Annotating Picture Description Task Responses for Content Analysis

Levi King & Markus Dickinson
Indiana University

Overview

- **Semantic Analysis of Image-based Learner Sentences (SAILS) Corpus**
  - 13,533 picture description task (PDT) responses
  - Both native (NS) & non-native speakers (NNS)
  - Annotated for five binary features
- **Goal:** Evaluate content of NNS sentences
  - Compare to gold standard (GS) of NS sentences
- **Need:** Adequate data, appropriately constrained
  - Large set of PDT responses
  - Varied task prompts & participant demographics
  - Annotation 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 intransitive, 10 trans, 10 ditrans

2 prompts

- **Targeted**: *What is <the subject> doing?*
- **Untargeted**: *What is happening?*

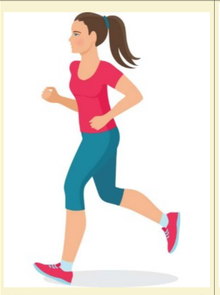
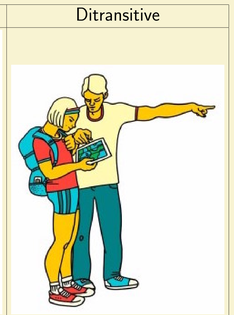
Intransitive	Transitive	Ditransitive
		
What is the woman doing?	What is the woman doing?	What is the man doing?

Table 1: Example PDT images with their **targeted** questions.

Administered as online survey (SurveyMonkey.com)

PDT Instructions

- Focus on the main action
- Respond in a complete sentence

Multiple versions

- Most participants completed 30 items
- Roughly equal number of targeted & untargeted responses
- NNSs provide one response per item
- NSs provide two non-identical responses per item (more robust GS)

Participants

499 total participants

- 141 NNSs: students in intermediate & advanced ESL writing courses at IU
  - 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 (assumedly higher quality)
  - 329 Crowdsourced Native Speakers (CNSs)
    - Responses purchased via SurveyMonkey (assumedly lower quality)

Responses

Response Counts

	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 & second response counts for SAILS Corpus participant groups

Type-Token Ratios (TTRs)

	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: TTRs for *complete responses* (not words), for full corpus

- Capitalization & final punctuation ignored
- Variation increases with:
  - Item complexity (intransitives < transitives < ditransitives)
  - Less targeting (targeted < untargeted)

Type-Token Ratios (TTRs): first vs. second responses (NSs only)

	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 (R1) & second responses (R2)

- TTRs for R2s considerably higher than for R1s

⇒ Asking for two responses increases variety of language available for use in GS

Annotation Scheme

Initial scheme: *accurate* + *native-like* > *accurate* + *not native-like* > *not accurate*)

Final scheme: five binary features related to accuracy & native-likeness:

1. **Core Event (C)**: Does response capture the core event depicted in image?
2. **Verifiability (V)**: Does response contain only true & verifiable info, based on image?
   - Inferences allowed only when necessary; e.g., familial relationships of persons in image
3. **Answerhood (A)**: Does response make a clear attempt to answer the question?
   - Generally requires a progressive verb
   - For targeted items: subject of question or appropriate pronoun must be response subject
4. **Interpretability (I)**: Does response evoke clear mental image (even if different from PDT)?
   - Any required verb arguments must be present & unambiguous
5. **Grammaticality (G)**: Is response free from errors of spelling & grammar?

Annotators

Two annotators:

- NSs (US English), both with language teaching experience (child & adult learners).
- Annotator 1 (A1): complete corpus
- Annotator 2 (A2): development & test sets, each with 1 intransitive, 1 trans, 1 ditrans

Annotation Results

Annotation Examples

What is the boy doing? (Targeted)	C	V	A	I	G
eating pizza	1	1	1	1	1
eating food.	0	1	1	1	1
eatting.	0	1	1	1	0
The child is eating pizza.	1	1	0	1	1
He may get fat eating pizza.	1	0	0	1	1
The boy is hungry.	0	1	0	0	1
Pizza is this boy's favorite food.	0	0	0	0	1
What is happening? (Untargeted)	C	V	A	I	G
The kid's eating pizza	1	1	1	1	1
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
A youngster anticipates the taste of pizza	1	1	0	1	1
Pepperoni pizza makes the boy smile	0	0	0	1	1
He sure is happy.	0	1	0	1	1

Table 5: Sample responses from development transitive item, with adjudicated annotations

Inter-Annotator Agreement

	Set	Total	A1Yes	A2Yes	AvgYes	Chance	Agree	Kappa
Verb Type	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
Prompt	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
Feature	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 test set

Observations from Table 6

- Average *yes* rates (*AvgYes*) show all features skew toward *yes* annotations
  - Cohen's kappa needed as measure of inter-annotator agreement
- Cohen's kappas well above conventional 0.67 threshold for meaningful agreement

⇒ Annotation scheme can be implemented reliably by following guidelines

- **Verb Type:** Agreement decreases with item complexity (intransitive > trans > ditrans)
- **Prompt:** Agreement slightly higher for untargeted than targeted items
  - Guidelines less complicated for untargeted items
- **Feature:** Answerhood has highest kappa, interpretability has lowest
  - Matches annotator reporting of easiest & hardest features to annotate

Accessing the SAILS Corpus

Download the entire annotated SAILS Corpus, PDTs, & annotation guidelines at:

https://github.com/sailscorpus/sails

SAILS corpus can be used for:

- Language testing & ICALL
- Question answering, dialog systems, pragmatic modeling, visual references

Possibilities for expansion from other researchers:

- New participants, items, approaches for processing

{leviking,md7}@indiana.edu