameter, but as a set of descriptive statistics including mean, median, minimum, maximum and standard deviation. These figures should give a better depiction of the shape of the results than a single mean score and thus better indicate potential trends and guide future research.

6.6.1 `Term normalization` experiments

In my pipeline, the NS model for each PDT item is comprised of some number of NS responses. The length of these responses can vary; some valid responses contain only one or two words[1], while the longest perfectly annotated responses top out at around 15 words and some less-than-perfect responses exceed 30 words. Understanding the impact of response lengths on a model is an important step in optimizing my response rating process. So far, I have treated each NS model as a flat "bag of terms" in which each term (roughly, *dependency*; see Sections 3.3.2 and 6.6.6) contributes equally, meaning longer responses carry more weight than shorter responses in the model. This has the potential to introduce noise.

Response A	Response B	`Norm` wt	`Non-norm` wt
The girl is singing	The girl in the cute 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, ERASED)	—	—
	det(the, dress)	0.050	0.071
	amod(cute, 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, ROOT)	root(sing, ROOT)	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 6.15: A "toy" model consisting of lemmatized syntactic dependencies from only two NS responses, each with perfect annotation scores. (Note that the version of Stanford typed dependencies used in this work collapses dependencies containing prepositions and incorporates prepositions in a label, resulting in the "prep_in" and "erased" dependencies above. See Section 3.2.2 for more on the parsing and lemmatization.)

My hypothesis is that system performance should improve by using NS models where

[1]Participants were instructed to provide complete sentences, but incomplete sentences were still judged valid where appropriate; see Chapter 4 and the Annotation Guide in Appendix B.

each term token is re-weighted by $1/n$ before it is added to the model, where n is the number of term tokens in the response containing the term token. In other words, I believe dependencies should be re-weighted to ensure that every NS *response*—not *term*—contributes equal weight to the model. The rationale here is simple. Every response used in the NS model is assumed to contain the information that is crucial for satisfying the PDT prompt, and the number of terms conveying this information is roughly equivalent from one response to another. Thus, as the number of terms in a response increases (above some minimum number), the likelihood that any given term in that response is crucial decreases.

This is illustrated by the responses in Table 6.15. If we take these two responses to constitute one NS model, Response A contributes four dependencies, each of which is necessary to satisfy the five annotated features and contributes meaningfully to the model. Response B, however, contributes 10 dependencies, some of which, like *amod(purple, dress)*, add non-critical detail. In a `non-normalized` setting, this dependency constitutes one out of a total 14 dependencies in the NS model, approximately 0.071. In a `normalized` setting, however, this dependency appears as zero of four (0.0) dependencies in Response A, and one of 10 (0.1) in Response B, making it 0.05 of the overall model (0.1 divided by two responses). This should have the effect of making extraneous information in the model less impactful on response ratings.

Taking this example further, consider the dependency *nsubj(girl, sing)*, also from Table 6.15. In the `non-normalized` setting, this dependency appears as two out of a total 14 dependencies, or 0.143 of the NS model. 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 NS model (0.35 divided by two responses). Because this dependency is critical, raising its weight from 0.143 to 0.175 should have a positive impact on system performance; NNS responses containing the dependency should rise in the rankings relative to those without it.

To test my hypothesis, I compared the performance of `normalized` and `non-norm-`

	NS model sample size = 14			NS model sample size = 50		
	Non-n	Norm	SBERT	Non-n	Norm	SBERT
count	360	360	120	360	360	120
mean	**0.340**	0.335	***0.487***	**0.349**	0.347	***0.509***
median	**0.332**	0.313	***0.507***	**0.348**	0.333	***0.523***
min	**-0.181**	-0.219	***-0.138***	**-0.185**	-0.230	***-0.090***
max	***0.900***	0.891	0.881	0.898	***0.899***	0.881
std dev	0.226	0.227	0.196	0.225	0.230	0.177

Table 6.16: Comparing Spearman rank correlation scores where all dependencies (terms) in Non-n(ormalized) NS models carry equal weight, and all dependencies in Norm(alized) NS models have their scores normalized proportionally to the length of the parent response. Results are shown using NS models of 14 responses and 50 responses. Each Norm(alized) and Non-n(ormalized) column represents 360 different rankings (12 system configurations $\times$ 30 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking. Each *SBERT* column represents 120 rankings (4 system configurations $\times$ 30 items; SBERT operates on plain text, so the term representation parameter does not apply).

alized models. The results of this experiment are shown in Table 6.16. This re-weighting does not apply to SBERT, but SBERT scores are included as a benchmark. For both sample sizes, the comparisons show slightly higher mean and median Spearman correlations for the non-normalized NS models, disproving my hypothesis. This means that the desired NNS response ranking—the weighted annotation ranking—is best approximated via similarity comparisons with the non-normalized NS responses. In other words, the level of noise in the NS responses is already optimal, and with its impact reduced via term normalization, the NS models are less like the NNS responses. Because the non-normalized NS models outperform their normalized counterparts, and for the sake of simplicity, this parameter was not included in the other experiments discussed in this chapter; only non-normalized configurations were used elsewhere.

6.6.2 Transitivity experiments

Here I examine the performance of my ranking system when applied to items that are canonically either intransitive, transitive, or ditransitive. Unlike the other variables through-

out this chapter, `transitivity` is not a parameter setting. Individual PDT items are assumed to fit predominately only one of the three types here. In other words, I cannot choose to process a given PDT item with any of the three `transitivity` settings. Rather, the experiments in this section examine my system's performance across three sets of 10 items each, representing `intransitive`, `transitive` and `ditransitive` events.

	NS model sample size = 14					
	`intransitives`		`transitives`		`ditransitives`	
	System	SBERT	System	SBERT	System	SBERT
count	120	40	120	40	120	40
mean	0.439	**0.497**	0.314	***0.563***	0.267	**0.400**
median	0.416	**0.479**	0.304	***0.555***	0.276	**0.444**
min	-0.119	***0.199***	-0.110	***0.199***	-0.181	**-0.138**
max	***0.900***	0.881	**0.777**	0.772	**0.710**	0.697
std dev	0.228	0.189	0.218	0.134	0.198	0.222

	NS model sample size = 50					
	`intransitives`		`transitives`		`ditransitives`	
	System	SBERT	System	SBERT	System	SBERT
count	120	40	120	40	120	40
mean	0.423	**0.516**	0.345	***0.566***	0.278	**0.446**
median	0.426	**0.517**	0.331	***0.561***	0.286	**0.471**
min	-0.076	**0.200**	-0.204	***0.222***	-0.185	**-0.090**
max	***0.898***	0.881	**0.778**	0.771	0.708	**0.709**
std dev	0.249	0.172	0.207	0.135	0.195	0.200

Table 6.17: Comparing Spearman rank correlation scores for `intransitive`, `transitive` and `ditransitive` PDT items, using NS models of either 14 or 50 random responses per item. Each *System* column represents 120 different rankings (12 system configurations $\times$ 10 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking. Each *SBERT* column represents 40 rankings (4 system configurations $\times$ 10 items; SBERT operates on plain text, so the `term representation` parameter does not apply).

Sets of descriptive statistics for the Spearman rank correlation scores produced by my system and SBERT using these models are presented in Table 6.17. There are 10 items per type. Each item has a `targeted` and `untargeted` version, which are separate datasets. For each dataset, I sample both `primary` and `mixed` models. My system con-

verts the response text to three different `term representations` (`ldh`, `xdh`, `xdx`). This results in 12 system configurations: 2 `targeting` settings × 2 `primacy` settings × 3 `term representations`. Using all 12 configurations results in 120 Spearman rank correlations per `transitivity` type. The system statistics presented in Table 6.17 cover these 120 scores. Because SBERT operates on plain text and cannot make use of the `term representation` variable, it involves only four configurations, resulting in 40 Spearman scores per `transitivity` type.

Table 6.17 shows some notable patterns. In all cases, SBERT outperforms my system on the most important metrics here: mean and median Spearman scores. This is expected, as SBERT is a powerful, highly developed, state-of-the-art approach to language modeling and sentence similarity and is trained on much more data. My similarity measure is far less sophisticated and involves tools trained on much smaller datasets, but also has the advantage of preserving a high degree of transparency and explainability. Although SBERT underperformed my system in predicting individual feature annotations in Section 6.5, the results here would suggest that SBERT is relatively good at scoring and ranking responses holistically in a way that correlates with the WAR (and thus with human annotation).

Notably, my system does reach a higher maximum Spearman score than SBERT in all but one case. However, leveraging this fact for unseen and unannotated items would require consistently predicting cases in which system scores outperform SBERT scores, and I have not uncovered any pattern that would enable this.

For both size models, my system performs best on `intransitives`, followed by `transitives` and then `ditransitives`. For `intransitives`, my system achieves its highest mean and median Spearman scores using the 14-response models. For `transitives`, the median is highest with the 14-response models, but the mean is highest with the 50-response models. For `ditransitives`, the mean and median are highest with the 50-response model. If we take standardized type-to-token ratio (STTR) as a metric of complexity (see Table 6.3), the observations here show a monotonic pattern in which

more complex items are best handled with larger models. Future research could explore this curve by sampling a wider range of model sizes, including models larger than 50 responses. Assuming the pattern holds, for each new PDT item, one could use the item's STTR to help determine the optimal model size for ranking NNS responses.

Like my system, SBERT achieves its highest mean and median Spearman scores with the smaller model. For both size models, SBERT performs best on `transitives`, however, followed by `intransitives` and then `ditransitives`. The relatively strong performance on `transitives` appears to be an inherent feature of BERT architectures (at least for English), as Thrush et al. (2020) showed by testing BERT's predictions with novel verbs. They found that when BERT is shown a single instance of a novel verb without an object during training, BERT expects the verb to also occur with an object. The reverse was not true, however; novel verbs seen once in a transitive context are not expected to occur intransitively. This shows what Thrush et al. (2020) call BERT's "transitivity bias," which likely explains why I found SBERT performance to be weaker for `intransitives`, despite the fact those responses are generally shorter than for `transitives`.

For `ditransitives`, SBERT's lower performance may be influenced by the increased complexity, but research suggests this is more directly related to sentence length, and `ditransitive` items, naturally, are the longest in my dataset, as they require more verb arguments (see Table 6.2). Warstadt and Bowman (2019) used classifiers trained on a corpus of sentences with corresponding BERT output and human grammaticality judgments and found that the classifiers' ability to predict human judgments from the BERT sentence embeddings correlates strongly with sentence length, with a sharp drop appearing for sentences of 11 words or more.

6.6.3 `Targeting` experiments

Comparing Spearman rank correlations across the `targeting` parameter reveals that `targeted` settings consistently outperform `untargeted` settings, as shown in Table 6.18.

That is, targeted settings allow my similarity-based ranking system to rank NNS responses in a way that better captures responses' overall quality as represented by the weighted annotation rankings. Responses in targeted settings tend to be shorter and have lower standardized type-to-token ratios, as shown in Tables 6.2 and 6.3; as seen with transitivity, these measures correlate with higher Spearman scores for targeted settings.

These results also show that SBERT generally outperforms my system at approximating the desired holistic response rankings. There are exceptions—as seen in the table, my system achieves higher maximum Spearman scores with both size models. Upon examining these cases, I found no discernible pattern that could predict when the system score will exceed the SBERT score, but it is possible a pattern would emerge with more data.

Comparing the results for the 14-response NS models and 50-response NS models here reveals that performance is consistently better with the 50-response models, both for my system and for SBERT. For my system scores, if we compare the differences for targeted settings versus the differences for untargeted settings across these model sizes, Table 6.18 shows that targeted settings benefit more from the increased model size than do untargeted settings. In fact, these mean and median differences for targeted settings are more than double the mean and median differences for untargeted settings. With more NS responses, one could experiment with larger NS models to find the sizes that lead to optimal system performance for targeted settings and untargeted settings, which could inform design choices for applications relying on a system like mine.

6.6.4 `Familiarity` experiments

I used the limited amount of familiar NS data available to compare system performance when using familiar versus crowdsourced models. The familiar and crowdsourced models discussed here are all mixed (first and second responses), because too few familiar responses are available for adequate primary response mod-

	NS model sample size = 14			
	`targeted`		`untargeted`	
	System	SBERT	System	SBERT
count	180	60	180	60
mean	0.380	***0.530***	0.300	**0.444**
median	0.369	***0.545***	0.314	**0.472**
min	-0.147	**-0.138**	-0.181	***-0.107***
max	0.840	**0.879**	***0.900***	0.881
std dev	0.241	0.192	0.204	0.191

	NS model sample size = 50			
	`targeted`		`untargeted`	
	System	SBERT	System	SBERT
count	180	60	180	60
mean	0.393	***0.550***	0.305	**0.469**
median	0.389	***0.564***	0.323	**0.496**
min	**-0.048**	-0.090	-0.185	***0.132***
max	0.872	**0.881**	***0.898***	0.880
std dev	0.234	0.173	0.208	0.172

Table 6.18: Comparing Spearman rank correlation scores for `targeted` and `untargeted` versions of the PDT data, using NS models of either 14 or 50 random responses per item. Each *System* column represents 180 different rankings (6 system configurations $\times$ 30 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking. Each *SBERT* column represents 60 rankings (2 system configurations $\times$ 30 items; SBERT operates on plain text, so the `term representation` parameter does not apply).

els. The results, presented in Table 6.19, show no practical difference between `familiar` NS models and `crowdsourced` NS models when it comes to system performance. The mean Spearman score is slightly higher for the `crowdsourced` NS model, but the median Spearman score is slightly higher for the `familiar` NS model. Future work should collect more `familiar` responses in order to explore whether this pattern holds for larger model sizes or not. If there are differences, it may be possible to pinpoint the optimal model size depending on whether the available NS responses come from `familiar` or `crowdsourced` NSs.

SBERT performance differs here in that it is clearly better for the `familiar` models

	NS model sample size = 14			
	`familiar` NS		`Crowd` NS	
	System	SBERT	System	SBERT
count	180	60	180	60
mean	0.338	***0.499***	0.339	**0.481**
median	0.329	***0.513***	0.326	**0.500**
min	-0.239	***-0.026***	-0.181	**-0.125**
max	***0.896***	0.880	0.875	**0.879**
std dev	0.217	0.173	0.224	0.185

Table 6.19: Comparing Spearman rank correlation scores where `familiar` NS models contain only responses from participants *familiar* to the researcher and `Crowd` NS models contain only responses from `crowdsourced` participants. Results are shown using NS models of 14 responses; note the models used here are `mixed` (containing first and second responses; see Section 6.6.5) due to the sparsity of `familiar` data. Each *System* column represents 180 different rankings (6 system configurations $\times$ 30 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking. Each *SBERT* column represents 60 rankings (2 system configurations $\times$ 30 items; SBERT operates on plain text, so the `term representation` parameter does not apply).

than the `crowdsourced` models. Standardized type-to-token ratio may be informative here. For the NNS test sets, the mean STTR is 0.505, which is closer to that of the `familiar` models (0.576) than that of the `crowdsourced` models (0.652; Table 6.3). This suggests that SBERT is more sensitive than my system to differences in complexity between the NS model and NNS test sample.

6.6.5 `Primacy` experiments

I also compared how well my system approximates the weighted annotation ranking when using models comprised of NSs' first responses (`primary` models) versus models comprised of a mix of NSs' first and second responses (`mixed` models). I collected these second responses in order to broaden the range of NS responses, which were observed in Chapter 3 to be highly convergent for some items (e.g., *raking*). I expected noticeably different performance from the `primary` and `mixed` settings, but the results, presented in Table 6.20, show that the parameter has little impact. For the 50-response models, the best

setting is unclear: the mean Spearman score is slightly higher for the `primary` setting, but the median Spearman score is slightly higher for the `mixed` setting. For the 14-response models, the mean and median Spearman scores are both slightly higher for the `mixed` setting. These differences are likely too small to claim statistical significance, but it is reasonable that at smaller sample sizes, performance is improved by including secondary responses, as in the `mixed` models. These secondary responses elicit greater variety, as reflected in standardized type-to-token ratios (Table 6.3).

	NS model sample size = 14			
	`primary`		`mixed`	
	System	SBERT	System	SBERT
count	180	60	180	60
mean	0.339	***0.493***	0.340	**0.481**
median	0.326	***0.517***	0.334	**0.500**
min	-0.181	**-0.138**	-0.158	***-0.125***
max	**0.875**	0.881	***0.900***	0.879
std dev	0.224	0.207	0.230	0.185

	NS model sample size = 50			
	`primary`		`mixed`	
	System	SBERT	System	SBERT
count	180	60	180	60
mean	0.354	***0.514***	0.344	**0.505**
median	0.345	***0.532***	0.350	**0.518**
min	-0.185	**-0.090**	-0.147	***0.049***
max	***0.898***	0.880	**0.894**	0.881
std dev	0.226	0.186	0.226	0.168

Table 6.20: Comparing Spearman rank correlation scores where `primary` models contain only first responses from NSs and `mixed` models contain an equal mix of first and second responses from NSs. Results are shown using NS models of 14 responses and 50 responses. Each *System* column represents 180 different rankings (6 system configurations $\times$ 30 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking. Each *SBERT* column represents 60 rankings (2 system configurations $\times$ 30 items; SBERT operates on plain text, so the `term representation` parameter does not apply).

Examining this parameter across the two NS model sizes, it is clear that the 50-response model always results in better system performance than the 14-response model, regardless

of the primary or mixed setting. The mean and median Spearman scores show that this increase in performance is greater for the primary setting than for the mixed setting. This is unsurprising; if increasing variety (to some extent) in the NS model can improve performance, the more constrained primary setting should benefit more from an increased sample size than should the (already relatively varied) mixed setting. More NS responses would be needed to fully explore this curve and establish statistical significance. The mixed models here all contain half primary responses and half secondary responses. This default 50-50 mix is unlikely to be optimal, and with enough data, it would even be possible to determine exactly how many primary and secondary responses to include when sampling NS models for new PDT items in order to optimize system performance.

As seen in the other holistic experiments, SBERT performance here generally exceeds system performance; given access to the same NS response models, SBERT's similarity measures result in rankings that better approximate the weighted annotation rankings. My system does achieve higher maximum Spearman scores, but with no clear pattern to predict such cases, these higher maximums are not useful.

6.6.6 `Term representation` experiments

The current system allows for different term representations, which are variations on syntactic dependencies. 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 (see Table 3.6). 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, my system uses a "bag of terms" approach, where 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 approximating a bag of words approach, although, as seen throughout this chapter, xdx models are among the top per-

formers in many cases.

	NS model sample size = 14			
	ldh	xdh	xdx	SBERT
count	120	120	120	120
mean	0.333	0.336	**0.351**	***0.487***
median	0.318	**0.344**	0.330	***0.507***
min	***-0.108***	-0.181	-0.158	-0.138
max	0.871	0.875	***0.900***	0.881
std dev	0.223	0.227	0.231	0.196

	NS model sample size = 50			
	ldh	xdh	xdx	SBERT
count	120	120	120	120
mean	**0.350**	0.349	0.348	***0.509***
median	0.364	**0.374**	0.331	***0.523***
min	-0.147	-0.185	***-0.062***	-0.090
max	0.892	0.893	***0.898***	0.881
std dev	0.229	0.236	0.213	0.177

Table 6.21: Comparing Spearman rank correlation scores where system configurations use different `term representations`: `ldh` (labeled dependencies), `xdh` (unlabeled dependencies), or `xdx` (dependents only; i.e., *words*). Results are shown using NS models of 14 responses and 50 responses. Each *System* and *SBERT* column represents 120 different rankings (4 system configurations $\times$ 30 items) of 70 NNS responses, where each ranking receives a Spearman score via comparison with the weighted annotation ranking.

The results in Table 6.21 show that regardless of `term representation`, performance is best with the 50-response NS models. This difference is greatest for the `ldh` models and smallest for the `xdx` models. This makes sense, given that the `ldh` representations are the most complex and `xdx` representations are least complex, because we can expect the simpler `xdx` models to become "saturated" with relevant NNS response terms at smaller sample sizes as compared to `ldh` models. In Table 6.3, this is reflected in standardized type-to-token ratios. For the `crowdsourced` models, the greatest difference in STTR is seen when moving from the `ldh` 14-response models to 50-response models. For the `xdh` representation, the two sample sizes have nearly identical STTRs, and for the `xdx` representation, the 14-response model has the highest STTR. This suggests the optimal

NS model size is correlated with the complexity. Considering Table 6.3, it appears that STTRs for the simpler `xdh` and `xdx` representations likely peak somewhere between the 14-response sample and 50-response sample, but `ldh` samples may see increased STTR for samples larger than 50 responses.

With more NS responses, future work could form larger samples and determine the size at which STTR reaches its maximum for `ldh` representations. The same could be done for `xdh` and `xdx` `term representations` as well. With such a "map" of the relationship between STTR and sample size for NS models, one could better explore the correlations between system performance and `term representations` across sample sizes. For processing new PDT items, this would make it possible to choose the optimal `term representation` based on the number of available NS responses.

6.7 Optimization conclusion

The experiments in this chapter uncovered a number of trends that are informative for optimizing my system's performance. I found that the smaller of the two NS model sizes works best for ranking NNS responses in a way that corresponds with individual feature annotations, as measured with mean average precision throughout Section 6.5. For the features most related to content or semantics (CORE EVENT, INTERPRETABILITY, and VERIFIABILITY) the most detailed representation (labeled dependencies) works best, while the dependent-only representation works best for ANSWERHOOD and GRAMMATICALITY.

In the experiments discussed in Section 6.6, I found the larger of the two NS model sizes generally works best for ranking NNS responses in a way that correlates with the weighted annotation score, a holistic measure of response quality. For `intransitives` and for `xdx` terms (dependents only), however, the smaller NS model size was found to work best. This is notable because these are the settings that result in the lowest standardized type-to-token ratios (STTRs), as seen in Table 6.3. This indicates that response complexity and optimal NS model size are correlated; for more complex items, larger mod-

els are needed. Among `targeting` and `primacy` settings, all settings show stronger Spearman scores with the larger models. This preference for larger models is greater for the settings with higher STTRs—`untargeted` and `mixed`—than for their less complex counterparts, `targeted` and `primary`, further highlighting the correlation between item complexity and NS model size. Because the 50-response NS models work best for most parameter settings here, future research should involve collecting more NS responses to see if even larger models can improve performance.

Finally, through comparisons with SBERT, these experiments revealed that my system can outperform a state-of-the-art language embedding tool at particular tasks. Namely, with regard to the five annotation features, my system ranks responses in a way that better captures the annotation values than does SBERT, as measured with mean average precision throughout Section 6.5. For approximating the ideal rankings, however, SBERT consistently outperformed my system, as measured with the Spearman rank correlations with the weighted annotation scores throughout Section 6.6. These findings would be useful to anyone implementing a system for processing NNS sentences in visual contexts. For relying on similarity with a NS model as a representation of overall NNS response quality, a sophisticated tool like SBERT better correlates with the desired rankings than a custom, dependency-based system, even after considerable attempts at optimizing settings. However, for focusing on specific, custom features of a response, as may be desirable in language learning contexts, a custom system can outperform such a tool. Moreover, because my system scores responses using syntactic dependencies rather than uninterpretable word embeddings, it is better suited for providing feedback. Taken together, these findings suggest that researchers and developers should consider these approaches to be complimentary, using them alone or jointly to suit the task at hand.

CHAPTER 7
CONCLUSION

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

Appendices

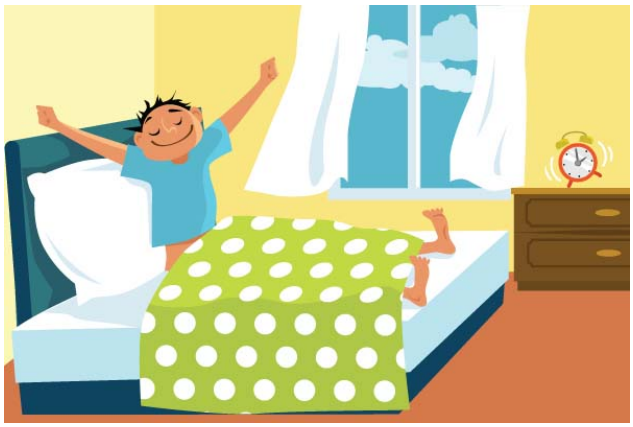
APPENDIX A

PDT ITEMS

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?"

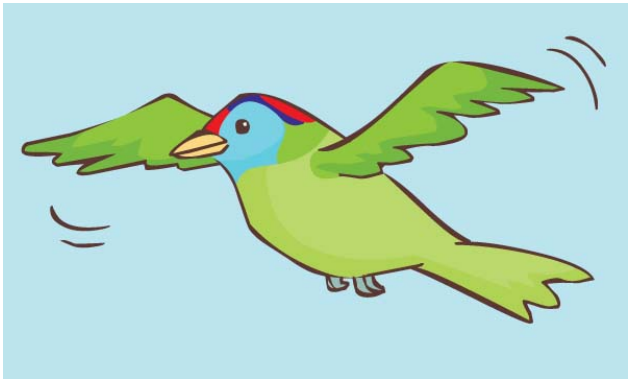
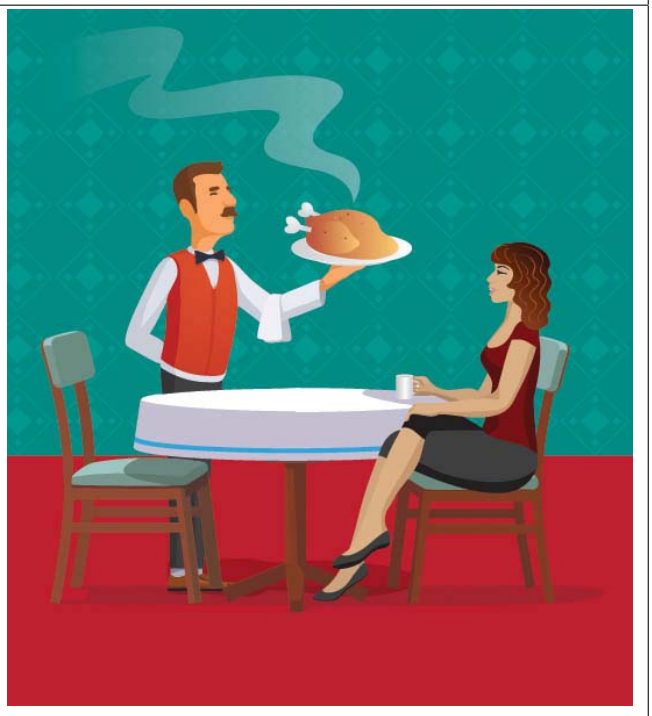
PDT Items

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?
		

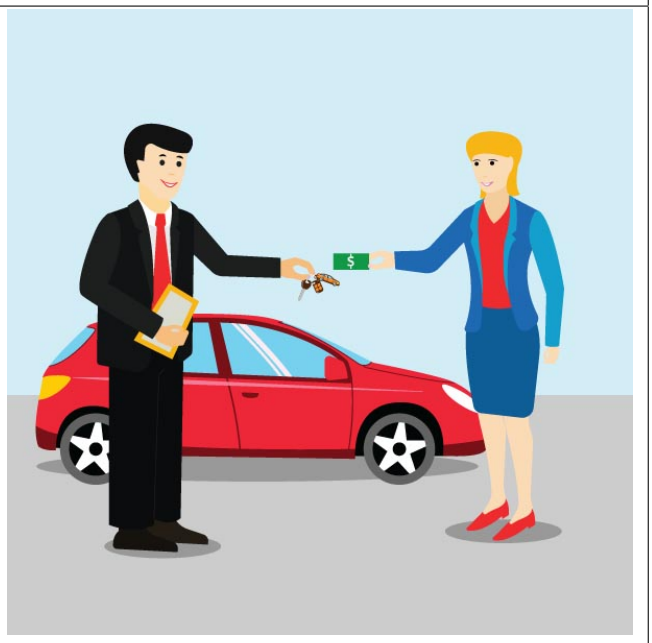
07: What is the bird doing?		08: What is the waiter doing?
		

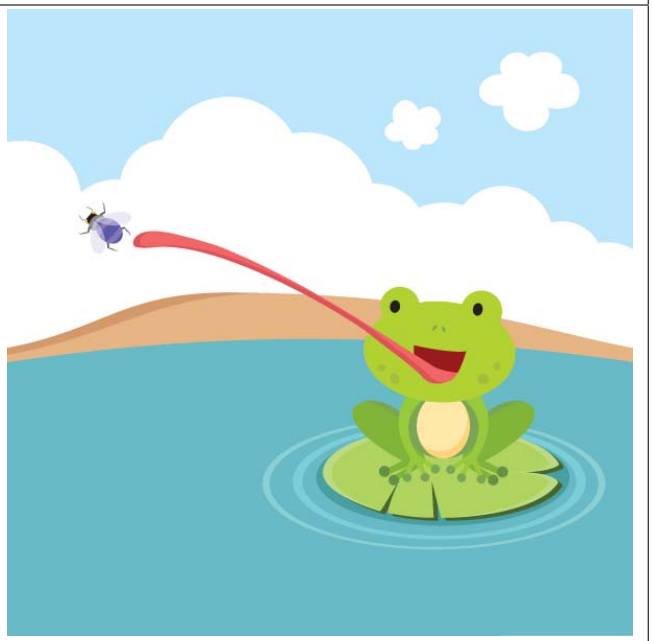
PDT Items

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

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

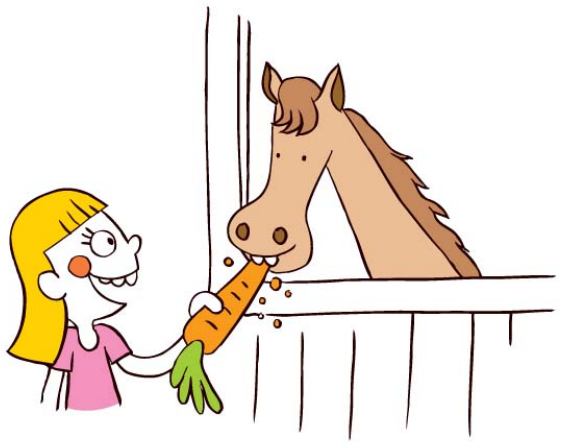
PDT Items

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?

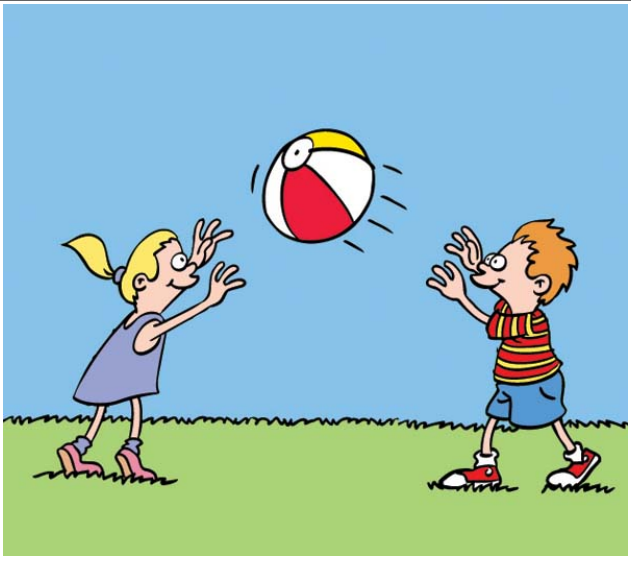
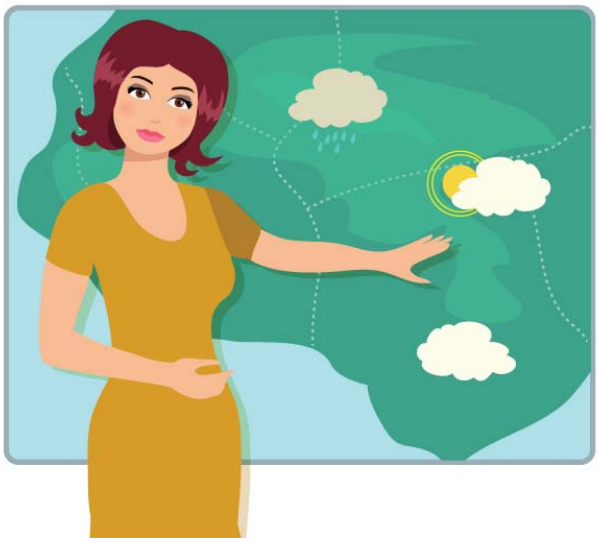


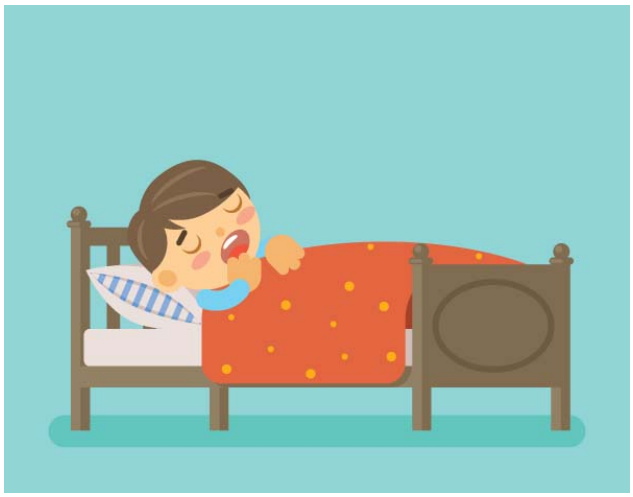
18: What is the man doing?

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 **3**

1.1 Overview . . . 3

1.2 Participants . . . 4

1.2.1 Non-native speakers . . . 4

1.2.2 Native speakers . . . 4

1.2.2.1 Familiar NSs . . . 4

1.2.2.2 Crowd-sourced NSs . . . 5

1.3 Instructions . . . 5

1.4 Item Examples (Targeted and Untargeted) . . . 6

2 Annotating Features **9**

2.1 Core event . . . 9

2.1.1 Contextuality of core event . . . 9

2.1.2 Defining `core event` . . . 9

2.1.2.1 Subjects . . . 9

2.1.2.2 Verb forms . . . 10

2.1.2.3 Content . . . 11

2.1.3 Alternative interpretations & inaccurate information . . . 12

2.1.4 Language problems . . . 12

2.1.5 Imprecise language . . . 13

2.1.6 Slang . . . 13

2.1.7 Intransitive vs. transitive core events . . . 13

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

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

1 Task Background

1.1 Overview

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

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

In order to focus attention on the main action, images contain very little background or other detail. Participants were instructed to provide one complete sentence capturing the main action in the image.

The data collected in the task will be used to analyze the differences in English **native speaker (NS)** and **non-native speaker (NNS)** language use. The researchers intend to study the many ways in which responses vary, and to compare these variations for NS and NNS responses. Ultimately, the researchers intend to use the NS responses to derive a kind of answer key or **gold standard (GS)**, which can be used to automatically evaluate the content of NNS responses.

1.2 Participants

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

1.2.1 Non-native speakers

NNS participants were recruited from intermediate and advanced level English as a Second Language (ESL) courses in the English Language Improvement Program at Indiana University. Roughly 140 NNS students completed the PDT. These participants all performed the task independently in a computer lab, with the researchers present. Responses from this group appear to be given in good faith.

1.2.2 Native speakers

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

1.2.2.1 Familiar NSs

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

in good faith.

1.2.2.2 Crowd-sourced NSs

Responses were also collected from roughly 330 different **crowd-sourced** NSs through the online platform, Survey Monkey. The researchers purchased survey responses from the platform's pool of users, who may win prizes or earn donations for charities in exchange for completing surveys. These participants all performed the task remotely, without the researchers present.

Crowd-sourced participants are less likely to complete a lengthy task, so the PDT was divided into four smaller tasks, and each crowd-sourced NS completed only one of these. Additionally, a sizable number of these participants completed only part of their task before abandoning it. The resulting data set is equivalent in size to roughly 150 completed familiar NS PDTs. Responses from the crowd-sourced group are of varying reliability; the majority are legitimate and in good faith, but some responses clearly are not. Some crowd-sourced NSs simply typed random characters in the response fields in order to move on to the next item and complete the task with minimal time and effort. Others responded with jokes, sarcasm or profanity.

1.3 Instructions

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

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

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

> Then, you will be asked to write a second, *different* answer, which is also a complete sentence. This might involve rewording or reorganizing your first sentence.

It does not need to be *completely* different; some words may be the same. If you cannot think of another way to answer the question, you may leave the second answer space empty, but any second responses you provide will be greatly appreciated.

1.4 Item Examples (Targeted and Untargeted)

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

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

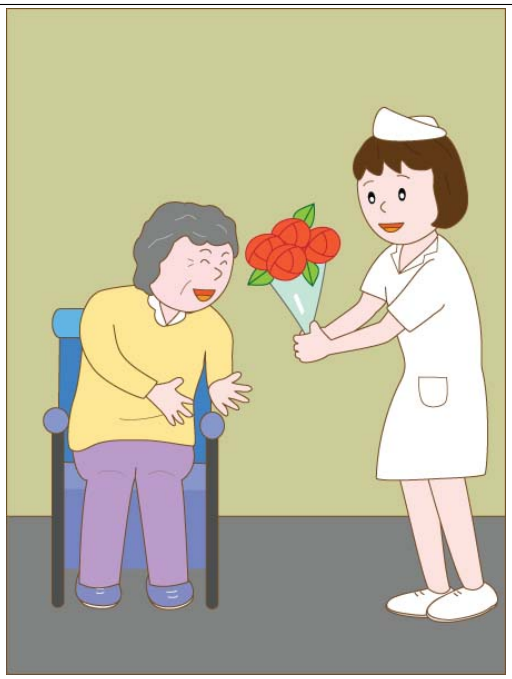
The terms *targeted* and *untargeted* are never used in the task, and participants are not explicitly informed of these differences. They are, however, provided with an example of each type immediately following the instructions, as seen in Figures 1 and 2 below.

Example 1

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.

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

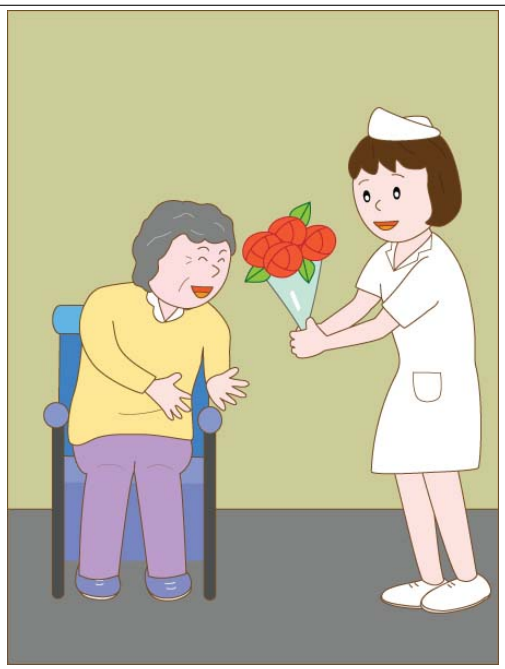
The answerhood feature primarily considers the following question: *Exactly as written, does the response make an attempt to answer the specific question asked?*

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

(*(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 and Detmar Meurers. 2011. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, 23(1):4–24.

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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Martin Bygate, Peter Skehan, and Merrill Swain. 2001. Researching pedagogical tasks: second language learning, teaching, and assessment.

Aoife Cahill, Binod Gyawali, and James Bruno. 2014. Self-training for parsing learner text. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 66–73, Dublin, Ireland. Dublin City University.

Marianne Celce-Murcia. 1991. Grammar pedagogy in second and foreign language teaching. *TESOL Quarterly*, 25:459–480.

Marianne Celce-Murcia. 2002. Why it makes sense to teach grammar through context and through discourse. In Eli Hinkel and Sandra Fotos, editors, *New perspectives on grammar teaching in second language classrooms*, pages 119–134. Lawrence Erlbaum, Mahwah, NJ.

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.

Karina Collentine. 2011. Learner autonomy in a task-based 3d world and production. *Language Learning & Technology*, 15(3):50–67.

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.

Luigi Di Caro and Matteo Grella. 2013. Sentiment analysis via dependency parsing. *Computer Standards & Interfaces*, 35(5):442–453.

Yadolah Dodge. 2008. *The concise encyclopedia of statistics*. Springer Science & Business Media.

Rod Ellis. 2000. Task-based research and language pedagogy. *Language Teaching Research*, 4(3):193–220.

Rod Ellis. 2003. *Task-based language learning and teaching*. Oxford university press.

Rod Ellis. 2006. Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly*, 40:83–107.

Katrina Forbes-McKay and Annalena Venneri. 2005. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological Sciences*, 26(4):243–254.

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.

Björn Granström. 2004. Towards a virtual language tutor. In *InSTIL/ICALL Symposium 2004*.

Jack Grieve. 2007. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3):251–270.

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

Wendell Johnson. 1944. Studies in language behavior: A program of research. *Psychological Monographs*, 56(2):1–15.

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.

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.

Diane Larsen-Freeman. 2002. Teaching grammar. In Diane Celce-Murcia, editor, *Teaching English as a second or foreign language*, 3rd edition, pages 251–266. Heinle & Heinle, Boston.

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.

David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical diversity and language development*. Springer.

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

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Steven Moran, Daniel McCloy, and Richard Wright. 2012. Revisiting population size vs. phoneme inventory size. *Language*, pages 877–893.

Noriko Nagata. 2002. BANZAI: An application of natural language processing to web based language learning. *CALICO Journal*, 19(3):583–599. `http://www.usfca.edu/japanese/CALICO02.pdf`.

Vivi Nastase, Jelber Sayyad Shirabad, and Maria Fernanda Caropreso. 2006. Using dependency relations for text classification. In *Proceedings of the 19th Canadian conference on artificial intelligence*, pages 12–25. Citeseer.

Rebecca L Oxford. 1993. Intelligent computers for learning languages: The view for language acquisition and instructional methodology. *Computer Assisted Language Learning*, 6(2):173–188.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Brian J Richards and David D Malvern. 2000. Accommodation in oral interviews between

foreign language learners and teachers who are not native speakers. *Studia Linguistica*, 54(2):260–271.

Mathias Schulze. 2010. Taking intelligent call to task. *Task-Based Language Learning and Teaching with Technology*, page 63.

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.

Tristan Thrush, Ethan Wilcox, and Roger Levy. 2020. Investigating novel verb learning in bert: Selectional preference classes and alternation-based syntactic generalization. *arXiv preprint arXiv:2011.02417*.

Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in Sport Sciences*, 21(1).

Shravan Vasishth and Bruno Nicenboim. 2016. Statistical methods for linguistic research: Foundational ideas–part i. *Language and Linguistics Compass*, 10(8):349–369.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Alex Warstadt and Samuel R Bowman. 2019. Linguistic analysis of pretrained sentence encoders with acceptability judgments. *arXiv preprint arXiv:1901.03438*.

Sara Cushing Weigle. 2013. English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, 18(1):85–99.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Kasumi Yamazaki. 2014. Toward integrative call: A progressive outlook on the history, trends, and issues of call. *TAPESTRY*, 6(1):6.

Jerrold H Zar. 1972. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 67(339):578–580.

Nicole Ziegler, Detmar Meurers, Patrick Rebuschat, Simon Ruiz, José L Moreno-Vega, Maria Chinkina, Wenjing Li, and Sarah Grey. 2017. Interdisciplinary research at the intersection of call, nlp, and sla: Methodological implications from an input enhancement project. *Language Learning*, 67(S1):209–231.

Daniel Zwillinger and Stephen Kokoska. 1999. *CRC standard probability and statistics tables and formulae*. Crc Press.

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.