CHAPTER 4
ANNOTATION & WEIGHTING

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

4.1 Annotation scheme

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

The annotation scheme was developed through an iterative process of annotation, discussion and revision, with input from two annotators and other language professionals. The annotation was developed on and applied to both NS and NNS responses. To avoid any potential bias, 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 4.3). Furthermore, the annotation lends insights into which aspects of a response are the most difficult to account for in my approach to content assessment.

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 Table 4.1, it is not clear how native-like *She is happy with the dog* is. Grammatically, it is native-like, but it does not seem like an appropriate answer to the question, *What is the woman doing?* Moreover, *The dog is so happy!* may be native-like in terms of language use, but does not seem appropriate in the context of the question. Thus, for the purpose of analyzing content in PDT responses, native-likeness seems to encompass considerations beyond language use and grammar.

Likewise, accuracy could not be satisfactorily defined as a simple *yes* or *no* construct. To illustrate, consider the ramifications of the response *hugging her dog Fluffy that she missed while on vacation* (Table 4.1) as either a NS or NNS response. The response does capture the main action of the item, but embellishes with unknowable details like the dog's name and the subject's motivation. This kind of response is undesirable in its own right, but could also lead to real problems during the automatic scoring. If included in a 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, it seems, is an inadequate construct for the approach to content assessment envisioned for this work. Clearly, *verifiability* is an important consideration as

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

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

well.

In order to handle the relevant kinds of variation observed in the responses, five binary features were eventually settled on, with each feature having some relation to the original concepts of accuracy and native-likeness. As with most annotation schemes in linguistics, the final SAILS scheme is a compromise. This scheme represents the minimal set of features that I believe necessary to accomplish two major goals of this work: investigating the use of NS responses as 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 re-

sponses. Such a discrimination was applicable to less than three percent of responses in the development set, however, so this feature was deemed too costly for the value it would provide.

A set of annotation guidelines were produced with definitions, rules and examples for each feature. For most features, the rules for targeted and untargeted items (See Section 3.1) 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 4.2.

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

2. **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!* (Table 4.1) 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 Table 4.1, *She is wear a blue dress* contains an ungrammatical verb form.

4. **INTERPRETABILITY**: Does the response evoke a clear mental image (even if different from the actual item image)? Any required verb arguments must be present and unambiguous. For example, *She loves her pet* (Table 4.1) is too vague to generate a

clear mental image. No action is specified (unless we force an unlikely reading of *loves* as a dynamic, simple present verb), and we cannot know if the *pet* is a dog, a 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 Table 4.1, *She is wear a blue dress* conveys information that is irrelevant to the core event but is nonetheless recoverable from the image (CORE EVENT=0, VERIFIABILITY=1), while *hugging her dog Fluffy that she missed while on vacation* fulfills the core event but also has information that cannot be verified from the picture (CORE EVENT=1, VERIFIABILITY=0).

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

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

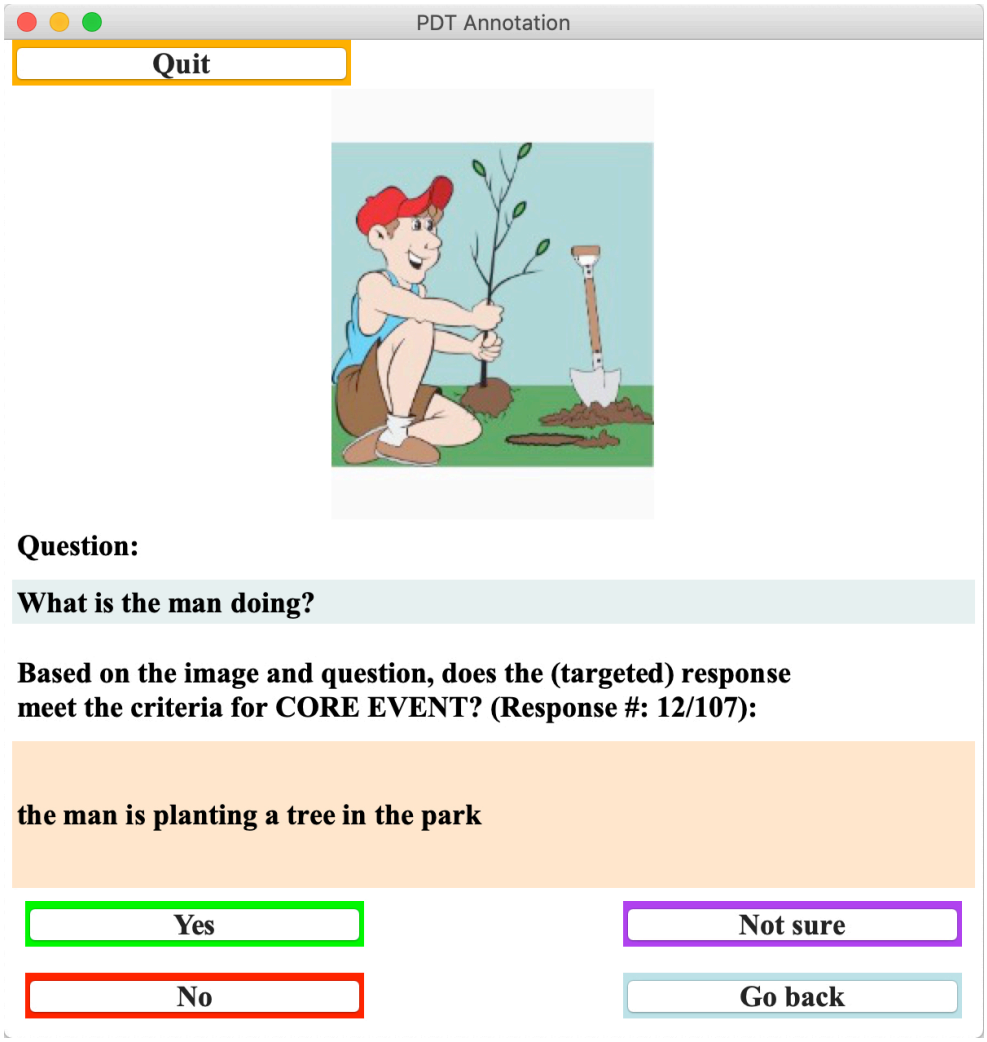


Figure 4.2: Interface used for feature annotations. Note that "Not sure" is not a final annotation value; it merely puts the response aside for a later decision.

4.2 Agreement

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

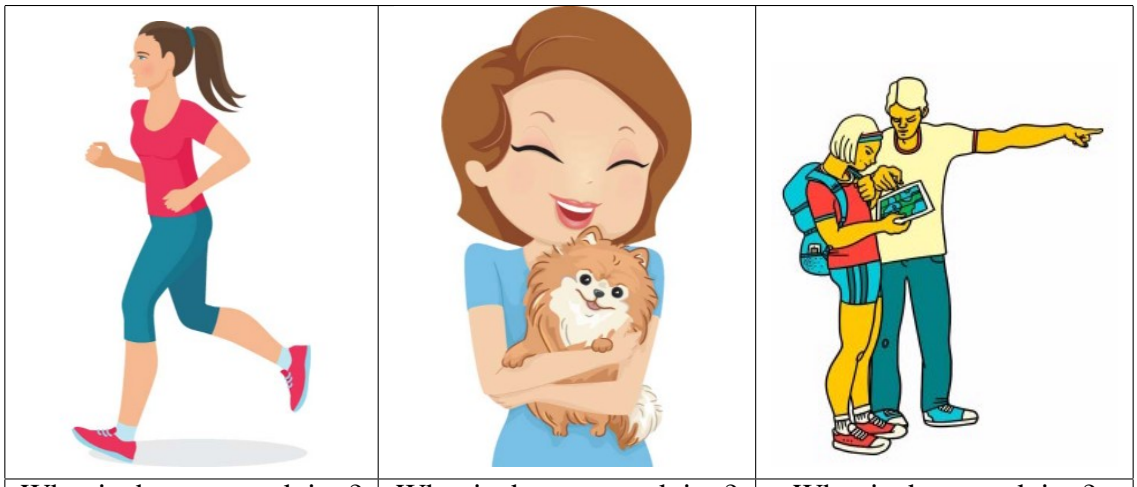
Three items were used as a development set for creating and revising the annotation scheme. These items were also used as examples in the guidelines. They represent one

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

Table 4.1: Targeted and untargeted sample responses from the development set transitive item, shown with adjudicated annotations for the five features: CORE EVENT (*C*), VERIFIABILITY (*V*), ANSWERHOOD (*A*), INTERPRETABILITY (*I*) and GRAMMATICALITY (*G*).

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

The test set parallels the development set and consists of one intransitive, one transitive and one ditransitive item; it is shown in Figure 4.3. Agreement and Cohen's kappa scores



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

Figure 4.3: 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.

are given in Table 4.2, 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.

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 4.2: 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*).

4.2.1 Transitivity

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

This is as expected, as ditransitive sentences are longer and have more verbal arguments, making for more opportunities for responses to vary (see Table 3.5), 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.

4.2.2 Targeting

Grouping the annotations into targeted and untargeted sets, the raw agreement scores are comparable (94.9% vs. 95.2%). However, despite a greater degree of response variation, the untargeted group has a higher kappa score (0.872 vs. 0.823).

When asked to compare the annotation process for targeted and untargeted items, A2 noted that targeted responses require more concentration and closer consultation of the guidelines. For example, ANSWERHOOD does not allow for targeted responses to modify the subject provided in the question in any way, whereas in answering *What is happening?*, the respondent is free to speak of characters in the pictures in many different ways. Both A1 and A2 thus describe the annotation of untargeted items as less restrictive and less time-consuming.

4.2.3 Features

Grouped by feature, the annotations all show raw agreement scores above 91% and Cohen's kappa scores above 0.74 (Table 4.2). For future use of this corpus in content assessment, these kappa scores are comfortably above the 0.67 suggested as a threshhold for meaning-

ful, 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 4.2.1).

The main issue in this case has to do with the amount of specificity required to capture the core event. The development set 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 4.3), and this resulted in a greater degree of variation. Many (particularly NNS) responses portray this not as a canonical *giving directions* event but as *pointing*, *helping a lost person* or *reading a map*, with A2 more likely to accept these less specific descriptions.

Similarly, but to a lesser extent, the transitive item, which shows a woman hugging a dog (Figure 4.3), 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 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 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 4.1);

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

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

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

Of 23 disagreements, seven stem from one annotator failing to enforce the requirement that a targeted response subject be either an appropriate pronoun or the exact subject given in the question, without adjectives, relative clauses or other modifiers. Given the question *What is **the woman** doing?*, for example, the responses *The **lady** is running* and *The woman*

who in pink *is running* were incorrectly accepted by one annotator each. While this criterion may seem strict, this subject-identity rule separates the task of identifying an attempt to answer the question from the task of verifying information (see VERIFIABILITY 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 for 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 4.3). 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 illustration as a cat. Each annotator incorrectly accepted one such response. One disagreement involved the misspelling of a crucial object: *The woman is holding the pat*. It is unclear whether *pet* or *cat* was intended. This should render the response unverifiable, but A1 accepted it.

The remaining disagreements are attributable to different opinions about inferences. 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.

4.2.4 NS & NNS responses

Response quality and annotation agreement were also calculated separately for NS and NNS responses, as shown in Table 4.3. 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.

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

Table 4.3: Comparing feature annotation agreement scores for NSs and NNSs: average *yes* annotations (*Average Yes*), total expected chance agreement (for *yes*es and *no*s) (*Chance Agree*), actual observed agreement (*Observed Agree*) and Cohen's kappa (*Kappa*).

Turning to the question of annotation quality, raw agreement scores are high among both groups, ranging from 91% to 99.3%. Notably, for CORE EVENT, VERIFIABILITY and INTERPRETABILITY, kappa scores are higher for NS responses than for NNS ones; i.e., annotators agree more on NS responses for these features. It may be no coincidence that these three features are the most closely tied to meaning, while ANSWERHOOD gets at

pragmatics and GRAMMATICALITY focuses on form.

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

4.3 Establishing Feature Weights

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

Clearly, weights must be assigned to each feature. One way to derive weights would be to manually rank responses from best to worst, then use the distribution of annotations across this ranking to determine some coefficient that represents the importance (weight) of each feature in the rankings. For each task item, the corpus contains roughly 150 NS responses and 70 NNS responses, so producing a manual ranking of the full set of responses is highly impractical. Manually ranking even a subset of 10 or 20 responses is frustrating and unreliable. Ranking a single pair of responses is a much more practical task, so I decided to have annotators perform 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 60 items (30 images presented with

two prompts each; see Section 3.3). 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 nothing 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 4.2.

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 the best response for the accompanying image and question. For our purposes, a good response is relevant and reasonable given the prompt. While you should consider form, please prioritize communicativeness and content. Naturally, you may consider what you know about the previously annotated features, but do not overthink them. These features are not of equal importance. A quick decision based on your own experience and intuition about communication is the goal here. If you feel that the responses are equally appropriate to the task, or if you cannot decide which is better, you may choose the "same/unsure"

option, but please do so sparingly.

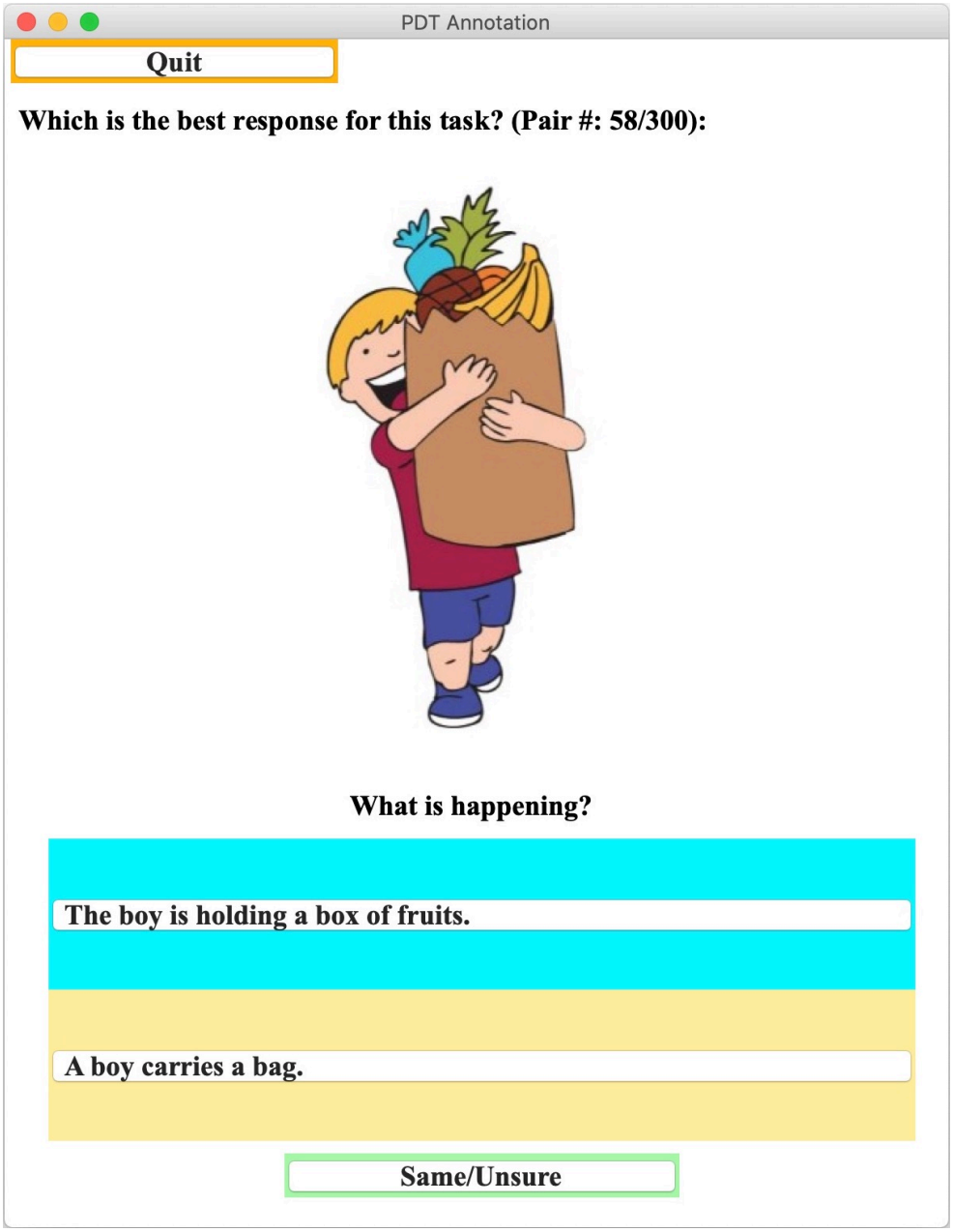


Figure 4.4: Annotation interface used for the preference test.

The preference test interface (Figure 4.4) 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 in-

cluded, 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 4.4. 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.

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

Table 4.4: Preference test sample responses pairs, annotator decisions (*A1 & A2*) and agreement for the item shown in Figure 4.4.

The two disagreements shown in the table are representative of the majority of the 35/300 disagreements in the sample in that one could make a reasonable argument for preferring either response (or marking them *same* in quality). Disagreement over *The boy is holding a box of fruits* and *A boy carries a bag* seems to involve the weighing of issues related to ANSWERHOOD (***is*** *hold**ing*** versus *carries*), CORE EVENT (i.e., the descriptiveness of *a box of fruits* versus *a bag*) and VERIFIABILITY (with *box* being quite clearly inaccurate). The disagreement in 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 A would likely not satisfy the CORE EVENT feature, while in response B, *towing* is a questionable verb choice, and *to the car* would arguably violate VERIFIABILITY because the image contains no car.

Agreement was calculated for the 300 response pairs judged by both annotators, presented in Table 4.5. The agreement rate of 0.883 with a Cohen's kappa of 0.692 confirms that high agreement on this task is both possible and reliable (Landis and Koch, 1977; Artstein and Poesio, 2008). With these scores, I am confident in using the full set of Annotator 1's 1,200 A/B decisions to derive the feature weights.

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

Table 4.5: 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 4.6.

	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 4.6: 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.

4.4 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 I believe the annotations and raw responses could find uses in question answering, dialog systems, pragmatic modeling, visual references and other challenges in natural language processing. The feature set could also be expanded to better suit other purposes, and the task could easily be extended to include new items. 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)).