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 (CITE), for linguistic patterns (CITE), for content analysis (CITE), or for interactions with intelligent computer-assisted language learning (ICALL) systems (CITE). 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 this question has hardly been investigated: how big is the problem of variability? And how does it differ from native speaker (NS) variability, given that all users of a language, native or non-native (NNS), can be creative in their language usage? To investigate these questions, we control for meaning by collecting a dataset of picture description task (PDT) responses among a number of NSs and NNSs, and annotating along a handful of dimensions, to capture the multifaceted ways in which responses can vary.

In other words, 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 (CITE), but we do not know the specific effect of these constraints. In some ways, the enterprises of ICALL and grammatical error correction (GEC) started off without knowing how easy or difficult the problems were, under different kinds of conditions.

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 makes 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, but has also added information not strictly present in the picture, may be treated differently than one who made no such inferences but seemed to be addressing a different questions about the picture (see section 3.1). The goals of this work are thus distinct from GEC (CITE), or even from more linguistically driven work such as parsing (CITE), in that the acceptability of a response must be taken as a function of several interacting features.

- Motivate this for people who don't do this kind of work; measuring 'goodness' of responses; decomposing them into features; (this is not like grammatical error detection or parsing work); Getting a handle on variability, especially under different circumstances; is a response 'native-like'? in what ways (by which features?); breaking variability down into our features;
- Cite previous work:
  - ours (King and Dickinson, 2013)
  - and others' (Somasundaran et al., 2015)

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 1. 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?*. A roughly equal number of targeted and untargeted responses were collected for each item.

MD: Why do we distinguish targeted from untargeted items?

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.

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.

2.1 Data Collection

PDT versions;
NS vs NNS (1 vs 2 responses);
NNSs demographics (ELIP recruited);
NSs demographics and sources (known persons vs crowdsourced);
Response counts (types, tokens)

3 Annotation

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 3.1, was developed to capture different facets of assessment; insights gained from the annotation, and in particular an interannotator agreement study, are covered in section 3.2.

3.1 Scheme

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 1, 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 1, *The woman is running* clearly captures the core event, while *She is wearing a red shirt* is irrelevant to the event happening.

2. **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, in Table 1, *XXX* is answering a question other than *What is the woman doing?*

MD: Add faulty example, as this feature is non-obvious

3. **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 3.2).

4. **Interpretability**: Does the response evoke a

clear mental image? Any required verb arguments must be present and unambiguous. For example, in Table 1, *XXX* is too vague to generate a clear mental image.

MD: Add faulty example, as this feature is non-obvious

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 relationship between persons depicted in the image. For example, in Table 1, *XXX* conveys information that is irrelevant to the core event but is nonetheless recoverable from the photo (annotation=1), while *YYY* has information that cannot be inferred from the picture.

MD: Add examples

3.2 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 with 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 1. Agreement and Cohen's kappa scores are given in Table 2, broken down by different criteria. We will now walk through these results.

3.2.1 Intransitives, transitives & ditransitives

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, and the lowest are found with the ditransitive item.

LK: A type-token ratio table would be helpful here.

This is as expected, as ditransitive sentences are longer with more verb arguments, meaning there are a greater number of opportunities for responses to vary, 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.

3.2.2 Targeted & untargeted

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. However, despite a greater degree of response variation, the untargeted group has a higher kappa score: 87.2% compared to 82.3%. This may appear counterintuitive, but can most likely be attributed to the stricter annotation rules for targeted items. For example, *answerhood* does not allow targeted responses to modify the subject provided in the question in any way. The untargeted questions elicit a greater degree of variation, but the annotation scheme accommodates this.

are kappa scores percentages?

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. Both A1 and A2 describe the annotation of untargeted items as less restrictive.

3.2.3 Features

Grouped by feature, the annotations all show raw agreement scores above 91% and Cohen's kappa scores above 74%. For the future use of this corpus in content assessment, these kappa scores are comfortably above the 67% commonly suggested as a baseline for meaningful, reliable agreement [CITATION].

Core event This 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 80.8%. This is somewhat lower than expected, as it is lower than the pre-adjudication development set score (88.9%). This appears to be largely attributable to the difficulty of the ditransitive item, which was challenging for both participants and annotators. 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, 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 more likely than A1 to accept these less specific descriptions.

Need to back this up with some numbers!

To a lesser extent, the transitive item, which shows a woman hugging a dog, 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.

Answerhood This feature was added largely as a way to identify responses that do not follow the instructions. Such responses tend to fall into three categories: responses that simply do not directly answer the given question; responses that are gibberish or very low-effort and entered only so the participant can proceed to the next item; and "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 (93.6%) scores among the five features.

Of 23 disagreements, seven stem from an 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. (This rule separates the task of verifying information from the task of identifying an attempt to answer the question.) 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. Another 10 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. This suggests that a annotator fatigue accounts for more the majority of answerhood disagreements.

Grammaticality This is the most heavily skewed feature, with an average *yes* rate of 83.3%.

Grammaticality has a raw agreement score of 93.0% and a kappa of 82.7%. 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, which are 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*.

Interpretability The average *yes* rate for interpretability is 80.2%; only core event is less skewed.

The raw agreement score is 91.9% and the kappa is 74.4%, and these are the lowest agreement scores among the five features. This was anticipated, because interpretability is the most difficult feature to define and leaves 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. The guidelines place some restrictions on this mental image. If one were to illustrate it, 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. A response like *A woman is receiving a package* fails these criteria, 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*, would be accepted. An illustrator could simply show a delivery person carrying a package; an indirect object would not be necessary.

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, like subject verb disagreement, making the number of subjects or objects uninterpretable. Among the true disagreements, the level of specificity is often the point of contention. 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.

Verifiability 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 88.4%. 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, as in *A man is pointing the way for the women*, when the image shows only one woman. Two other errors stem from inaccuracy, with respondents referring to a dog in the illustration as a cat. One error involves the uninterpretable misspelling of a crucial object, which renders the response unverifiable.

The remaining disagreements are attributable to different opinions about inferences, and in general, A2 is more strict with verifiability. For the ditransitive item, for example, both annotators accept responses that refer to the woman as a *hiker*, but only A1 accepts responses where both the man and woman are 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.).

Relevant tables:
Types, Tokens, TypeTokenRatio
(for targeted vs. untargeted)

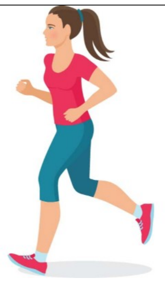
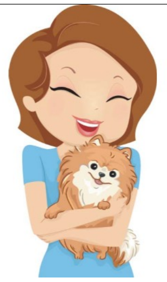
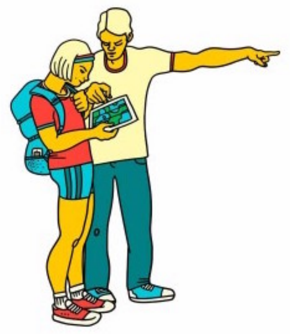
		
What is the woman doing? [Intrans.]	A1	A2
The woman is running.	1	1
She is wearing a red shirt.	0	0
Trying to run from her bad decisions.	1	0
		
What is the woman doing? [Trans.]	A1	A2
She's holding a puppy & looks happy.	1	1
She is happy with the dog.	0	0
The lady loves her dog.	1	0
		
What is the man 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 1: Test sample items and example responses with Core Event annotations from Annotators 1 and 2.

4 Discussion

4.1 Annotator feedback

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

4.2 Agreement & Disagreement Trends

(Do trends align with annotator feedback?)
What kinds of items & responses were challeng-

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	0.880
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
Answer	1293	0.834	0.831	0.833	0.721	0.982	0.936
Gramm	1293	0.861	0.872	0.866	0.768	0.960	0.827
Interp	1293	0.818	0.787	0.802	0.682	0.919	0.744
Verif	1293	0.845	0.817	0.831	0.719	0.968	0.884

Table 2: Agreement scores for different groupings of the test set, showing total annotations, *yes* annotations for Annotator 1 and 2, average *yes* annotations, total expected chance agreement (*yes* + *no*), actual raw agreement and Cohen's kappa.

ing?

Are the trends for individual features? Recurring disagreements?

4.3 Limitations

what we'd do differently:

which images are problematic and why (symbols, ambiguity);

Which features are problematic; useful/not useful;

4.4 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

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.

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.

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.