Annotating Picture Description Task Responses for Content Analysis

Levi King & Markus Dickinson
Indiana University

Overview & Background

- Introducing the **Semantic Analysis of Image-based Learner Sentences (SAILS) Corpus**
  - 13,533 picture description task (PDT) responses from native speakers (NS) and non-native speakers (NNS), each annotated for five binary features
- **Goal:** Evaluate content of NNS sentences
  - Compare to gold standard (GS) of NS sentences
- **Needs:** Adequate data, appropriately constrained
  - Large set of PDT responses from NS and NNS participants
  - Varied task prompts and participant demographics allowing for study of variability
  - Annotation allowing for content analysis

Picture Description Task

- PDT elicits natural productions but constrains form & content
- 60 **items**: 30 images *x* 2 prompts
  - 30 images
    - simple vector graphics
    - 10 transitive, 10 intransitive, 10 ditransitive actions
  - 2 prompt versions:
    - **targeted**: *What is <the subject> doing?*
    - **untargeted**: *What is happening?*

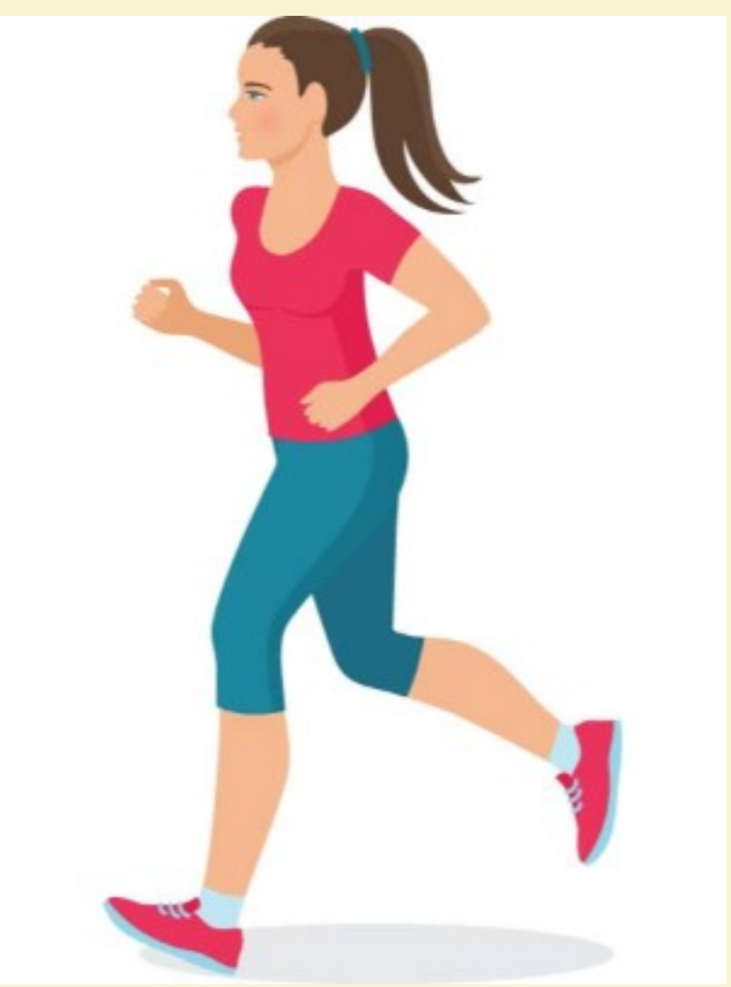
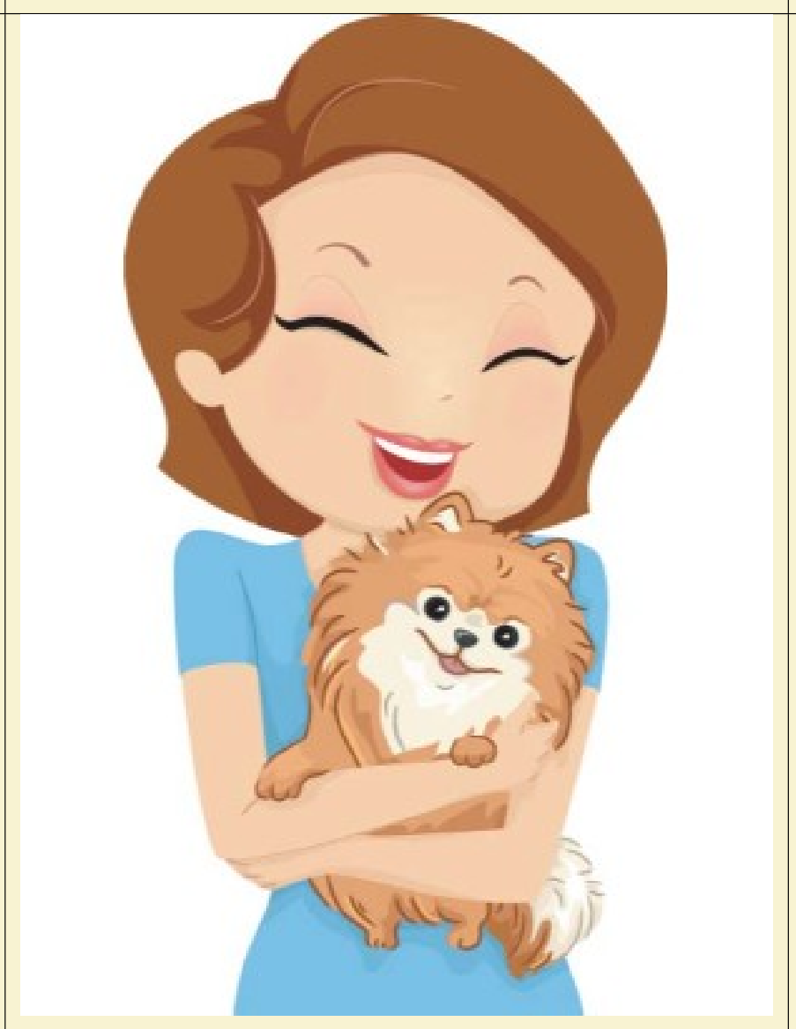
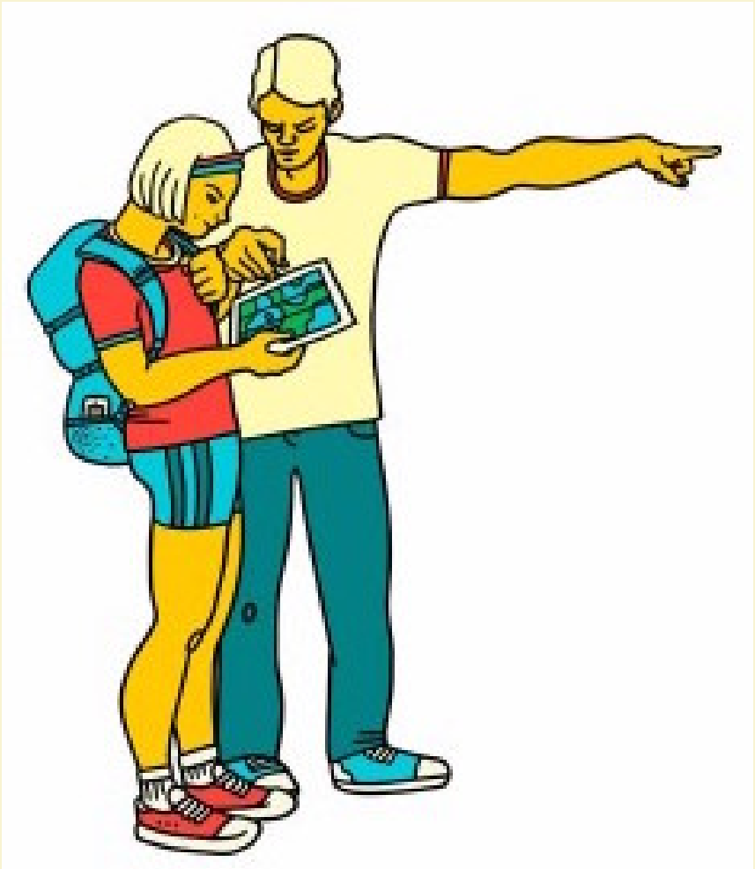
Intransitive	Transitive	Ditransitive
		
What is the woman doing?	What is the woman doing?	What is the man doing?

Table 1: Example PDT images with their **targeted** questions. In the **untargeted** form, the question for each is *What is happening?*

- PDT Instructions: Focus on main action; Respond with complete sentence
- Task administered as online survey (SurveyMonkey.com)
- Multiple versions
  - Most participants completed 30 items
  - Roughly equal number of targeted & untargeted responses collected per image
  - NNSs provide one response per item
  - NSs asked to provide two non-identical responses per item
    - Intended to increase variety of NS responses for a more robust GS

Participants

499 total participants

- 141 NNSs
  - Recruited from intermediate & advanced ESL writing courses at Indiana University
  - L1s: 125 Chinese (90%), 4 Korean, 3 Burmese, 2 Hindi; 1 each: Arabic, Indonesian, German, Gujarati, Spanish, Thai, Vietnamese
- 358 NSs
  - 29 Familiar Native Speakers (FNSs)
    - Relatives or friends of researchers; assumed to be high quality
  - 329 Crowdsourced Native Speakers (CNSs)
    - Responses purchased via SurveyMonkey; assumed to be lower quality

Responses

The SAILS Corpus contains a total of 13,533 PDT responses.

	Response Counts		
Group	First	Second	Total
NNS	4290	0	4290
NS (all)	4634	4609	9243
FNS	642	641	1283
CNS	3992	3968	7960
Total	8924	4609	13,533

Table 2: First and second response counts for the SAILS Corpus participant groups. Familiar (FNS) and crowdsourced (CNS) are subgroups of NS. NNS participants are not asked to provide a second response.

In order to examine the level of variation among responses, type to token ratios (TTRs) were calculated on the response level. Capitalization and final punctuation were ignored. We can see that variation increases with item complexity (intransitives < transitives < ditransitives) and that untargeted responses vary more than targeted responses.

	Targeted		Untargeted	
Set	NS	NNS	NS	NNS
Intransitives	0.628	0.381	0.782	0.492
Transitives	0.752	0.655	0.859	0.779
Ditransitives	0.835	0.817	0.942	0.936

Table 3: NS and NNS type-to-token ratios (TTR) for complete responses (not words), for the full corpus.

TTRs were also calculated to compare the variability among NSs' first responses versus their second responses. As the TTRs for second responses are considerably higher than those for first responses, asking for two non-identical responses appears to effectively increase the variety of NS responses available for use in a GS.

	Targeted		Untargeted	
Set	R1	R2	R1	R2
Intransitives	0.343	0.819	0.549	0.939
Transitives	0.509	0.895	0.682	0.926
Ditransitives	0.641	0.948	0.864	0.955

Table 4: TTRs for complete responses, separated by first responses (R1) and second responses (R2). The ratios here are calculated from all NS responses; NNS responses are not included.

Annotation Scheme

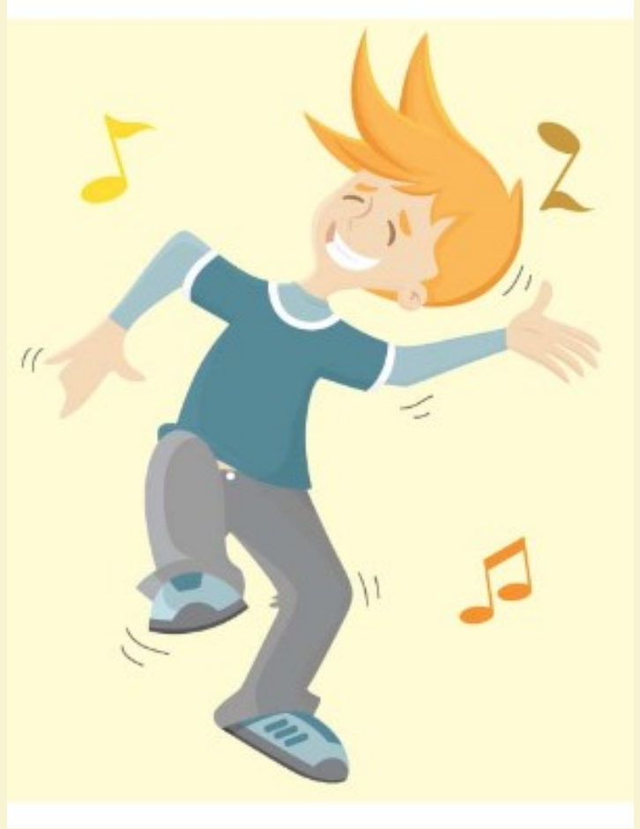
Two annotators:

- NSs (US English), both with language teaching experience (child & adult learners).
- Annotator 1 (A1) annotated the complete corpus, Annotator 2 (A2) annotated development set & test set, each containing 1 intransitive, 1 transitive, 1 ditransitive.

Initial scheme (*accurate* + *native-like* > *accurate* + *not native-like* > *not accurate*) proved problematic; evolved to five binary features related to accuracy & native-likeness.

1. **Core Event**: Does the response capture the core event depicted in the image?
2. **Verifiability**: Does the response contain only information that is true and verifiable based on the image? Inferences allowed only when necessary; e.g., familial relationships between persons in image.
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 or an appropriate pronoun must be used as the subject of the response.
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.
5. **Grammaticality**: Is the response free from errors of spelling and grammar?

Annotation Results



What is the boy doing? (Targeted)	C	V	A	I	G
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? (Untargeted)	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: 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*).

Using the test set items shown in Table 1, we calculated inter-annotator agreement for each feature, for targeted vs. untargeted items, and for the three verb types.

Set	Total	A1Yes	A2Yes	AvgYes	Chance	Agree	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
Verifiability	1293	0.845	0.817	0.831	0.719	0.968	0.884
Answerhood	1293	0.834	0.831	0.833	0.721	0.982	0.936
Interpretability	1293	0.818	0.787	0.802	0.682	0.919	0.744
Grammaticality	1293	0.861	0.872	0.866	0.768	0.960	0.827

Table 6: 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 *yeses* and *nos* (*Chance*), actual raw agreement (*Agree*) and Cohen's kappa (*Kappa*).

Observations from Table 6:

- Average *yes* rates are included to show that all features skew toward *yes* annotations; Cohen's kappa is thus used as a measure of inter-annotator agreement;
- Cohen's kappas well above the conventional 0.67 threshold for meaningful agreement, so we believe the annotation scheme can be implemented reliably by following the guidelines;
- Inter-annotator agreement decreases with item complexity, from intransitive to transitive to ditransitive verbs;
- Agreement is slightly higher for untargeted items than targeted items, likely due to the fact that annotation guidelines are less complicated for untargeted items;
- Among the features, answerhood has the highest kappa and interpretability has the lowest. This is unsurprising, as annotators reported these to be the easiest and hardest features to annotate, respectively.

Accessing the SAILS Corpus

The entire annotated SAILS Corpus, the PDTs and annotation guidelines are available to anyone at:

https://github.com/sailscorpus/sails

We believe the corpus can be a useful resource for language testing and ICALL, as well as other areas of research like question answering, dialog systems, pragmatic modeling, and visual references. We hope that other researchers will make use of the existing data or expand on it with new participants, items, and approaches for processing.

BEA13, 5 June 2018; New Orleans, LA

{leviking,md7}@indiana.edu