Annotating Picture Description Task Responses for Content Analysis

Anonymous NAACL submission

Abstract

My abstract ...

1 Introduction

The (written) data of second language learners poses many challenges, whether it is being analyzed for grammatical errors (Leacock et al., 2014), for linguistic patterns (Kyle and Crossley, 2015), for content analysis (Weigle, 2013), or for interactions with intelligent computer-assisted language learning (ICALL) systems (Amaral and Meurers, 2007). One of the core issues in doing anything with learner data is the inherent amount of variability in how linguistic forms are used to convey meaning (Meurers and Dickinson, 2017). It may indeed seem like learners can use an infinite variety of forms to express a particular meaning. But the question of how large the problem of variability is for computational processing has rarely been investigated. More specifically, within the space of possible language productions, there is the further question of determining which are acceptable ones for a given setting.

Our overarching goal is to investigate these questions of variability and acceptability, both for non-native speakers (NNSs) and native speakers (NSs), given that all users of a language can be creative in their language usage. To that end—and taking *variability* to concern different mappings between linguistic form and its meaning—in this paper we control for meaning by collecting a dataset of picture description task (PDT) responses from a number of NSs and NNSs, and we annotate a handful of dimensions, thereby capturing the multifaceted ways in which responses can vary and can be acceptable. Outlining the decisions to be made highlights questions that need to be addressed by anyone working with learner language properties like variability, acceptability and native-likeness.

Given the form-meaning aspect of variability, we are interested in how variable linguistic behavior is *for the same content*, both within and among NSs and NNSs. There is a long-standing notion that systems processing learner data would be wise to constrain the data in some way (Heift and Schulze, 2007; Somasundaran and Chodorow, 2014; Somasundaran et al., 2015), but we do not know how much constraint is truly needed, and at what the cost is in terms of losing particular learner behavior for a constraint, without knowing more about the ways in which variation happens (see in particular Bailey and Meurers, 2008). Even today, the enterprises of ICALL and grammatical error correction (GEC), under their different conditions, would benefit by knowing more about the sources of difficulty in processing the range of learner data they do.

In annotating different dimensions of PDT responses, not only are we able to see how variable they are, but we are able to get a better handle on what could make a particular response "better" or "worse" for different kinds of purposes. For example, knowing that a person has gotten the main content of a picture correct, while adding information not present in the picture, may be treated differently than one who has made no such inferences but seems to be addressing a question about a different person in the picture (see section 3). The goals of this work are thus distinct from GEC (Leacock et al., 2014) and from more linguistically driven work such as parsing (e.g., Cahill et al., 2014; Ragheb and Dickinson, 2014), in that the acceptability of a response is taken as a function of several interacting features, most of which relate the text to the known semantic content. At the same time, providing the dimensions of acceptability and elucidating how they can be applied provides insight for any enterprise that desires to connect learner text with semantic content, in ad-

dition to unpacking the nature of variation more generally.

Taking a cue from King and Dickinson (2013), in section 2 we outline the picture description task (PDT) we use, designing items that elicit specific types of linguistic behavior. Section 3 outlines the annotation, specifically tackling the five-dimensional scheme; inter-anntotator agreement results are provided in section 4. While agreement seems reliable, highlighting areas of disagreement showcases difficult areas for establishing a link between form and meaning (cf., e.g., Meurers and Dickinson, 2017).

2 Picture Description Task

The PDT is built around 30 cartoon-like vector graphics. The images were modified to remove any non-essential detail or background; some examples are in Table 5. To factor out the influence of previous linguistic context, images are devoid of any text or symbols, with the exceptions of two images containing numerals, two with music notes, and one with a question mark. Each image depicts an ongoing or imminent action, performed by a person or an animal. The images are divided evenly into intransitive, transitive and ditransitive actions.

Two main versions of the PDT were used. In each version, the first half contains "targeted" items, where questions take the form of *What is <subject> doing?*, with the subject provided (e.g., *the boy*, *the bird*). The second half contains "untargeted" items, where each question asks *What is happening?*. Collecting both versions allows for the examination of response variation with and without a subject constraint, which should help inform approaches to task design and automatic content assessment. A roughly equal number of targeted and untargeted responses were collected for each item.

Each half (targeted and untargeted) is introduced with instructions, including an example item and responses. The instructions ask participants to focus on the main event depicted in the image and to respond with one complete sentence. The PDT was presented as an online survey and all participants typed their own responses. Participants were instructed not to use any reference materials, but they were permitted to use browser-based spell checking.

2.1 Data Collection

Responses were collected from 499 participants. Of these, 141 were NNSs, recruited from intermediate and advanced writing courses for English as a Second Language students attending a large public university in the US. These participants 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.

Of the 358 NS participants, 29 were personally known to and recruited by the researchers. Responses from the remaining 329 NSs were purchased via an online survey platform where participants earn credits they can redeem for gift cards and prizes. Due to length restrictions for paid surveys, these NSs each completed only half of the task, so their data is equivalent to 164.5 full participants.

In previous similar work (King and Dickinson, 2013, 2016), NSs were found to produce less variation than NNSs. Many NSs provided the same or very similar response with the most canonical way of expressing the main action. One purpose of the current corpus is to be able to assess NNS response content by comparing it against the NS responses; thus, NSs were asked to provide two non-identical responses, in the hopes that this would result in more examples of native-like responses for the variability of NNS responses to compare against.

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

To examine the degree of variation among the NS and NNS groups in the current study, type-to-token ratios (TTR) were calculated on the response level (ignoring case and final punctuation) for the entire data set. These ratios are shown in Table 1. NS responses far outnumber NNS responses in the data set, so for each item, a TTR for a random sample of 50 responses was calculated ten times. These were averaged for each item, then these averages were grouped as intransitive, tran-

sitive and ditransitive and averaged to obtain the ratio shown. These scores show that in all cases, the NS set shows a greater degree of response variation, meaning that when asked for two responses, NSs provide more unique response types per response.

The ratios show the direct relationship between the complexity of the event portrayed (represented here as intransitive, transitive and ditransitive) and the degree of variation elicited. In all cases, TTR increases with this complexity. Interestingly, this trend seems more pronounced in the NNS responses; in the targeted NNS responses, the TTRs for intransitive and ditransitive items, are 0.381 and 0.817, respectively, compared to 0.628 and 0.835 for NS responses. The ratios also show that in all cases, 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.

3 Annotation scheme

The data were annotated with the aim of providing information that would be useful for the automatic assessment of NNS responses via comparison with the NS responses. A five-dimension annotation, discussed in section 3, was developed to capture different facets of assessment; insights gained from the annotation, and in particular an interannotator agreement study, are covered in section 4.

The annotation scheme was developed through an iterative process of annotation, discussion and revision, with input from two annotators and multiple language professionals. The initial scheme was planned as a three-point scale, ranging from *accurate and native-like* (2) to *accurate but not native-like* (1) to *not accurate* (0). This proved problematic, however, as *accuracy* and *native-likeness* could not be adequately defined and applied to the data. For example, in the middle picture of Table 5, it is not clear how accurate or 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?*

To address the specifics of appropriate answers, five binary features were eventually settled on, with each feature having some relation to the original concepts of accuracy and native-likeness. A set of annotation guidelines were produced with definitions, rules and examples for each feature. The features and brief descriptions are listed here and discussed further in the following sections:

1. **Core Event**: Does the response capture the core event depicted in the image? Core events are not pre-defined but should be fairly obvious given the nature of the images. The response should link an appropriate subject to the event. In the top picture of Table 5, *The woman is running* clearly captures the core event, while *She is wearing a red shirt* is irrelevant to the event happening.

2. **Verifiability**: Does the response contain only information that is true and verifiable based on the image? Inferences should not be speculations and are allowed only when necessary and highly probable, as when describing a familial relationship between persons depicted in the image. For example, in Table 5, *She is wearing a red shirt* conveys information that is irrelevant to the core event but is nonetheless recoverable from the image (annotation=1), while *Trying to run from her bad decisions* has information that cannot be inferred from the picture.

3. **Answerhood**: Does the response make a clear attempt to answer the question? This generally requires a progressive verb. For targeted items, the subject of the question must be used as the subject of the response; appropriate pronouns are also acceptable. For example, *The dog is happy* is answering a question other than *What is the woman doing?* (Table 5).

4. **Interpretability**: Does the response evoke a clear mental image? Any required verb arguments must be present and unambiguous. For example, *The map is hard to read* is too vague to generate a clear mental image (see Table 5).

5. **Grammaticality**: Is the response free from errors of spelling and grammar? While the focus of GEC work, in our data set, this is a relatively straightforward feature to annotate (see section 4).

4 Agreement

Two annotators participated in the annotation. Both are native speakers of 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 intransitive, one transitive and one ditransitive item. Both annotators annotated the full development set multiple times throughout the process, discussing and adjudicating disagreeing annotations before moving on to the test set, which was completed without consultation between the annotators.

The test set parallels the development set and consists of one intransitive, one transitive and one ditransitive item; it is shown in Table 5. Agreement and Cohen's kappa scores are given in Table 3, broken down by different criteria. We will now walk through these results.

4.1 Transitivity

Comparisons of the intransitive, transitive and ditransitive items reveal an association between lower item complexity and higher agreement. The highest raw agreement and Cohen's kappa scores are found with the intransitive item (97.8%, $\kappa = 0.91$), and the lowest are found with the ditransitive item (92.4%, $\kappa = 0.76$).

This is as expected, as ditransitive sentences are longer and have more verb arguments, meaning there are a greater number of opportunities for responses to vary (see Table 1), and thus more opportunities for annotators to disagree on a given response. This trend also matches annotator feedback; both ranked the ditransitive item as the most difficult to annotate (for all features) and the intransitive as the easiest.

4.2 Targeted & untargeted prompts

When the annotations are grouped into targeted and untargeted sets, the raw agreement scores are comparable: 94.9% for targeted and 95.2% for untargeted items (Table 3). However, despite a greater degree of response variation, the untargeted group has a higher kappa score: 0.872 compared to 0.823.

When asked to compare the annotation of targeted and untargeted responses, 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.

4.3 Features

Grouped by feature, the annotations all show raw agreement scores above 91% and Cohen's kappa scores above 0.74. For future use of this corpus in content assessment, these kappa scores are comfortably above the 0.67 commonly suggested as a baseline for meaningful, reliable agreement (Landis and Koch, 1977; Artstein and Poesio, 2008). We discuss each feature in turn, highlighting difficulties in coming to an agreement, as such disagreements illustrate some of the sources of variability.

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

`Core event` has the second lowest inter-annotator agreement kappa score, at 0.808. This is somewhat lower than expected, as the pre-adjudication development set score was 0.889. This appears to be largely attributable to the difficulty of the ditransitive item, which was challenging for both participants and annotators (section 4.1).

The main issue in this case has to do with the amount of specificity required to be the core event. The development set item depicts a man delivering a package to a woman, and many participants were familiar with the appropriate vocabulary and constructions. The test set item shows a man giving directions to a woman (Table 5), and this resulted in a greater degree of variation, as many NNSs did not describe this in a canonical way. Rather than constructions like *asking X for directions* or *giving directions to X*, many NNSs describe the item with phrases like *pointing*, *guiding*, *helping a lost person* or *reading a map*, and most disagreeing core event annotations involve such responses, with A2

MD: Still thinking through the connection with AvgYes scores ...

LK: Need to back this up with some numbers!

MD: Should we report Table 3 broken down by NS and NNS agreements?

more likely than A1 to accept these less specific descriptions.

Similarly, but to a lesser extent, the transitive item, which shows a woman hugging a dog (Table 5), 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 easily 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.

MD: Would determining the appropriate level of specificity require knowing the end use of the annotation?

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 uninterpretable misspelling of a crucial object: *The woman is holding the pat*. It is unclear whether *pet* or *cat* was intended. This should render the response unverifiable, but A1 accepted it.

The remaining disagreements are attributable to different opinions about inferences, with A2 being, in general, more strict. For the ditransitive item, for example, both annotators accept responses that refer to the woman as a *hiker*, but only A1 accepts responses where the man and woman are collectively referred to as hikers. For the intransitive item depicting a woman running, A1 accepts multiple responses that refer to this as *a race*, as well as responses that infer the runner's motivation (fitness, leisure, etc.).

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 three categories: i. responses that do not directly answer the given question, perhaps by reframing the perspective so that it seems like a different question was asked; ii. responses that are gibberish or very low-effort and entered only so the participant can proceed to the next item; and iii. "troll" responses that attempt to be funny or obscene at the cost of attempting a direct answer.

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. While it may seem like a strict criterion, 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, which is generally required here 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*). This suggests that annotator fatigue accounts for the majority of `answerhood` disagreements.

Interpretability The average *yes* rate for `interpretability` is 0.802; only core event is less skewed: `interpretability` is thus also a feature where responses were more likely to be unacceptable.

The raw agreement score is 91.9% and kappa

is 0.744, the lowest agreement scores among the five features. This was anticipated, because interpretability is perhaps the most difficult feature to define, leaving much 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 contain all `core event` information and be completely `verifiable` may still fall short, in that the man could be picking fruit, building a bridge, etc.

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. Any verb arguments in the response would be identifiable in the image – not obscured or out of the frame. 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.

MD: I'm wondering how this relates to the specificity discussion for core event?

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 out of the frame. *A man is delivering a package*, on the other hand, would be accepted. An illustrator could simply show a delivery person carrying a package, as an indirect object would not be 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 becomes uninterpretable. Among the true disagreements, the level of specificity is often the point of contention, as with the `core event` feature. For example, A1 accepted several transitive item responses with the verb *love*, as in *The woman loves her dog.* In discussion, A2 explained that these are too vague to illustrate as an action; A1 disagreed, and this seems to indicate differing judgments regarding the use of *love* as a dynamic verb.

Grammaticality The `grammaticality` feature is the most heavily skewed one, with an average *yes* rate of 83.3%. As the only non-semantic annotation, this is perhaps not surprising.

Grammaticality has a raw agreement score of 93.0% and a kappa of 0.827. Among 52 disagreements, annotators agreed 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 disagreements, annotators could not agree on the acceptability of prepositions, as in *The girl is asking for help **to** the man* and *The girl is hugging **with** her cat*.

MD: Cite paper by Joel & others where they proposed a weighted metric for preposition error detection because people can't agree!

5 Discussion

5.1 Annotator feedback

Annotators' impressions of the task;
what is difficult? what is easy?

5.2 Trends (Agreement, etc)

(Do trends align with annotator feedback?)
What kinds of items & responses were challenging?
Are the trends for individual features? Recurring disagreements?

5.3 Limitations

what we'd do differently:
which images are problematic and why (symbols, ambiguity);
Which features are problematic; useful/not useful;

5.4 Potential Uses

Use of corpus in the next phase of my work;
Suggestions of other projects and research questions for this corpus.

Acknowledgments

(Advisors, annotators)

References

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

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4).

Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In *The 3rd Workshop on Innovative Use of NLP for Building Educational Applications (ACL08-NLP-Education)*, pages 107–115, Columbus, OH.

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

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

Levi King and Markus Dickinson. 2013. Shallow semantic analysis of interactive learner sentences. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–21, Atlanta, Georgia.

Levi King and Markus Dickinson. 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.

Kristopher Kyle and Scott A Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. *Automated Grammatical Error Detection for Language Learners*, second edition. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Detmar Meurers and Markus Dickinson. 2017. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning. Special Issue on Language learning research at the intersection of experimental, corpus-based and computational methods*, 67(S1):66–95.

Marwa Ragheb and Markus Dickinson. 2014. The effect of annotation scheme decisions on parsing learner data. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, Tübingen, Germany.

Swapna Somasundaran and Martin Chodorow. 2014. Automated measures of specific vocabulary knowledge from constructed responses ('Use these words to write a sentence based on this picture'). In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, Baltimore, Maryland.

Swapna Somasundaran, Chong Min Lee, Martin Chodorow, and Xinhao Wang. 2015. Automated scoring of picture-based story narration. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–48, Denver, Colorado. Association for Computational Linguistics.

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

What is the woman doing? [Intrans.]	A1	A2
The woman is running.	1	1
She is wearing a red shirt.	0	0
Trying to run from her bad decisions.	1	0

What is the woman doing? [Trans.]	A1	A2
She's holding a puppy & looks happy.	1	1
She is happy with the dog.	0	0
The lady loves her dog.	1	0

What is the man doing? [Ditrans.]	A1	A2
giving directions to a woman.	1	1
The man is reading a map.	0	0
The man is is telling her where to go.	1	0

Table 2: Test sample items and example responses with Core Event annotations from Annotators 1 and 2.

Set	Total	A1Yes	A2Yes	AvgYes	Chance	Agree	Kappa
Intrans	2155	0.863	0.855	0.859	0.758	0.978	0.910
Trans	2155	0.780	0.774	0.777	0.653	0.949	0.853
Ditrans	2155	0.812	0.786	0.799	0.678	0.924	0.764
Target	3390	0.829	0.818	0.824	0.709	0.949	0.823
Untarg	3075	0.806	0.790	0.798	0.678	0.952	0.872
Core	1293	0.733	0.717	0.725	0.601	0.923	0.808
Verif	1293	0.845	0.817	0.831	0.719	0.968	0.884
Answer	1293	0.834	0.831	0.833	0.721	0.982	0.936
Interp	1293	0.818	0.787	0.802	0.682	0.919	0.744
Gramm	1293	0.861	0.872	0.866	0.768	0.960	0.827

Table 3: Agreement scores broken down by different properties of the test set: total annotations (*Total*), *yes* annotations for Annotator 1 and 2 (*A1Yes*, *A2Yes*), average *yes* annotations (*AvgYes*), total expected chance agreement for *yes*es and *no*s (*Chance*), actual raw agreement (*Agree*) and Cohen's kappa (*Kappa*).

	Total		AvgYes		Chance		Agree		Kappa	
Set	NS	NNS	NS	NNS	NS	NNS	NS	NNS	NS	NNS
Intrans	1450	705	0.814	0.952	0.697	0.909	0.974	0.987	0.914	0.859
Trans	1450	705	0.768	0.796	0.643	0.675	0.949	0.949	0.857	0.843
Ditrans	1450	705	0.794	0.808	0.672	0.689	0.922	0.928	0.762	0.767
Target	2340	1050	0.812	0.849	0.695	0.743	0.948	0.950	0.829	0.807
Untarg	2010	1065	0.768	0.855	0.643	0.753	0.949	0.959	0.856	0.833
Core	870	423	0.686	0.805	0.569	0.686	0.922	0.927	0.819	0.767
Verif	870	423	0.807	0.882	0.688	0.791	0.970	0.962	0.904	0.819
Answer	870	423	0.800	0.899	0.680	0.819	0.977	0.993	0.928	0.961
Interp	870	423	0.764	0.881	0.638	0.789	0.910	0.936	0.752	0.697
Gramm	870	423	0.902	0.793	0.823	0.671	0.962	0.955	0.786	0.863

Table 4: Comparing agreement for NS and NNS responses, with agreement scores broken down by different properties of the test set: total annotations (*Total*), average *yes* annotations (*AvgYes*), total expected chance agreement for *yes*es and *no*s (*Chance*), actual raw agreement (*Agree*) and Cohen's kappa (*Kappa*).

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 5: Targeted and untargeted sample responses from the development set transitive item, shown with adjudicated annotations for the five features: core event, verifiability, answerhood, interpretability and grammaticality.