

CHAPTER 3

DATA COLLECTION

In Chapter 1 I explained that a major motivation of this work is to investigate relatively low-resource mechanisms for content analysis that can help shift the focus of ICALL from form to meaning. In Chapter 2, I examined related work in testing and ICALL. While numerous creative approaches to contextual content analysis are discussed in the literature, the data they rely on is typically not available to other researchers. With these considerations in mind, I decided to collect a corpus of picture description task responses for use in my experiments. This chapter will discuss the data collection task, participants and responses.

3.1 Picture Description Task

The picture description task (PDT) is built around 30 images. Each image is a simple, cartoon-like vector graphic. These images were purchased from Shutterstock, a web-based graphics library¹. In order to constrain response contents to the main action of each image, the images were modified to remove any non-essential detail or background; an example is shown in Figure 3.1. Vector graphics are ideal for this use, because they tend to have an illustrational style with very little detail, as compared to photographs or drawings. Moreover, most consist of layers of graphic objects, and these objects can be easily moved, resized, deleted, combined or otherwise modified to compose the desired stimulus. More example images are presented in Figure 3.2 and the full set is found in Appendix A.

To factor out the influence of previous linguistic context, images are intentionally devoid of any text. In a few cases, symbols are used: two images have music notes; one displays a legible analog clock; one uses numerals in an arithmetic problem and one shows a question mark. The symbols were intended to elicit abstract concepts that are otherwise

¹<https://www.shutterstock.com/vectors>

difficult to portray visually, like TEACHING MATH and ASKING A QUESTION.

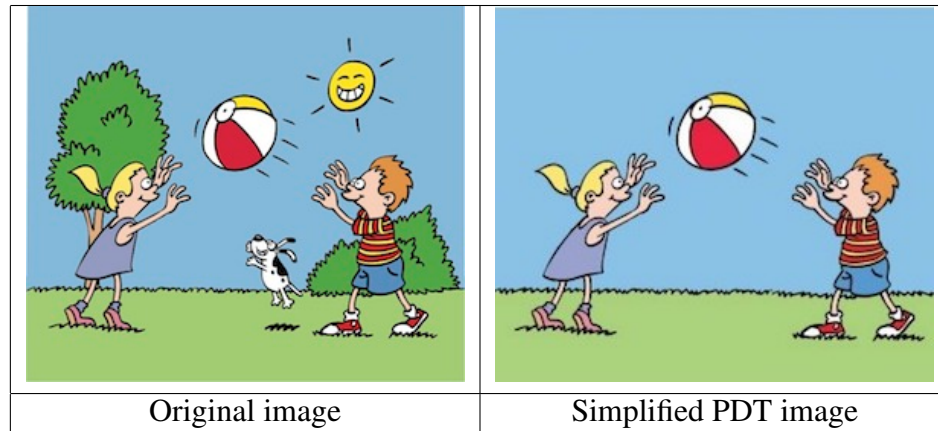


Figure 3.1: All non-essential details were removed from the PDT images in order to focus participants' attention on the main action.

Each image was chosen for its depiction of an ongoing or imminent action (as opposed to a static image) performed by a person or an animal. The images are divided evenly into actions that are canonically intransitive, transitive or ditransitive in English. I chose these three categories because they indicate the number of actors and objects in a given event, and my approach to scoring responses should be able to handle this range of complexity. It should be noted that this categorization is imperfect, however, as some events in the PDT can be expressed in multiple ways, like *The girl is riding a horse* (transitive construction) versus *The girl is horseback riding* (intransitive construction). I attempted to minimize ambiguity (especially between intransitives and transitives) by avoiding images with possible light constructions, like *He is taking a shower* versus *He is showering*.

Each PDT image is used in two different contexts: **targeted** and **untargeted**. An **item** consists of an image and a prompt question. For **targeted** items, questions take the form of *What is <subject> doing?*, with the subject provided (e.g., *the girl*, *the boy*; see Figure 3.2). For all **untargeted** items, the question is *What is happening?* Collecting these targeted and untargeted responses allows for the examination of response variation with and without a subject constraint. To elaborate, for targeted items, I expect less variation among responses;



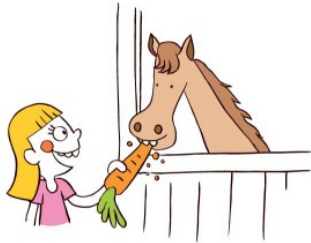
		
What is the girl doing?	What is the boy doing?	What is the girl doing?

Figure 3.2: PDT example images 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.

defining the subject in the prompt means all responses should reuse this subject and only vary in how they express the predicate. For untargeted items, some image prompts might allow for variation of the subject, however. For the image in Figure 3.1, for example, valid responses could include *The boy is throwing a ball to the girl* as well as *The girl is catching a ball from the boy*. Understanding the effect of the subject constraint could help inform approaches to task design and automatic content assessment (Foster and Tavakoli, 2009; Cho et al., 2013).

Different participants performed different versions of the PDT, and multiple versions were necessary to collect roughly equal numbers of targeted and untargeted **responses** for each image. These versions vary in which images are presented as targeted items and which images are presented as untargeted items. Additionally, native speakers (NSs) were asked to provide two non-identical responses to each item (see Section 3.2), but non-native speakers (NNSs) were asked to provide only one response per item, so different PDT versions were used for these groups. The PDTs were hosted online via Survey Monkey², and all participants submitted their responses through this platform.

²<https://www.surveymonkey.com>

In each (full-length) PDT, targeted items are presented in the first half, and untargeted items are presented in the second half. This targeted-untargeted ordering is intended to avoid the possibility that in an untargeted-targeted task, respondents might notice that the question for each untargeted item is always the same in the first half and finish the task hastily without noticing that later targeted items specify the subject. Each half is introduced with instructions, including an example item with sample responses. The instructions ask participants to focus on the main event depicted in the image and for each response to be one complete sentence. Because the PDT was presented as an online survey, all participants typed their responses. Participants were instructed not to use any reference materials, but browser-based spell checking was not disabled, and participants are assumed to have used it as necessary.

The main task instructions are presented in (1). Additional instructions provided to NSs are presented in (2). The full set of PDT versions is available for download with the SAILS Corpus.³

- (1) **Instructions:** In this task, you will view a set of images. For each image, please write **one sentence** to answer the question provided with the image. It is important to write a **complete sentence**, not a word or phrase.
- (2) **Additional Instructions for NSs:** Then, you will be asked to write a second, *different* answer, which is also a complete sentence. This might involve rewording or reorganizing your first sentence. It does not need to be *completely* different; some words may be the same. If you cannot think of another way to answer the question, you may leave the second answer space empty, but any second responses you provide will be greatly appreciated.

³<https://github.com/sailscorpus/sails>

3.2 Participants

This study involved a total of 499 PDT participants. Of these, 141 were NNSs, recruited from intermediate and advanced writing courses in the English Language Improvement Program at Indiana University. These participants performed the task in a computer lab with a researcher present. They 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. Because nearly 90% of these recruits were native speakers of Chinese, care should be taken when drawing conclusions from the corpus; patterns observed among the NNSs here might not apply broadly to all NNSs.

Responses from 329 of the NSs were purchased via Survey Monkey, where survey takers can earn credits that they can redeem for prizes or convert to donations to charities. Asking NSs to provide two responses doubles the length of the survey, exceeding the platform's limits on survey length for purchases responses, so the task was divided into two separate surveys for NSs. Thus while each NS and NNS provided 30 responses, each NNS responded to all 30 PDT items while each NS responded to only 15.

The remaining 29 NS participants were people known to me personally. Due to this relatively small number of participants, their data was not used for modeling or evaluating NNS responses, but it was annotated and is included in the SAILS Corpus. The NS data discussed throughout this dissertation is the crowdsourced data. Where relevant, however, I refer to these two groups as the **Familiar Native Speakers (FNSs)** and the **Crowdsourced Native Speakers (CNSs)**. Future work should include collecting much more FNS data and comparison of the two groups to better understand the differences in quality, as CNSs are almost certainly less likely to perform the task in good faith.

All participants completed a background questionnaire at the beginning of the PDT. This included questions about first and second languages, gender, age, national origin, amount of English language instruction and length of residency in English-speaking loca-

tions. This questionnaire is included as part of the PDT, and the background information provided by participants is included in the SAILS Corpus files. A summary of some of the demographic information is shown in Table 3.1.

	NNS	CNS	FNS
Mean age	18.7	45.0	39.1
Median age	18.0	44.0	35.0
Male	56 (39.7%)	138 (41.9%)	17 (58.8%)
Female	76 (53.2%)	172 (52.3%)	11 (37.9%)
Unknown	9 (6.4%)	19 (5.8%)	1 (3.4%)

Table 3.1: Age and gender information for the three participant groups (Non Native Speakers, Crowdsourced Native Speakers and Familiar Native Speakers).

In previous similar work (King and Dickinson, 2013), NSs were found to produce less variation than NNSs. Many NSs provided identical responses or responses that hew very closely to the most canonical way of expressing the main action. A major motivation for collecting the current corpus was the notion of assessing NNS response content by comparing it against the NS responses. Among other things, this involves the matching of words or syntactic dependencies and thus benefits from a broad set of acceptable responses in the gold standard. For this reason, NSs were asked to provide two non-identical responses, in the hopes that this would result in a wide range of examples of native-like responses for the NNS responses to be compared against.

3.3 Response Totals

A total of 13,533 responses were collected. The response counts for each participant group are presented in Table 3.2. Including the second responses collected from NSs, roughly two thirds of the corpus come from the NS groups. The overwhelming majority of responses appear to be given in good faith, but a small number of responses (primarily from the CNS group) are problematic in this regard, as shown in Table 3.3. These may contain gibberish or obscenities or are otherwise inappropriate for the task. Such responses would

also be expected in an ICALL environment, so they were not removed from the corpus. Instead, these responses were simply annotated like all others (see Chapter 4). Indeed, automatically assigning low scores to inappropriate responses is a central challenge and goal in this project (see Chapter 5).

	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 3.2: 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.

Exemplar: <i>The girl is laughing.</i>
Girl ate a 2x4 and is vomiting toothpicks.
I have to poop so bad.
Exemplar: <i>The boy is eating pizza.</i>
How is the pizza staying perfectly horizontal when the boy is holding it so close to the tip?
see my last statement
Exemplar: <i>The girl is feeding a carrot to a horse.</i>
Creepy clown child grinding her carrot down on poor Ed's beaver teeth
Hoj

Table 3.3: Crowdsourced responses for the items shown in Figure 3.2, showing one exemplar response and two examples of problematic or bad faith responses for each item.

3.4 Response Variation

Type-to-token ratios (TTR) are commonly used as an indication of how varied or homogeneous a set of data is. This number ranges between 0 and 1. In a set of data where most instances or *tokens* are unique (*types*), the number of types per tokens approaches 1. In a set of data where most tokens are identical, the number of types per tokens approaches 0.

With regard to language data, TTRs are often used on the word level, to calculate the lexical density of a document, for example (Granger et al., 2002). In this study, however, type-to-token ratios (TTR) were calculated on the response level for the entire set of items. For this calculation, final punctuation was ignored, and all responses were converted to lowercase. To illustrate, the first three response *tokens* in Table 3.4 would constitute a single response *type*.

Types	Tokens	Response
1	1	The woman is holding a dog
1	2	the woman is holding a dog!
1	3	The Woman is holding a Dog.
2	4	The woman is hugging a puppy.
3	5	The woman squeezed a dog.

Table 3.4: This toy dataset shows how TTR is calculated on the response (sentence) level. Ignoring punctuation and capitalization, the first three response tokens here constitute a single response type. The TTR for this set would be 3:5, or 0.6.

The TTRs for the corpus are presented in Table 3.5. For each cell in this table, the corpus contains 10 items, for each of which there are roughly 150 NS responses and 70 NNS responses. TTR is highly sensitive to text length, so to control for the imbalance between NS and NNS responses, the TTR was calculated for each item based on a random sample of 50 responses (Grieve, 2007). This was repeated 10 times and then averaged to produce a final TTR for each item. These item TTRs were then averaged as intransitives, transitives and ditransitives. 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. The ratios also show that in all cases, as expected, 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, as expected.

Additionally, from Table 3.5 we can see that the NS set contains a greater degree of response variation than does the NNS set. Note that the TTRs here are calculated on *all*

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

responses, and the NS participants each provided two responses per item, whereas the NNS participants were only asked to provide one response per item. This suggests that asking for two responses is an effective way of collecting a broader range of NS responses. This variability can be more closely examined in Table 3.6, which presents separate TTRs for all NS participants’ first responses and second responses. The numbers show that in general, first responses are far less varied than second responses. As we can see, among first responses, variability increases along with item complexity. The pattern holds for targeted second responses, although it is not as pronounced. For untargeted second responses, this monotonic increase in variability is not present, but all three TTRs vary by less than three percent, suggesting that a ceiling effect may be at work.

Finally, for ease of comparison, Table 3.7 presents the (NS only) first response TTRs from Table 3.6 alongside the NNS first (and only) response TTRs from Table 3.5. These comparisons should be made with caution, however, as they cannot account for the possibility of task effects arising from the different instructions given to NS and NNS participants. In other words, it is possible that the anticipation of providing a second response influences a NS participant’s choice of first response, and any such effect would be absent for NNS participants. A future study in which NSs are asked to provide only one response per item could be useful in examining the possibility of such a task effect. As it stands, the table suggests that NNSs generally do exhibit greater response variability than NSs; the only exception to this trend appears among the intransitive untargeted items. This trend is in keeping with the observations from previous work (King and Dickinson, 2013), which

found that NSs tend toward canonical forms, while NNSs use whatever language may be available to them, resulting in greater variation. As described above, this was the motivation for asking NSs for two responses.

	Targeted		Untargeted	
Set	R1	R2	R1	R2
Intrans	0.343	0.819	0.549	0.939
Trans	0.509	0.895	0.682	0.926
Ditrans	0.641	0.948	0.864	0.955

Table 3.6: 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.

	Targeted		Untargeted	
Set	NS	NNS	NS	NNS
Intrans	0.343	0.381	0.549	0.492
Trans	0.509	0.655	0.682	0.779
Ditrans	0.641	0.817	0.864	0.936

Table 3.7: TTRs for complete responses, comparing first responses only.

Having examined response variation in a rather abstract sense here, Chapter 4 will focus on annotating response features to obtain a more fine-grained view of the ways in which responses can vary.