Shallow Semantic Analysis of Interactive Learner Sentences

Levi King
Indiana University
Bloomington, IN USA
leviking@indiana.edu

Markus Dickinson
Indiana University
Bloomington, IN USA
md7@indiana.edu

Abstract

Focusing on applications for analyzing learner language which evaluate semantic appropriateness and accuracy, we collect data from a task which models some aspects of interaction, namely a picture description task (PDT). We parse responses to the PDT into dependency graphs with an an off-the-shelf parser, then use a decision tree to classify sentences into syntactic types and extract the logical subject, verb, and object, finding 92% accuracy in such extraction. The specific goal in this paper is to examine the challenges involved in extracting these simple semantic representations from interactive learner sentences.

1 Motivation

While there is much current work on analyzing learner language, it usually focuses on grammatical error detection and correction (e.g., Dale et al., 2012) and less on semantic analysis. At the same time, Intelligent Computer-Assisted Language Learning (ICALL) and Intelligent Language Tutoring (ILT) systems (e.g., Heift and Schulze, 2007; Meurers, 2012) also tend to focus more on grammatical feedback. An exception to this rule is *Herr Komissar*, an ILT for German learners that includes rather robust content analysis and sentence generation (DeSmedt, 1995), but this involves a great deal of hand-built tools and does not connect to modern NLP. Some work addresses content assessment for short answer tasks (Meurers et al., 2011), but this is still far from naturalistic, more conversational interactions (though, see Petersen, 2010).

Our overarching goal is to facilitate ILTs and language assessment tools that maximize free interaction, building as much as possible from existing NLP resources. While that goal is in the distant future, the more immediate goal in this paper is to pinpoint the precise challenges which interactive learner sentences present to constructing semantic analyses, even when greatly constrained. We approximate this by collecting data from a task which models some aspects of interaction, namely a picture description task (PDT), parsing it with an off-the-shelf parser, extracting semantic forms, and noting the challenges throughout.

The focus towards interaction is in accord with contemporary theory and research in Second Language Acquisition (SLA) and best practices in second language instruction, which emphasize the limiting of explicit grammar instruction and feedback in favor of an approach that subtly integrates the teaching of form with conversation and task-based learning (Celce-Murcia, 1991, 2002; Larsen-Freeman, 2002). Indeed, Ellis (2006) states, "a traditional approach to teaching grammar based on explicit explanations and drill-like practice is unlikely to result in the acquisition of the implicit knowledge needed for fluent and accurate communication." For our purposes, this means shifting the primary task of an ICALL application from analyzing grammar to evaluating semantic appropriateness and accuracy.

The data for error detection work is ideal for developing systems which provide feedback on essays, but not necessarily for more interactive communication. Thus, our first step is to collect data similar to what we envision processing in something like an

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 11–21,
Atlanta, Georgia, June 13 2013.

ILT game, data which—as far as we know—does not exist. While we desire relatively free production, there are still constraints; for games, for example, this comes in the form of contextual knowledge (pictures, rules, previous interactions). To get a handle on variability under a set of known constraints and to systematically monitor deviations from target meanings, we select a PDT as a constrained task that still promotes interactive communication. Collecting and analyzing this data is our first major contribution, as described in section 3.

Once we have the data, we can begin to extract semantic forms, and our second major contribution is to outline successes and pitfalls in obtaining shallow semantic forms in interactive learner data, as described in section 4, working from existing tools. Although we observe a lot of grammatical variation, we will demonstrate in section 5 how careful selection of output representations (e.g., the treatment of prepositions) from an off-the-shelf parser and a handful of syntax-to-semantics rules allow us to derive accurate semantic forms for most types of transitive verb constructions in our data. At the same time, we will discuss the difficulties in defining a true gold standard of meanings for such a task. This work paves the way for increasing the range of constructions and further exploring the space between free and constrained productions (see also the discussion in Amaral and Meurers, 2011).

2 Related Work

In terms of our 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, learners take on the role of a detective tasked with interviewing suspects and witnesses. The system relies largely on a custom-built database of verb classes and related lexical items. Likewise, Petersen (2010) designed a system to provide feedback on questions in English, extracting meanings from the Collins parser (Collins, 1999). Our work is is in the spirit of his, though our starting point is to collect data of the type of task we aim to analyze, thereby pinpointing how one should begin to build a system.

The basic semantic analysis in this paper parallels work on content assessment (e.g., ETS's c-rater system (Leacock and Chodorow, 2003)). Different from our task, these systems are mostly focused on essay and short answer scoring, though many focus on semantic analysis under restricted conditions. 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 on the reading prompt, including dependency parsing and lexical analysis from WordNet (Fellbaum, 1998), then attempts to align elements of the sentence with those of the (similarly annotated) reading prompt, the question, and target answers to determine whether a response is adequate or what it might be missing. Our scenario is based on images, not text, but our future processing will most likely need to include similar elements, e.g., determining lexical relations from WordNet.

3 Data Collection

The data involved in this study shares much in common with other investigations into semantic analysis of descriptions of images and video, such as the Microsoft Research Video Description Corpus (MSRvid; Chen and Dolan (2011)) and the SemEval-2012 Semantic Textual Similarity (STS) task utilizing MSRvid as training data for assigning similarity scores to pairs of sentences (Agirre et al., 2012). However, because our approach requires both native speaker (NS) and non-native speaker (NNS) responses and necessitates constraining both the form and content of responses, we assembled our own small corpus of NS and NNS responses to a PDT. 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 more interactive 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).

The NNSs were intermediate and upper-level adult English learners in an intensive English as a Second Language program at Indiana University. We rely on visual stimuli here for a number of rea-

sons. Firstly, computer games tend to be highly visual, so collecting responses to visual prompts is in keeping with the nature of our desired ILT. 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. For this initial experiment, we chose or developed each of the visual stimuli because it presents an event that we believe to be 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 us to investigate pure interlanguage without the influence of verbal prompts and shows learner language in a functional context, modeling real language use.

Response (L1)
He is droning his wife pitcher. (Arabic)
The artist is drawing a pretty women. (Chinese)
The artist is painting a portrait of a lady. (English)
The painter is painting a woman's paint. (Spanish)

Figure 1: Example item and responses

The PDT consists of 10 items (8 line drawings and 2 photographs) 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, and care was taken to explain to participants that either past or present tense (and simple or progressive aspect) was acceptable. Responses were typed by the participants themselves (without automatic spell checking). To date, we have collected responses from 53 informants (14 NSs, 39 NNSs), for a total of 530 sentences. The distribution of first languages (L1s) is as follows: 14 English, 16 Arabic, 7 Chinese, 2 Japanese, 4 Korean, 1 Kurdish, 1 Polish, 2 Portuguese, and 6 Spanish.

4 Method

We parse a sentence into a dependency representation (section 4.1) and then extract a simple semantic form from this parse (section 4.2), to compare to gold standard semantic forms.

4.1 Obtaining a syntactic form

We start analysis with a dependency parse. Because dependency parsing focuses on labeling dependency relations, rather than constituents or phrase structure, it easily finds the subject, verb and object of a sentence, which can then map to a semantic form (Kübler et al., 2009). Our approach must eventually account for other relations, such as negation and adverbial modification, but at this point, since we focus on transitive verbs, we take an naïve approach in which subject, verb and object are considered sufficient for deciding whether or not a response accurately describes the visual prompt.

We use the Stanford Parser for this task, trained on the Penn Treebank (de Marneffe et al., 2006; Klein and Manning, 2003).[1] Using the parser's options, we set the output to be Stanford typed dependencies, a set of labels for dependency relations. The Stanford parser has a variety of options to choose from for the specific parser ouput, e.g., how one wishes to treat prepositions (de Marneffe and Manning, 2012). We use the `CCPropagatedDependencies` / `CCprocessed` option to accomplish two things:[2] 1) omit prepositions and conjunctions from the sentence text and instead add the word to the dependency label between content words; and 2) propagate relations across any conjunctions. These decisions are important to consider for any semantically-informed processing of learner language.

[1] `http://nlp.stanford.edu/software/lex-parser.shtml`

[2] `http://nlp.stanford.edu/software/dependencies_manual.pdf`

To see the impetus for removing prepositions, consider the learner response (1), where the preposition *with* is relatively unimportant to collecting the meaning. Additionally, learners often omit, insert, or otherwise use the wrong preposition (Chodorow et al., 2007). The default parser would present a `prep` relation between *played* and *with*, obscuring what the object is; with the options set as above, however, the dependency representation folds the preposition into the label (`prep_with`), instead of keeping it in the parsed string, as shown in Figure 2.

(1) The boy played with a ball.

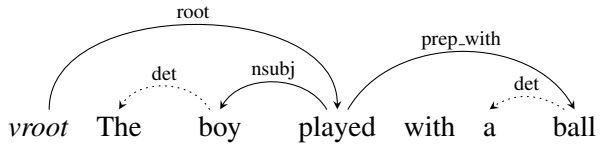

Figure 2: The dependency parse of (1)

This is a very lenient approach to prepositions, as prepositions certainly carry semantic meaning—e.g., *the boy played in a ball* means something quite different than what (1) means. However, because we ultimately compare the meaning to an expected semantic form (e.g., *play(boy,ball)*), it is easier to give the benefit of the doubt. In the future, one may want to consider using a semantic role labeler (e.g., SENNA (Collobert et al., 2011)).

As for propagating relations across conjunctions, this ensures that each main verb connects to its arguments, as needed for a semantic form. For example, in (2), the default parser returns the relation between the first verb of the conjunction structure, *setting* and its subject, *man*, but not between *reading* and *man*. The options we select, however, return an `nsubj` relation between *setting* and *man* and also between *reading* and *man* (similarly for the object, *paper*).

(2) The man is setting and reading the paper.

In addition to these options, many dependency relations are irrelevant for the next step of obtaining a semantic form. For example, we can essentially ignore determiner (`det`) relations between a noun and its determiner, allowing for variability in how a learner produces or does not produce determiners.

4.2 Obtaining a semantic form

4.2.1 Sentence types

We categorized the sentences in the corpus into 12 types, shown in Table 1. We established these types because each type corresponds to a basic sentence structure and thus has consistent syntactic features, leading to predictable patterns in the dependency parses. We discuss the distribution of sentence types in section 5.1.

4.2.2 Rules for sentence types

A sentence type indicates that the logical (i.e., semantic) subject, verb, and object can be found in a particular place in the parse, e.g., under a particular dependency label. For example, for simple transitive sentences of type A, the words labeled `nsubj`, `root`, and `dobj` exactly pinpoint the information we require. Thus, the patterns for extracting semantic information—in the form of *verb(subj,obj)* triples—reference particular Stanford typed dependency labels, part-of-speech (POS) tags, and interactions with word indices.

More complicated sentences or those containing common learner errors (e.g., omission of the copula *be*) require slightly more complicated extraction rules, but, since we examine only transitive verbs at this juncture, these still boil down to identifying the sentence type and extracting the appropriate triple. We do this by arranging a small set of binary features into a decision tree to determine the sentence type, as shown in Figure 3.

To illustrate this process, consider (3). We pass this sentence through the parser to obtain the dependency parse shown in Figure 4. The parsed sentence then moves to the decision tree shown in Figure 3. At the top of the tree, the sentence is checked for an `expl` (expletive) label; having none, it moves rightward to the `nsubjpass` (noun subject, passive) node. Because we find an `nsubjpass` label, the sentence moves leftward to the `agent` node. This label is also found, thereby reaching a terminal node and being labeled as a type F2 sentence.

(3) A bird is shot by a man.

With the sentence now typed as F2, we apply specific F2 extraction rules. The logical subject is taken from under the `agent` label, the verb from

Type	Description	Example	NS	NNS
A	Simple declarative transitive	The boy is kicking the ball.	117	286
B	Simple + preposition	The boy played with a ball.	5	23
C	Missing tensed verb	Girl driving bicycle.	10	44
D	Missing tensed verb + preposition	Boy playing with a ball.	0	1
E	Intransitive (No object)	A woman is cycling.	2	21
F1	Passive	An apple is being cut.	4	2
F2	Passive with agent	A bird is shot by a man.	0	6
Ax	Existential version of A or C	There is a boy kicking a ball.	0	0
Bx	Existential version of B or D	There was a boy playing with a ball.	0	0
Ex	Existential version of E	There is a woman cycling.	0	0
F1x	Existential version of F1	There is an apple being cut.	0	1
F2x	Existential version of F2	There is a bird being shot by a man.	0	0
Z	All other forms	The man is trying to hunt a bird.	2	6

Table 1: Sentence type examples, with distributions of types for native speakers (NS) and non-native speakers (NNS)

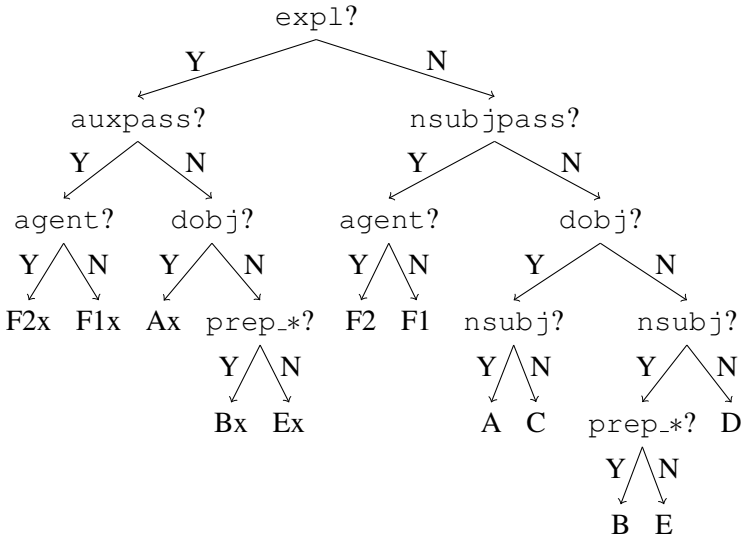

Figure 3: Decision tree for determining sentence type and extracting semantic information

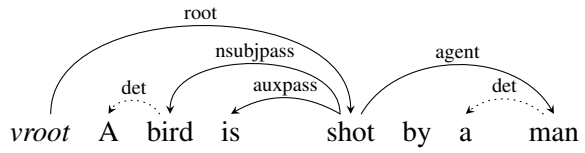

Figure 4: The dependency parse of (3)

root, and the logical object from nsubjpass, to obtain *shot(man,bird)*, which can be lemmatized to *shoot(man,bird)*. Very little effort goes into this process: the parser is pre-built; the decision tree is small; and the extraction rules are minimal.

We are able to use little effort in part due to the constraints in the pictures. For figure 1, for example, *the artist*, *the man in the beret*, and *the man* are all acceptable subjects, whereas if there were multiple men in the picture, *the man* would not be specific enough. In future work, we expect to relax such constraints on image contents by including rules to handle relative clauses, adjectives and other modifiers in order to distinguish between references to simi-

lar elements, e.g., *a man shooting a bird* vs. *a man reading the newspaper*.

5 Evaluation

To evaluate this work, we need to address two major questions. First, how accurately do we extract semantic information from potentially innovative sentences (section 5.2)? Due to the simple structures of the sentences (section 5.1), we find high accuracy with our simple system. Secondly, how many semantic forms does one need in order to capture the variability in meaning in learner sentences (section 5.3)? We operationalize this second question by asking how well the set of native speaker semantic forms models a gold standard, with the intuition that a language is defined by native speaker usage, so their answers can serve as targets. As we will see, this is a naïve view.

5.1 Basic distribution of sentences

Before a more thorough analysis, we look at the distribution of sentence types, shown in Table 1, broken down between native speakers (NSs) and non-native speakers (NNSs). A few sentence types clearly dominate here: if one looks only at simple declaratives, with or without a main verb (types A and C), one accounts for 90.7% of the NS forms and 84.6% of the NNS ones, slightly less. Adding prepositional forms (types B and D) brings the total to 94.3% and 90.8%, respectively. Although there will always be variability and novel forms (cf. type Z), this shows that, for situations with basic transitive actions, developing a system (by hand) for a few sentence types is manageable. More broadly, we see that clear and simple images nicely constrain the task to the point where shallow processing is feasible.

5.2 Semantic extraction

For the purpose of evaluating our extraction system, we define two major classes of errors. The first are *triple errors*, responses for which our system fails to extract one or more of the desired subject, verb, or object, based on the sentence at hand and without regard to the target content. Second are *content errors*, responses for which our system extracts the desired subject, verb and object, but the resulting triple does not accurately describe the image (i.e., is an error of the participant's). We are of course concerned with reducing the triple errors. Examples are in Table 2.

Triple errors are subcategorized as *speaker*, *parser*, or *extraction* errors, based on the earliest part of the process that led to the error. Speaker errors typically involve misspellings in the original sentence, leading to an incorrect POS tag and parse. Parser errors involve a correct sentence parsed incorrectly or in such a way as to indicate a different meaning from the one intended; an example is given in Figure 5. Extraction errors involve a failure of the extraction script to find one or more of the desired subject, verb or object in a correct sentence. These typically involve more complex sentence structures such as conjoined or embedded clauses.

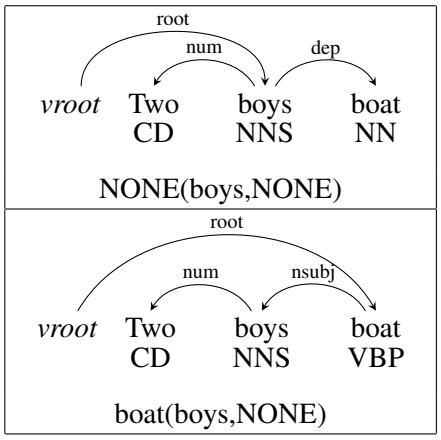

Figure 5: A parser error leading to a triple error (top), and the desired parse and triple (bottom).

As shown in table 2, we obtain 92.3% accuracy on extraction for NNS data and roughly the same for NS data, 92.9%. However, many of the errors for NNSs involve misspellings, while for NSs a higher percentage of the extraction errors stem only from our hand-written extractor, due to native speakers using more complex structures. For a system interacting with learners, spelling errors are thus more of a priority (cf. Hovermale, 2008).

Content errors are subcategorized as *spelling* or *meaning* errors. Spelling errors involve one or more of the extracted subject, verb or object being misspelled severely enough that the intended spelling cannot be discerned. A spelling error here is unlike those included in *speaker* errors above in that it does not result in downstream errors and is a well-

		Error type	Example Sentence	Example Triple	Count (%)
Triple error	NNS	Speaker	A man swipped leaves.	leaves(swipped,man)	16 (4.1%)
		Parser	Two boys boat.	NONE(boys,NONE)	5 (1.3%)
		Extraction	A man is gathering lots of leafs.	gathering(man,lots)	9 (2.3%)
		Total (390)			**30 (7.7%)**
	NS	Speaker	(*None*)		0 (0%)
		Parser	An old man raking leaves on a path.	leaves(man,path)	2 (1.4%)
		Extraction	A man has shot a bird that is falling from the sky.	shot(bird,sky)	8 (5.7%)
		Total (140)			**10 (7.1%)**
Content error	NNS	Spelling	The artiest is drawing a portret.	drawing(artiest,portret)	36 (9.2%)
		Meaning	The woman is making her laundry.	making(woman,laundry)	23 (5.9%)
		Total (390)			**59 (15.1%)**
	NS	Spelling	(*None*)		0 (0%)
		Meaning	A picture is being taken of a girl on a bike.	taken(NONE,picture)	3 (2.1%)
		Total (140)			**3 (2.1%)**

Table 2: Triple errors and content errors by subcategory, with error rates reported (e.g., 7.7% error = 92.3% accuracy)

formed triple except for a misspelled target word. Meaning errors involve an inaccurate word within the triple. This includes misspellings that result in a real but unintended word (e.g., *shout(man,bird)* instead of *shoot(man,bird)*).

The goal of a system is to identify the 15.1% of NNS sentences which are content errors, in order to provide feedback. Currently, the 7.7% triple errors would also be grouped into this set, showing the need for further extraction improvements. Also notable is that three content errors were encountered among the NS responses. All three were meaning errors involving some meta-description of the image prompt rather than a direct description of the image contents, e.g., *A picture is being taken of a girl on a bike* vs. *A girl is riding a bike*.

5.3 Semantic coverage

Given a fairly accurate extraction system, as reported above, we now turn to evaluating how well a gold standard represents unseen data, in terms of semantic matching. To measure coverage, we take the intuition that a language is defined by native speaker usage, so their answers can serve as targets, and use NS triples as our gold standard. The set of NS responses was manually arbitrated to remove any unacceptable triples (both *triple* and *content* errors), and the remaining set of lemmatized triples was taken as a gold standard set for each item.

Similarly, with the focus on coverage, the NNS triples were amended to remove any triple errors. From the remaining NNS triples, we call an appropriate NNS triple found in the gold standard set a **true positive (TP)** (i.e., a correct match), and an appropriate NNS triple *not found* in the gold standard set a **false negative (FN)** (i.e., an incorrect non-match), as shown in Table 4. We adopt standard terminology here (TP, FN), but note that we are investigating what *should be* in the gold standard, making these false negatives and not false positives. To address the question of how many (NS) sentences we need to obtain good coverage, we define **coverage** (=recall) as *TP/(TP+FN)*, and report, in Table 3, 23.5% coverage for unique triple types and 50.8% coverage for triple tokens.

		NNS +	NNS −
NS	Y	TP	FP
	N	FN	TN

Table 4: Contingency table comparing presence of NS forms (Y/N) with correctness (+/−) of NNS forms

We define an inappropriate NNS triple (i.e., a content error) *not found* in the gold standard set as a **true**

Item	NS	NNS	TP	TN	FN	Coverage		Accuracy	
						Ty.	Tok.	Ty.	Tok.
1	5	14	3	2	9	3/12	23/38	5/14	25/39
2	6	14	3	5	6	3/9	15/28	8/14	20/32
3	6	19	5	7	7	5/12	23/30	12/19	30/36
4	4	8	2	2	4	2/6	32/37	4/8	34/39
5	4	24	1	8	15	1/16	3/25	9/24	11/33
6	8	22	3	5	14	3/17	16/31	8/22	21/36
7	7	23	5	4	14	5/19	14/35	9/23	18/39
8	6	23	5	6	11	5/16	10/30	11/22	17/36
9	7	33	3	12	18	3/21	3/23	15/33	15/35
10	5	21	2	13	6	2/8	14/24	15/21	27/35
Total	58	201	32	64	104	32/136	153/301	96/200	218/360
						23.5%	50.8%	48.0%	60.6%

Table 3: Matching of semantic triples: *NS*/*NNS*: number of unique triples for NSs/NNSs. Comparing NNS types to NS triples, *TP*: number of true positives (types); *TN*: number of true negatives; *FN*: number of false negatives. *Coverage* for *Ty*pes and *Tok*ens = $\frac{TP}{TP+FN}$; *Accuracy* for *Ty*pes and *Tok*ens = $\frac{TP+TN}{TP+TN+FN}$

negative (TN) (i.e., a correct non-match). **Accuracy** based on this gold standard—assuming perfect extraction—is defined as *(TP+TN)/(TP+TN+FN)*.[3] We report 48.0% accuracy for types and 60.6% accuracy for tokens.

The immediate lesson here is: NS data alone may not make a sufficient gold standard, in that many correct NNS answers are not counted as correct. However, there are a couple of issues to consider here.

First, we require exact matching of triples. If maximizing coverage is desired, extracting individual subjects, verbs and objects from NS triples and recombining them into the various possible *verb(subj,obj)* combinations would lead to a sizable improvement. An example of triples distribution and coverage for a single item, along with this recombination approach is presented in Table 5.

It should be noted, however, that automating this recombination without lexical knowledge could lead to the presence of unwanted triples in the gold standard set. Consider, for example, *do(woman,shirt)*—an incorrect triple derived from the correct NS triples, *wash(woman,shirt)* and *do(woman,laundry)*. In addition to handling pronouns (e.g., *cut(she,apple)*) and lexical relations (e.g., *apple* as a type of *fruit*), one approach might be

Type	NNS	NS	Coverage
cut(woman,apple)	*5*	*0*	*(5)*
cut(someone,apple)	4	2	4
cut(somebody,apple)	3	0	
cut(she,apple)	3	0	
slice(someone,apple)	2	5	2
cut(person,apple)	2	1	2
cut(NONE,apple)	*2*	*0*	*(2)*
slice(woman,apple)	1	1	1
slice(person,apple)	1	1	1
slice(man,apple)	1	0	
cut(person,fruit)	1	0	
cut(people,apple)	1	0	
cut(man,apple)	1	0	
cut(knife,apple)	1	0	
chop(woman,apple)	1	0	
chop(person,apple)	1	0	
slice(NONE,apple)	0	2	
Total	30	12	10 *(17)*

Table 5: Distribution of valid tokens across types for a single PDT item. Types in italics do not occur in the NS sample, but could be inferred to expand coverage by recombining elements of NS types that do occur.

[3] Accuracy is typically defined as (TP+TN)/(TP+TN+FN+FP), but false positives (FPs) are cases where an incorrect learner response was in the gold standard, and we have already removed such cases (i.e., FP=0).

to prompt NSs to give multiple alternative descriptions of each PDT item.

A second issue to consider is that, even when only examining cases where the meaning is literally correct, NNSs produce a wider range of forms to describe the prompts than NSs. For example, for a picture showing what NSs overwhelmingly described as a *raking* action, many NNSs referred to a man *cleaning* an area. Literally, this may be true, but it is not native-like. This behavior is somewhat expected, given that learners are encouraged to use words they know to compensate for gaps in their vocabularies (Agustín Llach, 2010). This also parallels the observation in SLA research that while second language learners may attain native-like grammar, their ability to use pragmatically native-like language is often much lower (Bardovi-Harlig and Dörnyei, 1998). The answer to what counts as a correct meaning will most likely lie in the purpose of an application, reflecting whether one is developing native-ness or whether the facts of a situation are expressed correctly. In other words, rather than rejecting all non-native-like responses, an ILT may need to consider whether a sentence is native-like or non-native-like as well as whether it is semantically appropriate.

6 Summary and Outlook

We have begun the process of examining appropriate ways to analyze the semantics of language learner constructions for interactive situations by describing data collected for a picture description task. We parsed this data using an off-the-shelf parser with settings geared towards obtaining appropriate semantic forms, wrote a small set of semantic extraction rules, and obtained 92–93% extraction accuracy. This shows promise at using images to constrain the syntactic form of a "free" learner text and thus be able to use pre-built software. At the same time, we discussed how learners give responses which are literally correct, but are non-native-like. These results can help guide the development of ILTs which aim to process the meaning of interactive statements: there is much to be gained with a small amount of computational effort, but much work needs to go into delineating a proper set of gold standard forms.

There are several ways to take this work. First, given the preponderance of spelling errors in NNS data and its effect on downstream processing, the effect of automatic spelling correction must be taken into account. Secondly, we only investigated transitive verbs, and much needs to be done to investigate interactions with other types of constructions, including the definition of more elaborate semantic forms (Hahn and Meurers, 2012). Finally, to better model ILTs and the interactions found in activities and games, one can begin by modeling more complex visual prompts. By using video description tasks or story retell tasks, we can elicit more complex narrative responses. This would allow us to investigate the possibility of extending our current approach to tasks that involve greater learner interaction.

Acknowledgments

We would like to thank the task participants, David Stringer for assistance in developing the task, Kathleen Bardovi-Harlig, Marlin Howard and Jayson Deese for recruitment help, and Ross Israel for evaluation discussion. For their helpful feedback, we would also like to thank the three anonymous reviewers and the attendees of the Indiana University Linguistics Department Graduate Student Conference.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: a pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 385–393. Association for Computational Linguistics, Stroudsburg, PA, USA.

Maria Pilar Agustín Llach. 2010. Lexical gap-filling mechanisms in foreign language writing. *System*, 38(4):529 – 538.

Luiz Amaral and Detmar Meurers. 2011. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, 23(1):4–24.

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.

Kathleen Bardovi-Harlig and Zoltán Dörnyei. 1998. Do language learners recognize pragmatic violations? Pragmatic versus grammatical awareness in instructed L2 learning. *TESOL Quarterly*, 32(2):233–259.

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*, pages 119–134. Lawrence Erlbaum, Mahwah, NJ.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*. Portland, OR.

Martin Chodorow, Joel Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30. Prague.

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

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 12:2461–2505.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62. Montréal.

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.

Marie-Catherine de Marneffe and Christopher D. Manning. 2012. *Stanford typed dependencies manual*. Originally published in September 2008; Revised for Stanford Parser v. 2.0.4 in November 2012.

William DeSmedt. 1995. Herr Kommissar: An ICALL conversation simulator for intermediate german. In V. Holland, J. Kaplan, and M. Sams, editors, *Intelligent Language Tutors. Theory Shaping Technology*, pages 153–174. Lawrence Erlbaum Associates, Inc., New Jersey.

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

Rod Ellis. 2006. Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly*, 40:83–107.

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.

Michael Hahn and Detmar Meurers. 2012. Evaluating the meaning of answers to reading comprehension questions: A semantics-based approach. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*, pages 326–336. Association for Computational Linguistics, Montreal, Canada.

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.

Trude Heift and Mathias Schulze. 2007. *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.

DJ Hovermale. 2008. Scale: Spelling correction adapted for learners of English. Pre-CALICO Workshop on "Automatic Analysis of Learner

Language: Bridging Foreign Language Teaching Needs and NLP Possibilities". March 18-19, 2008. San Francisco, CA.

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

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Morgan & Claypool Publishers.

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

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

Detmar Meurers. 2012. Natural language processing and language learning. In Carol A. Chapelle, editor, *Encyclopedia of Applied Linguistics*. Blackwell. to appear.

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*, pages 227–248. Seameo Regional Language Centre.