# Shallow Semantic Reasoning from an Incomplete Gold Standard for Learner Language

**Levi King** and **Markus Dickinson**
Indiana University
Bloomington, IN
{leviking,md7}@indiana.edu

## Abstract

We investigate questions of how to reason about learner meaning in cases where the set of correct meanings is never entirely complete, specifically for the case of picture description tasks (PDTs). To operationalize this, we explore different models of representing and scoring non-native speaker (NNS) responses to a picture, including bags of dependencies, automatically determining the relevant parts of an image from a set of native speaker (NS) responses. In more exploratory work, we examine the variability in both NS and NNS responses, and how different system parameters correlate with the variability. In this way, we hope to provide insight for future system development, data collection, and investigations into learner language.

## 1 Introduction and Motivation

Although much current work on analyzing learner language focuses on grammatical error detection and correction (e.g., Leacock et al., 2014), there is a growing body of work covering varying kinds of semantic analysis (e.g., Meurers et al., 2011; Bailey and Meurers, 2008; King and Dickinson, 2014, 2013; Petersen, 2010), including assessment-driven work (e.g., Somasundaran et al., 2015; Somasundaran and Chodorow, 2014). One goal of such work is to facilitate intelligent language tutors (ILTs) and language assessment tools that maximize communicative interaction, as suggested by research in second language instruction (cf. Celce-Murcia, 1991, 2002; Larsen-Freeman, 2002). Whether for feedback or for assessment, however, there are lingering questions about the semantic analysis to address. We investigate questions of how to reason about learner meaning in cases where the set of correct meanings is never entirely complete.

Focusing on semantic analysis requires a sense of what counts as a semantically appropriate utterance from a language learner. Consider when a learner has to describe the contents of a picture (see section 3). There are a number of questions to address in such a situation: 1) Does a semantically correct answer have to sound nativelike or only convey the correct facts? 2) Which facts from the picture are more or less relevant? 3) Are responses strictly correct or not, or is it better to treat correctness as a gradable phenomenon? Additionally, a gold standard of correct responses cannot capture all possible variations of saying the correct content (cf. paraphrases, Barzilay, 2003). We thus must address the specific question of how one can reason about semantic correctness from a (necessarily) incomplete gold standard of answers.

In this paper, we build from our previous work (King and Dickinson, 2013, 2014) and move towards finding a “sweet spot” of semantic analysis (cf. Bailey and Meurers, 2008) for such image-based learner productions. In particular, using available NLP tools, we move away from specific correct semantic representations and an exact definition of correctness, to more abstract data representations and more gradable notions of correctness (section 4). A benefit of more abstract representations is to allow correct and relevant information to be derived from a relatively small set of native speaker responses, as opposed to deriving them by hand, in

*Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 112–121,
San Diego, California, June 16, 2016. ©2016 Association for Computational Linguistics

addition to allowing for a range of sentence types.

We should note, in this context, that we are discussing semantic analysis given a gold standard of native sentences. Image description tasks can often rely on breaking images into semantic primitives (see, e.g., Gilberto Mateos Ortiz et al., 2015, and references therein), but for learner data, we want to ensure that we can account not just for correct semantics (the *what* of a picture), but natural expressions of the semantics (the *how* of expressing the content). In other words, we want to reason about meaning based on specific linguistic forms.

A second issue regarding semantic analysis, beyond correctness, stems from using an incomplete gold standard, namely: assessing the degree of semantic variability, both for native speakers (NSs) and non-native speakers (NNSs). In addition to providing insight into theoretical research on variability in learner language (cf. Ellis (1987), Kanno (1998)), analyzing variability can help determine the best parameters for an NLP system for different kinds of responses. That is, different types of image content might require different mechanisms for processing. Additionally, knowing how different pictures elicit different kinds of content can provide feedback on appropriate types of new data to collect. We approach this issue by clustering responses in various ways (section 5) and seeing how the clusters connect to system parameters.

For both the experiments involving the accuracy of different system parameters (section 4) and the clustering of different responses (section 5), we present results within those sections that show the promise of moving to abstract representations, but in different ways for different kinds of data.

## 2 Related Work

In terms of the overarching goals of developing an interactive ILT, a number of systems exist (e.g., TAGARELA (Amaral et al., 2011), e-Tutor (Heift and Nicholson, 2001)), but few focus on matching semantic forms. *Herr Komissar* (DeSmedt (1995)) is one counter-example; in this game, German learners take on the role of a detective interviewing suspects and witnesses. The system relies largely on a custom-built database of verb classes and related lexical items. Likewise, Petersen (2010) has a system to provide feedback on questions in English, extracting meanings from the Collins parser (Collins, 1999). We also rely on reusing modern NLP software, as opposed to handcrafting a system.

The basic semantic analysis in this paper parallels work on content assessment (e.g., c-rater (Leacock and Chodorow, 2003)). These systems are mostly focused on relatively open-ended short answer scoring, with some systems employing task-based restrictions. As one example, Meurers et al. (2011) evaluate English language learners' short answers to reading comprehension questions, constrained by the topic at hand. Their approach performs multiple levels of annotation, including dependency parsing and lexical analysis from WordNet (Fellbaum, 1998), then aligns elements of the sentence with those of the (similarly annotated) reading prompt, the question, and target answers to determine whether a response is adequate. We explore here a looser notion than alignment for matching NNS responses to a gold standard.

In research closer to our own image-based work, Somasundaran and Chodorow (2014) analyze learner responses to a PDT where the responses were constrained by requiring the use of specific words. The pictures were annotated by experts, and the relevance of responses was calculated through the overlap of the response and annotation contents. Somasundaran et al. (2015) present similar work analyzing responses to sequences of pictures. While they score via a machine learning system, we stick closer to the original forms in trying to find an appropriate way to analyze the data; the notion of overlap for relevance, however, is very similar in spirit to our count-based methods (section 4.2).

We build directly from King and Dickinson (2013, 2014), where the method to obtain a semantic form from a NNS production is: 1) obtain a syntactic dependency representation from the off-the-shelf Stanford Parser (de Marneffe et al., 2006; Klein and Manning, 2003), and 2) obtain a semantic form from the parse, via a small set of hand-written rules. It is this method we attempt to generalize (section 4).

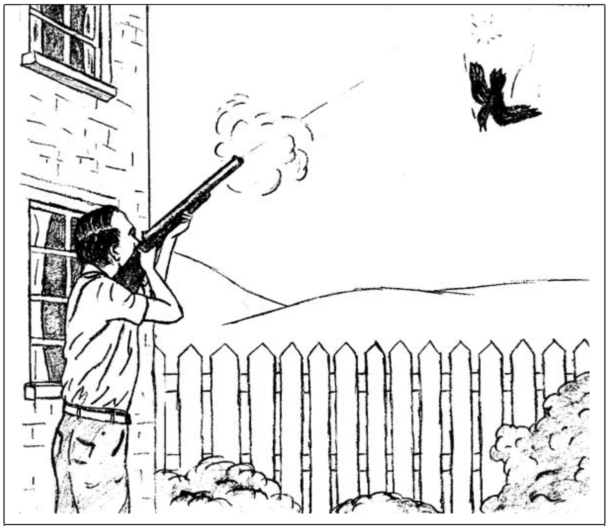
## 3 Data Collection

Because our approach requires both NS and NNS responses and necessitates constraining both the form and content of responses, we previously assembled a small corpus of NS and NNS responses to a PDT (King and Dickinson, 2013). Research in SLA often relies on the ability of task design to induce particular linguistic behavior (Skehan et al., 1998), and the PDT should induce context-focused communicative behavior. Moreover, the use of the PDT as a reliable language research tool is well-established in areas of study ranging from SLA to Alzheimer's disease (Ellis, 2000; Forbes-McKay and Venneri, 2005).

We rely on visual stimuli here for a number of reasons. First, an overarching goal of our work is the development of an ILT that feels like more like a computer game than a grammar drill, and visual stimuli are essential to many games. Secondly, by using images, the information the response should contain is limited to the information contained in the image. Relatedly, particularly simple images should restrict elicited responses to a tight range of expected contents. The current visual stimuli present events that are mainly transitive in nature and likely to elicit responses with an unambiguous subject, verb and object, thereby restricting form in addition to content. Finally, this format allows one to investigate pure interlanguage without the influence of verbal prompts and shows learner language being used to convey meaning and not just manipulate forms.

The PDT consists of 10 items (8 line drawings and 2 photographs[1]) intended to elicit a single sentence each; an example is given in Figure 1. Participants were asked to view the image and describe the action in past or present tense. The data consist of responses from 53 informants (14 NSs, 39 NNSs), for a total of 530 sentences, with the NNSs being intermediate and upper-level adult English learners in an intensive English as a Second Language program. The distribution of first languages (L1s) is: 14 English, 16 Arabic, 7 Chinese, 2 Japanese, 4 Korean, 1 Kurdish, 1 Polish, 2 Portuguese, and 6 Spanish.

Responses were typed by the participants themselves, with spell checking disabled in some cases. Even among the NNSs that used spell checking, a number of spelling errors resulted in real words. To address this, we use a spelling correction tool to obtain candidate spellings for each word, prune the candidates using word lists from the NS responses, recombine candidate spellings into candidate sentences, and evaluate these with a trigram language model (LM) to select the most likely intended response (King and Dickinson, 2014).



| **Response (L1)** |
|---|
| The man killing the beard. (Arabic) |
| A man is shutting a bird. (Chinese) |
| A man is shooting a bird. (English) |
| The man shouted the bird. (Spanish) |

**Figure 1:** Example item and responses

Once the responses had been collected, the NNS responses were annotated for correctness, with respect to the content of the picture. The lead author marked spelling and meaning errors which prevent a complete mapping to correct information (see King and Dickinson, 2013). On the one hand, minor misspellings are counted as incorrect (e.g., *The artiest is drawing a **portret***), while, on the other hand, the annotation does not require distinguishing between between spelling and meaning errors. In the future, we plan on fine-tuning the annotation criteria.

## 4 Generalizing the Methods

The previous work assumed that the assessment of NNS responses involves determining whether the gold standard (GS) contains the same semantic triple that the NNS produced, i.e., whether a *triple* is *covered* or *non-covered*. In such a situation the GS need only be comprised of *types* of semantic triples. But the GS is comprised of the small set of NS responses

[1] We have not observed substantial differences between responses for the drawings and the photographs.

and is thus incomplete—meaning that exact matching is going to miss many cases, and indeed in King and Dickinson (2013), we note that GS coverage is only at 50.8%. Additionally, relying on matching of triples limits the utility of the method to specific semantic requirements, namely transitive sentences. By moving to bags of dependencies and tallying the counts of (NS) responses in the GS, we can move into a gradable, or ranking, approach to NNS responses.

We want to emphasize the degree to which a response conveys the same meaning as the GS, necessitating an approach which can automatically determine the importance of a piece of information in the GS. We break this down into how we represent the information (section 4.1) and then how we compare NNS information to GS information (section 4.2), allowing us to rank responses from least to most similar to the GS.[2] We also discuss the handling of various other system parameters (section 4.3).

### 4.1 Representation

To overcome the limitations of an incomplete GS, we represent each response as a list of *terms* taken from the dependency parse (de Marneffe et al., 2006), the terms referring to individual dependencies (i.e., relations between words). This eliminates the complications of extracting semantic triples from dependency parses, which could only handle a very restricted set of grammatical forms and resulted in errors in 7–8% of cases (King and Dickinson, 2013). Operating directly on individual dependencies from the overall tree also means the system can allow for "partial credit"; it distributes the matching over smaller, overlapping pieces of information rather than a single, highly specific triple.

Specifically, representations take one of five forms. We first tokenize and lemmatize the response to a list of lemmas that represents the response. The five term representations are then variations on dependencies. The full form concatenates the label, head and dependent, as in `subj#boy#kick`. We call this **ldh** (label, dependent, head). The remaining four forms abstract over either the label, head and/or dependent, as in `X#boy#kick`. We refer to these forms as **xdh**, **lxh**, **ldx**, and **xdx**. The `xdx` model is on a par with treating the sentence as a bag of lemmas, except that some function words not receiving parses (e.g., prepositions) are not included (see King and Dickinson, 2013). In our current experiments, we test each of these term representations separately, but we expect to ultimately make use of some weighted combination. Future representations may also incorporate WordNet relations or semantic role labeler output.

### 4.2 Scoring Responses

Taking the term representations from the previous section, the next task is to combine them in a way which ranks responses from least to most appropriate. Responses are scored with one of four approaches, using one of two methods to **weight** response terms combined with one of two methods to **compare** the weighted NNS terms with the GS.

For weighting, we use either a simple frequency measure (**F**) or one based on **tf-idf** (**T**) (Manning et al., 2008, ch. 6). We explore tf-idf as a measure of a term's importance with the hope that it is able to reduce the impact of semantically unimportant terms—e.g., determiners like *the*, frequent in the GS, but unimportant for evaluating the semantic contents of NNS responses—and to upweight terms which may be salient but infrequent, e.g., only used in a handful of GS sentences. For example, for an item depicting a man shooting a bird (see Table 1 and Figure 1), of 14 GS responses, 12 described the subject as *man*, one as *he* and one as *hunter*. Since *hunter* is infrequent in English, even one instance in the GS should get upweighted via tf-idf, and indeed it does. This is valuable, as numerous NNS responses use *hunter*.

Calculating tf-idf relies on both *term frequency* ($tf$) and *inverse document frequency* ($idf$). Term frequency is simply the raw count of an item, and for tf-idf of terms in the GS, we take this as the frequency within the GS. Inverse document frequency is derived from some reference corpus, and it is based on the notion that appearing in more documents makes a term less informative with respect to distinguishing between documents. The formula is in (1) for a term $t$, where $N$ is the number of documents in the reference corpus, and $df_t$ is the number of documents featuring the term ($idf_t = \log \frac{N}{df_t}$).

[2] Although rankings often go from highest to lowest, we prioritize identifying problematic cases, so we rank accordingly.

A term appearing in fewer documents will thus obtain a higher $idf$ weight, and this should readjust frequencies based on semantic importance.

(1) $tfidf(t) = tf_{GS} \log \frac{N}{df_t}$

After counting/weighting, the frequencies are then either **averaged** to yield a response score (**A**), or NNS term weights and GS term weights are treated as vectors and the response score is the **cosine distance** (**C**) between them. This yields:

**Frequency Average (FA).** Within the GS, the frequency of each term is calculated. Each term in the NNS response is then given a score equal to its frequency in the GS; terms missing from the GS are scored zero. The response score is the average of the term scores, with higher scores closer to the GS.

**Tf-idf Average (TA).** This involves the exact same averaging as with model FA, but now the terms in the GS are assigned tf-idf weights before averaging.

**Frequency Cosine (FC).** The frequency of each term is calculated within the GS and (separately) within the NNS response. The term scores are then treated as vectors, and the response score is the cosine distance between them, with lower scores being closer to the GS.

**Tf-idf Cosine (TC).** This involves the exact same distance comparison as with model FC, but now the terms of both the GS and NNS responses are assigned tf-idf weights before comparison.

### 4.3 System Parameters

In addition to the four approaches, we have term representations and two sets of parameters, listed below, to vary, resulting in a total of 60 settings for processing responses (see also Table 2).

**Term form.** As discussed in section 4.1, the terms can take one of five representations: `ldh`, `xdh`, `lxh`, `ldx`, or `xdx`.

**Scoring approach.** As discussed in section 4.2, the NNS responses can be compared with the GS via models `FA`, `TA`, `FC`, or `TC`.

**Reference corpus.** The reference corpus for deriving tf-idf scores can be either the Brown Corpus (Kucera and Francis, 1967) or the Wall Street Journal (WSJ) Corpus (Marcus et al., 1993). These are abbreviated as `B` and `W` in the results below; `na` indicates the lack of a reference corpus, as this is only relevant to approaches `TA` and `TC`. The corpora are divided into as many documents as originally distributed (`W`: 1640, `B`: 499). The WSJ is larger, but Brown has the benefit of containing more balance in its genres (vs. newstext only). Considering the narrative nature of PDT responses, a reference corpus of narrative texts would be ideal, but we choose manually parsed reference corpora as they are more reliable than automatically parsed data.

**NNS source.** Each response has an original version (`NNSO`) and the output of a language model spelling corrector (`NNSLM`) (see section 3).

### 4.4 Results

#### 4.4.1 Evaluation metrics

We ran 60 response experiments, each with different system settings (section 4.3). Within each experiment, we rank the 39 scored NNS responses from least to most similar to the GS. For assessing these settings themselves, we rely on past annotation, which counted unacceptable responses as errors (see section 3).[3] As the lowest rank indicates the greatest distance from the GS, a good system setting should ideally position the unacceptable responses among those with the lowest rankings. Thus, we assign each error-containing response a score equal to its rank, or, if necessary, the average rank of responses sharing the same score.

In Table 1, an excerpt of sentence responses is shown for one item, ranked from lowest to highest. To take one example, the third-ranked sentence, *the man is hurting duck*, has a score of 0.996, and it is annotated as an error (1 in the *E* column). Thus, the evaluation metric adds a score of 3 to the overall sum. The sentence ranked 18, by contrast, is not an error, and so nothing is added. In the case of the top rank, two responses with errors are tied, covering rank 1 and 2, so each adds a score of 1.5.

[3] The source of the error is also labeled—stemming from NNS unintelligibility or a system error (from spelling correction, parsing, or some downstream component)—but we do not currently use this annotation.

| *R* | *S* | Sentence | *E* | *V* |
|---|---|---|---|---|
| 1 | 1.000 | she is hurting. | 1 | 1.5 |
|  | 1.000 | man mull bird | 1 | 1.5 |
| 3 | 0.996 | the man is hurting duck. | 1 | 3.0 |
| 4 | 0.990 | he is hurting the bird. | 1 | 3.0 |
| 11 | 0.865 | the man is trying to hurt a bird | 1 | 11.0 |
| 12 | 0.856 | a man hunted a bird. | 0 | 0.0 |
| 17 | 0.775 | the bird not shot dead. | 1 | 17.0 |
| 18 | 0.706 | he shot at the bird | 0 | 0.0 |
| 19 | 0.669 | a bird is shot by a un | 1 | 19.0 |
| 20 | 0.646 | the old man shooting the birds | 0 | 0.0 |
| 37 | 0.086 | the old man shot a bird. | 0 | 0.0 |
| 38 | 0.084 | a old man shot a bird. | 0 | 0.0 |
| 39 | 0.058 | a man shot a bird | 0 | 0.0 |
| Total (Raw) |  |  | 17 | 169 |
| Average Precision |  |  | 0.75084 |  |

**Table 1:** Rankings for Item 10 from the best system setting (TC_B_NNSLM_ldh) based on average precision scores. *R*: rank; *S*: sentence score; *E*: error; *V*: rank value.

The sum of these scores is taken as the **Raw** metric for that experimental setting. In many cases, one version of a response (NNSO or NNSLM) contains an error, but the other version does not. Thus, for example, an NNSO experiment may result in a higher error count than the NNSLM equivalent, and in turn a higher Raw score. In this sense, Raw scores emphasize error reduction and incorporate item difficulty.

However, it is possible that the NNSO experiment, even with its higher error count and Raw score, does a better job ranking the responses in a way that separates good and erroneous ones. To account for this, we also use **(mean) average precision ((M)AP)** (Manning et al., 2008, ch. 8), which emphasizes discriminatory power.

For average precision (AP), one calculates the precision of error detection at every point in the ranking, lowest to highest. In Table 1, for example, the precision for the first cut-off (1.000) is 1.0, as two responses have been identified, and both are errors ($\frac{2}{2}$). At the 11th- and 12-ranked response, precision is 1.0 ($\frac{11}{11}$) and 0.917 (=$\frac{11}{12}$), respectively, precision dropping when the item is not an error. AP averages over the precisions for all $m$ responses ($m = 39$ for our NNS data), as shown in (2), with each response notated as $R_k$. Averaging over all 10 items results in the Mean AP (MAP).

$$(2) \quad AP(item) = \tfrac{1}{m} \sum_{k=1}^{m} Precision(R_k)$$

As mentioned, the Raw metric emphasizes error reduction, as it reflects not just performance on identifying errors, but also the effect of the overall number of errors. In this way, it may be useful for predicting future system performance, an issue we explore in the evaluation of clustering items (section 5.3). MAP, on the other hand, emphasizes finding the optimal separation between errors and non-errors and is thus more of the focus in the evaluation of the best system parameters next.

### 4.4.2 Best system parameters

To start the search for the best system parameters, it may help to continue our single example, in Table 1. The best setting, as determined by the Normalized metric, uses the tf-idf cosine (TC) approach with the Brown Corpus (B), the spelling corrected response (NNSLM), and the full form of the dependencies (ldh). It ranks highest because errors are well separated from non-errors; the highest ranked of 17 total errors is at rank 19. Digging a bit deeper, we can see in this example how the verb *shoot* is common in all the highest-ranked cases shown (#37–39), but absent from all the lowest, showing both the effect of the GS (as all NSs used *shoot* to describe the action) and the potential importance of even simple representations like lemmas. In this case, the ldh representation is best, likely because the word *shoot* is not only important by itself, but also in terms of which words it relates to, and how it relates (e.g., dobj#bird#shoot).

Table 3 shows the five best and five worst system settings averaged across all 10 PDT items, as ranked by MAP. Among the trends that pop out is a favoritism towards NNSLM models (i.e., spelling correction). This is due to the fact that higher numbers of errors inflate the MAP scores, and somewhat counterintuitively, the spelling correction module introduces more errors than it corrects, meaning there are more errors present overall in the NNSLM responses than in the NNSO responses.[4]

[4] Note that among the remaining parameter classes, variation does not effect the number of errors.

| Approach | | Term Form | | Ref. Corpus (TA/TC) | | NNS Source | |
|---|---|---|---|---|---|---|---|
| 0.51577 | TC | xdh | 0.51810 | Brown | 0.51534 | NNSLM | 0.51937 |
| 0.50780 | FC | ldh | 0.51677 | WSJ | 0.50798 | NNSO | 0.49699 |
| 0.50755 | TA | lxh | 0.51350 | | | | |
| 0.49464 | FA | xdx | 0.49901 | | | | |
| | | ldx | 0.49352 | | | | |

**Table 2:** Approaches and parameters ranked by mean average precision for all 10 PDT items.

Another feature among the best settings is the inclusion of heads in the dependency representations. In fact, the top 17 ranked settings all include heads (lxh, xdh, ldh); xdx first enters the rankings at 18, and xdx and ldx are common among the worst performers. This is likely due to the salience of the verbs in these transitive sentences; they constitute the heads of the subjects and objects, in relatively short sentences with few dependencies. Furthermore, the tf-idf weighted models dominate the rankings, especially TC. It is also clear that for our data tf-idf works best with the Brown Corpus (B).

| Rank | MAP | Settings |
|---|---|---|
| 1 | 0.5534 | TC_B_NNSLM_lxh |
| 2 | 0.5445 | TA_B_NNSLM_lxh |
| 3 | 0.5435 | TC_W_NNSLM_lxh |
| 4 | 0.5422 | TC_B_NNSLM_xdh |
| 5 | 0.5368 | TC_B_NNSLM_ldh |
| 56 | 0.4816 | TA_B_NNSO_xdx |
| 57 | 0.4796 | FA_na_NNSLM_ldx |
| 58 | 0.4769 | FC_na_NNSO_lxh |
| 59 | 0.4721 | TA_W_NNSO_xdx |
| 60 | 0.4530 | FA_na_NNSO_lxh |

**Table 3:** Based on Mean Average Precision, the five best and five worst settings across all 10 PDT items.

We also summarize the rankings for the individual parameter classes, presented in Table 2, confirming the trends in Table 3. For a given parameter, e.g., ldh, we averaged the experiment scores from all settings including ldh across all 10 items. Notably, TC outperforms the other models, with FC and TA close behind (and nearly tied). Performance falls for the simplest model, FA, which was in fact intended as a baseline. With TC>FC and TA>FA, tf-idf weighting seems preferable to basic frequencies.

Again, the importance of including heads in dependencies is apparent here; the three dependency representations containing heads constitute the top three, with a sizable drop in performance for the remaining two forms (xdx and ldx). Moreover, given the content and narrative style of the PDT responses, it is unsurprising that the Brown Corpus serves as a better reference corpus than the WSJ Corpus for tf-idf. Finally, the NNSLM source significantly outperforms the NNSO source.

Despite the strength of these overall trends, variability does exist among the best settings for different items, a point obscured in the averages. In Tables 4 and 5, we present the best and worst ranked settings for two of the least similar items, 1 and 5. Their dissimilarity can be seen at a glance, simply from the range of the AP scores (0.05–0.31 for item 1 vs. 0.52–0.81 for item 5), which in itself reflects a differing number of erroneous responses (2 [NNSO] or 6 [NNSLM] for item 1 vs. 23 or 24 for item 5).

For item 1, a drawing of a boy kicking a ball, we see considerable variability in the best approach just within the top five settings: all four approaches are in the top five. Contrary to the overall trends, we also see the ldx form—without any head information—in the two best settings. Note also that, even though tf-idf weighting (TA/TC) is among the best settings, it is consistently the worst setting, too.

For item 5 in Table 5, a drawing of a man raking leaves, the most noticeable difference is that of xdx being among three of the top five settings. We believe that part of the reason for the superior performance of xdx (cf. lemmas), is that for this item, all the NSs use the verb *rake*, while none of the NNSs use this word. For item 1 (the boy kicking a ball), there is lexical variation for both NSs and NNSs.

These types of differences—for these items and others—lead us to explore the clustering of item patterns, in order to leverage these differences and auto-

| Rank | AP | Settings |
|---|---|---|
| 1 | 0.30997 | TC_B_NNSLM_ldx |
| 2 | 0.30466 | TA_B_NNSLM_ldx |
| 3 | 0.30015 | TA_B_NNSLM_xdh |
| 4 | 0.29704 | FC_na_NNSLM_xdh |
| 5 | 0.29650 | FA_na_NNSLM_ldh |
| 56 | 0.06474 | TC_B_NNSO_ldx |
| 57 | 0.06174 | TC_W_NNSO_ldx |
| 58 | 0.06102 | TA_W_NNSO_lxh |
| 59 | 0.05603 | TA_W_NNSO_xdx |
| 60 | 0.05094 | TA_W_NNSO_ldx |

**Table 4:** Based on Average Precision, the five best and five worst settings for item 1.

| Rank | AP | Settings |
|---|---|---|
| 1 | 0.80965 | FA_na_NNSLM_xdx |
| 2 | 0.80720 | TA_B_NNSLM_lxh |
| 3 | 0.80473 | TC_B_NNSLM_lxh |
| 4 | 0.79438 | TC_B_NNSLM_xdx |
| 5 | 0.78108 | TC_W_NNSLM_xdx |
| 56 | 0.56495 | FC_na_NNSO_xdh |
| 57 | 0.56414 | TC_B_NNSO_lxh |
| 58 | 0.55890 | TC_W_NNSO_lxh |
| 59 | 0.54506 | FC_na_NNSO_lxh |
| 60 | 0.52013 | FA_na_NNSO_lxh |

**Table 5:** Based on Average Precision, the five best and five worst settings for item 5.

matically choose the optimal settings for new items; we turn to this next.

## 5 Clustering

Given the variability of NS and NNS responses, and the possible correlation with different system parameters, we have begun exploring connections by clustering the different items. The clustering uses, for one set, *response features*, i.e., features observable from the responses, and, separately, *performance features*, i.e., the performance of different system settings on the responses. Although the work is very exploratory, our goal is to get a handle on learner variability for different items and explore correlations between response and performance clusters.

### 5.1 Response Clustering

We cluster the 10 PDT items using simple features taken from the responses themselves. Specifically, we use various combinations of type counts, token counts, and type-to-token ratios for each term form (ldh, xdh, lxh, ldx, xdx), taken from each response source (GS, NNSO, NNSLM).

### 5.2 Performance Clustering

From the system output, we cluster items using per-item Raw scores for various settings. That is, for each of the 10 items, we calculate an average error score for each approach (FA, TA, FC, TC), each term form (ldh, xdh, lxh, ldx, xdx), each reference corpus (B, W), and each response source (NNSO, NNSLM). As mentioned in section 4.4.1, Raw scores should account for the number of errors produced by NNSs for each item, which should correlate with future system performance.

### 5.3 Results

Although there is noise in some experiments, some patterns do seem to emerge in many of the clusterings; we present some of the most common patterns here. Figure 2 shows a clustering based on response features that shares some characteristics with Figure 3, a clustering based on performance features. (Note that clustering heights are not to scale.) In both examples, items 5 and 9 form a cluster attaching to the root. These are described in the GS as *A man is raking leaves* and *Two boys are rowing a boat*. These were also the two most difficult items for NNSs. While other items involved common verbs like *kick*, *paint* and *cut*, the actions depicted in these items were more specific and required words outside the vocabulary of many participants. For example, while all 14 NSs used either *row* or *paddle*, only five of 39 NNSs used these verbs; the rest used verbs like *boat*, *sail*, *sit*, *play* or *ride*.

Items 1 and 4 also appear as a cluster in both cases. In GS examples, these are described as *The boy is kicking a ball* and *A man is reading a newspaper*. The images portray actions that language learners often learn in beginner courses, and in fact, these were the easiest items for NNSs. The simple actions and objects mean that both token counts and type counts are relatively low. With regard to fea-

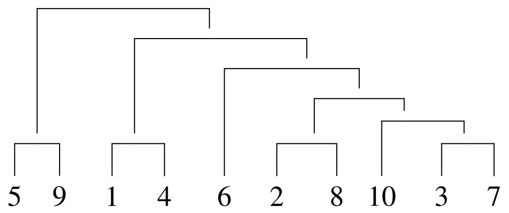


**Figure 2:** PDT items clustered by type and token counts of all NS, NNSO and NNSLM responses.

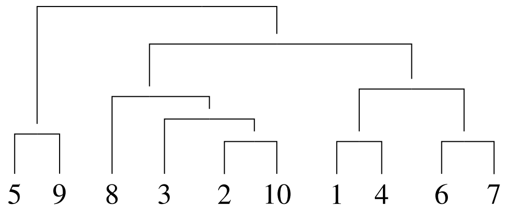


**Figure 3:** PDT items clustered by parameter performance.

ture performance, for both items the same parameters perform highly (`TC`/`TA`, `ldx`/`ldh`/`xdh`), suggesting that a future item which clusters with these two would benefit from the same processing.

## 6 Summary and Outlook

We have investigated ways to reason about learner meaning in cases where the set of correct meanings is incomplete, namely in the case of picture description tasks (PDTs). Specifically, we have explored different models of representing and scoring NNS responses to a picture—different linguistic representations, term weighting schemes, reference corpora, and use of spelling correction—in order to automatically determine the relevant parts of an image from a set of NS responses. In more exploratory work, we have also examined the variability in both NS and NNS responses, and how different system parameters correlate with the variability.

Already, the results are providing insight for future system development, data collection, and investigations into learner language. A big next step of the work is to collect more data, and examining the variability in the NS/NNS data has provided feedback on the types of new data to gather, to better ensure a wide range of behavior from NNSs. Getting a range of items, with different sentence types and variability in responses, will help us properly find our envisioned sweet spot of semantic analysis. In that vein, we plan on exploring more parameters (e.g., semantic role information) and holding out data to better gauge the impact of clustering a new item with the existing items and selecting the processing parameters on that basis. Beyond that loom large questions about how to annotate gradability in learner responses and how to map system processing to accurate semantic feedback.

## Acknowledgments

We would like to thank the members of the CLingDing discussion group and the three anonymous reviewers for their helpful feedback.

## References

Luiz Amaral, Detmar Meurers, and Ramon Ziai. 2011. Analyzing learner language: Towards a flexible NLP architecture for intelligent language tutors. *Computer Assisted Language Learning* 24(1):1–16.

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)*. Columbus, OH, pages 107–115.

Regina Barzilay. 2003. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University.

Marianne Celce-Murcia. 1991. Grammar pedagogy in second and foreign language teaching. *TESOL Quarterly* 25:459–480.

Marianne Celce-Murcia. 2002. Why it makes sense to teach grammar through context and through discourse. In Eli Hinkel and Sandra Fotos, editors, *New perspectives on grammar teaching in second language classrooms*, Lawrence Erlbaum, Mahwah, NJ, pages 119–134.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*. Genoa, Italy.

William DeSmedt. 1995. Herr Kommissar: An ICALL conversation simulator for intermediate German. In V. Melissa Holland, Jonathan Kaplan, and Michelle Sams, editors, *Intelligent Language Tutors: Theory Shaping Technology*, Lawrence Erlbaum, Mahwah, NJ, pages 153–174.

Rod Ellis. 1987. Interlanguage variability in narrative discourse: Style shifting in the use of the past tense. *Studies in Second Language Acquisition* 9:1–19.

Rod Ellis. 2000. Task-based research and language pedagogy. *Language Teaching Research* 4(3):193–220.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Katrina Forbes-McKay and Annalena Venneri. 2005. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological Sciences* 26(4):243–254.

Luis Gilberto Mateos Ortiz, Clemens Wolff, and Mirella Lapata. 2015. Learning to interpret and describe abstract scenes. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 1505–1515.

Trude Heift and Devlan Nicholson. 2001. Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education* 12(4):310–325.

Kazue Kanno. 1998. Consistency and variation in second language acquisition. *Second Language Research* 14(4):376–388.

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*. Atlanta, Georgia, pages 11–21.

Levi King and Markus Dickinson. 2014. Leveraging known semantics for spelling correction. In *Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning*. Uppsala, Sweden, pages 43–58.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL-03*. Sapporo, Japan.

Henry Kucera and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Diane Larsen-Freeman. 2002. Teaching grammar. In Diane Celce-Murcia, editor, *Teaching English as a second or foreign language*, Heinle & Heinle, Boston, pages 251–266. 3rd edition.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and Humanities* pages 389–405.

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

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. CUP.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics* 19(2):313–330.

Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey Bailey. 2011. Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *Special Issue on Free-text Automatic Evaluation. International Journal of Continuing Engineering Education and Life-Long Learning (IJCEELL)* 21(4):355–369.

Kenneth A. Petersen. 2010. *Implicit Corrective Feedback in Computer-Guided Interaction: Does Mode Matter?*. Ph.D. thesis, Georgetown University, Washington, DC.

Peter Skehan, Pauline Foster, and Uta Mehnert. 1998. Assessing and using tasks. In Willy Renandya and George Jacobs, editors, *Learners and language learning*, Seameo Regional Language Centre, pages 227–248.

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*. Baltimore, Maryland, pages 1–11.

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*. Association for Computational Linguistics, Denver, Colorado, pages 42–48.