Semantic Analysis of Learner Sentences

Levi King
Dissertation Proposal

January 9, 2017

1 Summary

This work is motivated by a disconnect between the fields of second language acquisition (SLA) and intelligent computer-assisted language learning (ICALL), with most SLA research supporting communicative and task-based learning methods, while many existing ICALL applications focus on explicit grammar instruction and correction. By combining existing language resources and natural language processing (NLP) tools to form an evaluation system, the work aims to demonstrate that reliable, automatic, contextual, meaning-based analysis of NNS (non-native speaker) sentences (whether for testing, tutoring or otherwise) is possible without the need for developing expensive new data sets or processing tools, and thus encourage the use of such approaches in ICALL systems. The dissertation will be an expansion of previous work that attempted to automatically evaluate the semantic appropriateness of English NNSs' responses to a picture description task (PDT) by comparing the responses to those of native speakers (NSs). This work relies heavily on dependency parsing and statistical comparisons of syntactic dependencies to determine how well a NNS response matches the collection of NS responses.

A major focus of the research will be establishing representations of NS and NNS sentences and a corresponding gold standard (GS) by which NNS responses can be evaluated. Many ICALL systems primarily handle grammar, in which case a grammar model is sufficient for conducting the analysis and providing feedback. Other systems are menu based, with users choosing responses from a pre-defined set. However, as the system under development allows for novel responses and addresses content over form, the defining of a meaning-based GS for communicative tasks will be both a central challenge and a major contribution to ICALL and to other areas of NLP that handle user

sentences in communicative and visual contexts, such as dialog systems, translation systems and speech-to-text engines. Other major tasks will include refining the overall approach to handle a wide variety of PDT items eliciting many different sentence forms, establishing methods for discriminating between acceptable and non-acceptable responses, and automatically determining the optimal system settings for processing responses. The designing of a robust, pedagogically sound feedback module is beyond the scope of this dissertation, but the system will include a lightweight feedback module that at minimum provides the user with one or more of the NS responses that is most similar to the user's response. Beyond dependency parsing, this work will also examine the effect of incorporating semantic role labeling (SRL) tools and lexical information (hyper-/hypo-/synonyms via WordNet or similar resources) on coverage and overall performance.

2 Research questions

Given both the practical need to develop a GS and the theoretical issues surrounding nativeness and ultimate attainment, the first research question is:

1. Are the responses of intermediate and advanced L2 English learners sufficiently similar to those of NSs to allow automatic evaluation based on a collection of NS responses? In other words, do learners demonstrate significant overlap with native-like usage in a PDT setting?

The images of the communication task serve as a rough simulation of real world scenes; given the lack (and desirability) of learner tools that analyze language content in visual contexts, the second research question is:

2. In the constrained communicative environment of a PDT, what are appropriate response and GS representations for the purpose of providing meaning-oriented feedback or evaluation? In other words, which linguistic components are crucial and which are superfluous?

As mentioned above, one goal of this project is to show that content-based evaluation of learner sentences is possible without the expense of developing major new tools or language resources; in this vein, the third research question is:

3. What kinds of existing NLP tools and language resources can be integrated to

form a content analysis system for open response language learning tasks?

As discussed later, this work attempts to take statistical methods traditionally used to analyze the frequencies of individual words in sentences and apply those methods to the frequencies of syntactic dependencies in sentences, as one means of deriving semantic information from syntactic tools. Thus, the fourth research question is:

4. How do "bag-of-words" and "bag-of-dependencies" approaches compare in terms of performance? Is a bag-of-words approach alone adequate for our needs?

Given that the system has thus far relied primarily on a parser, lemmatizer and spelling correction module, without the inclusion of semantic tools, the fifth research question is:

5. Can the accuracy of the system be improved by the inclusion of semantic information from tools like semantic role labelers or WordNet?

To evaluate the system, it will be necessary to have humans annotate the NNS responses, then compare this annotation to system output. In light of the specific nature of the PDT and the goal of analyzing semantic appropriateness, the sixth research question is:

6. What is the annotation scheme for this task and can the system perform within the range of inter-annotator reliability? Relatedly, what does it mean for a response to be *appropriate* and how can this be captured with annotation?

3 Organization

The dissertation will be organized as shown in section 3.1. Discussion of the core issues in each section of the dissertation follows, in section 3.2.

3.1 Overview

1. **Introduction.** Overview of the project; motivations; goals.
2. **Literature Review.** ICALL, NLP and related topics in SLA.
3. **Data Collection & Description.** Description of the PDT, data, & annotation.
4. **Method.** Overview of current method & our previous work.

(a) **Previous Work.** Discussion of previous rule-based semantic triple methods and effectiveness.

(b) **Gold Standard.** Description of GS and decisions involved.

(c) **Response Processing and Representation.** Description of the representation and (pre)processing to derive it.

(d) **Response Evaluation.** Description of methods to grade NNS responses via comparison with GS; translating these comparison scores into decisions.

5. **Experiments & Results.** Presentation of experiments to determine the strongest approach and optimize the system.

6. **Feedback.** Explanation of our feedback module and its output.

7. **Conclusion.** Summary of the work highlighting any important findings.

3.2 Discussion

Introduction. The work will begin with an overview of the project, the broader philosophical motivations of bridging the SLA-ICALL divide and the specific goals for the system. I will also explain how the system differs from existing ICALL systems, what major challenges are anticipated, and how the work fits into current and past work in ICALL and related areas.

Literature Review. This work sits at the intersection of NLP, SLA and language testing, and related work from these fields will be examined. This includes work on PDTs, ICALL, dependency parsing, spelling correction, SRL and automatic evaluation of essays and short answers. This work will touch on related research in other domains, such as image processing in NLP (as in automatic image captioning). Despite much current interest in analyzing relationships between visual and linguistic information, this work will involve learner language, which has generally been overlooked in such research (for exceptions, see Somasundaran and Chodorow [2014] and Somasundaran et al. [2015]).

Through its focus on language in context, the current project is in accord with contemporary theory in SLA and second language instruction, which suggest 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]. 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." Indeed, Bailey and Meurers [2008] observe that in contrast to these principles, "existing research on ICALL systems has focused primarily on providing practice with grammatical forms." For current purposes, bridging this divide means shifting the primary task of an ICALL application from analyzing grammar to evaluating semantic appropriateness and accuracy. The work involves two somewhat conflicting major challenges to this goal: ensuring that the system is consistent with SLA-informed approaches to language teaching, and sufficiently constraining the input to the system to allow for automatic evaluation. As discussed in Amaral and Meurers [2011], "in order to obtain tractable and reliable NLP supporting the analysis of both form and meaning, it is necessary to restrict the ill-formed and well-formed variation in learner input that an ICALL system needs to deal with." Thus the learner input and the context in which it is collected are central considerations.

Data Collection & Description. The use of PDTs in language research is well-established in areas of study ranging from SLA to Alzheimer's disease [Ellis, 2000, Forbes-McKay and Venneri, 2005]. PDTs are well suited to connecting visual contexts and linguistic information, and the careful selection of images can help to constrain the expected responses, as is necessary for the content analysis of free responses. In the current work, PDTs serve not only as a research tool but as a proxy for language use in visual settings, extending the impact of this work beyond ICALL and second language testing and into many areas of NLP where contextual NNS language may require processing.

Data will be collected largely through ESL courses, with participants completing the task individually in a computer lab, under supervision. If necessary, additional respondents meeting the needed NNS or NS profiles may be collected remotely through an online version of the task, with instructions mirroring the supervised setting as closely as possible. NNS participants will be chosen primarily from intermediate and advanced levels in the local Intensive English Program; additional participants with corresponding proficiency levels will also be used if necessary. Lessons learned from previous the PDT experiments will be considered in the development of the current PDT. For example, for some PDT items, NSs overwhelmingly used a particular verb or construction, leading to poor coverage of accurate but non-nativelike responses. To avoid this pitfall, each NS will be instructed to provide multiple responses (which will

be weighted differently in the GS), leading to a wider range of responses being covered. Relevant anonymized participant information will also be collected, such as the length of English study and native language.

Given the goal of handling a variety of sentence forms, developing an effective PDT involves selecting or creating images that require the use of linguistic variation. For example, eliciting transitive sentences requires images that depict a clear subject and object. Eliciting ditransitives will require images that add an indirect object, and illustrating these complex ditransitive events clearly will require greater care. Other variations in the PDT material are planned. For example, two "minimal pair" items may depict the same transitive event but reverse the role of the subject and object; such items will later be used to compare the effectiveness of simple word-based approaches with dependency-based approaches.

A portion of the data will be held out for testing the completed system. For the rest, observations about the actual responses collected will be explored and may influence decisions in the system development. Such observations will include the distribution of various sentence forms and constructions (*declaratives* vs. *passives*; *intransitives, transitives, ditransitives*; *relative clauses*, etc.) and how these differ among the NS and NNS data. The work will also consider qualitative aspects of the data, such as the rate of spelling errors among NNS responses and any problems among NS responses.

By design, annotation is not required for this system to function, but in order to evaluate its performance, the data must be annotated by human raters. The development of an appropriate annotation scheme will be important here. Most likely this will be a simple "yes" (acceptable) or "no" (unacceptable) rating for annotators, but as the system is developed, a Likert scale may prove more appropriate. (The full annotation guidelines will be included in the appendix.) Examples of annotated responses will be provided, with special attention paid to difficult or ambiguous cases. Inter-annotator reliability measures will also be examined here.

Method. This chapter will consist of four sections covering the following: *my previous work in this area, the gold standard (GS), response processing and representation, and response evaluation.*

Previous Work. This part of the work will begin with an overview of my previous work with shallow processing of NNS responses to PDT items, including descriptions of the PDT, the data, the system and the performance (see King and Dickinson [2013]

and King and Dickinson [2014]). Notably, this past approach used a markedly different method, relying on custom rules for extracting *subject-verb-object* semantic triples and attempting to match these NNS triples against NS triples. This approach was made possible in large part by the fact that the PDT included only images of transitive events; this allowed for constraints on the syntactic form of the sentences and the development of rules for extracting subjects, verbs and objects based on predictable dependency labels and part-of-speech tags. For example, most responses to an item depicting a boy kicking a ball resulted in the triple `kick(boy,ball)`. For relatively unambiguous items like this, the performance of the system was satisfactory. For other items, however, such as a close-up photograph of a hand cutting an apple; respondents chose a wider variety of verbs and disagreed on the gender of the hand, and many passivized the sentence (e.g., "An apple is being cut"); the lexical variation lead to somewhat lower coverage, and crucially, the unexpected passivization moved the object to the subject position, changing the triple drastically and severely harming performance. Such unexpected responses have pointed toward major revisions to the approach; significant findings from this past work will be examined, such as the need for methods to expand lexical coverage, and the greater robustness of dependencies over semantic triples, explaining how such insights will guide the current work.

Gold Standard. The GS for communicative tasks is arguably the keystone of this project. The GS for a given PDT item will essentially be a collection of representations of individual NS responses to that item. As mentioned above, the system (and its GS) should focus on content over form and be capable of handling novel responses. The exact nature of the GS and its purpose will be explored here: how it is directly related to the PDT and the NS instructions, what it is intended to represent and how it will be used in the system, as well as its evaluative power and its limitations. This part of the work will experiment with automatically extrapolating additional GS content from the NS responses. For example, if the set of all NS responses for an item consists of the sentences *The woman mailed letters* and *The lady sent mail*, the GS contains the dependencies `[subj, woman, mailed]`, `[obj, mailed, letters]`, `[subj, sent, lady]`, `[obj, sent, mail]` (among others). However, by recombining subjects, objects and verbs, additional dependencies can be added, like `[subj, woman, sent]` and `[obj, sent, letters]`.

The "philosophy" of this work suggests that the GS be automatically derived from the NS responses; the practicality of this notion will be explored here, along with any potential complications and any steps taken to arbitrate the NS responses before gen-

erating the GS.

Response Processing & Representation. This portion of the research will consist of the process of deriving an evaluable representation from a NNS response. First, a set of pre-processing steps will be taken to obtain a normalized form (or forms) of the response, with regard to spelling, morphology, and minor syntactic variation. Variations in the order of conjuncts will addressed; for example, in the sentence, *The men chopped carrots and potatoes*, methods will be employed for avoiding the potential decrease in coverage introduced by the conjunction. That is, if NS responses only use the ordering *carrots and potatoes*, but some NNS responses use *potatoes and carrots*, these NNS responses would not be covered by the GS. Overcoming this issue could include propagating relationships across conjunctions and storing all resulting relationships in the response representation; i.e., this would result in something like *chop(man,carrot)* and *chop(man,potato)*. Alternatively, it could involve including concatenated versions of all possible orderings, i.e., *chop(man,carrot_and_potato)* and *chop(man,potato_and_carrot)*. The effectiveness of such methods for handling conjunctions will be explored here.

Spelling correction has been the biggest preprocessing concern, and the system for correcting spelling errors before further processing will be presented here. Through my previous work in this area, I developed a system that uses an existing spelling tool to generate candidate spellings for each word. The NS responses and lists of "stop words" (common function words) are then used to look for matches among candidates, effectively filtering out many candidates and significantly reducing the runtime. All possible sentences are generated from combinations of the remaining candidates, and these resulting candidate sentences are then given probabilities via an ngram language model (also using existing tools). An ngram language model, or *LM*, is a tool that iterates a sliding window of n words over some very large text, tallying the number of times each unique sequence of n words (an *ngram*) is encountered and ultimately converting these counts to a model of ngrams and their frequencies, which is theoretically representative of the whole language. When given an unseen text, the LM compares the frequency of the text's ngrams to those in the model and returns the probability of the unseen text; an incoherent string of words should receive a low score, while a well-formed, coherent sentence should receive a high score. In my current system, some predetermined number of the highest-scored candidate sentences are then passed on for further processing. This system works presently, but I expect to explore methods to improve it, primarily through using more appropriate training text for the ngram language model, which currently relies on newspaper text.

Finally, preprocessing will involve lemmatization. This allows various forms or inflections of a word to be mapped to a single form, which improves coverage and reduces the need for an exhaustive GS. For instance, from the previous example, this yields *The man chop carrot and potato.* This step relies on existing tools and is somewhat less probabilistic and more straightforward than spelling correction. However, as some information (such as plurality and verb tense) is lost via lemmatization, any decisions to preserve such information elsewhere in the representation will be fully explored here.

In this dissertation, I plan to introduce two previously unused tools to the system and explore their effects on performance: semantic role labelers (SRLs) and WordNet. [1] Existing SRLs vary, but they all attempt to show semantic relationships in the sentence. This generally means either explicitly labeling entities in the sentence with theta roles like *agent, patient*, etc., or using indexing to indicate the semantic relation between a verb and its arguments. My goal is to use a SRL in order to identify cases where a word's syntactic and semantic roles may vary across responses. In a transitive sentence, for example, the syntactic subject is usually an *agent* acting on a *patient*, which is the syntactic object. Describing the same event with a passive sentence, however, means the *patient* becomes the syntactic subject. If implemented correctly, a SRL should help map such variations to a single representation.

WordNet is a hierarchical database of English words. I will explore ways to implement this resource in order to boost response coverage. For example, if the set of NNS responses (and in turn, the GS) for a PDT item contains only the verbs *scrub* and *brush*, but a NS response uses the (hypernym) verb *clean*, the system would ideally to be able to recognize the close relationship between these words and consider it in the evaluation, rather than outright reject the response. Thoughtful implementation of WordNet or related tools employing it may allow for this kind of lexical expansion.

Determining the ultimate representation of the response is a central problem in this dissertation, but past experiments suggest this will continue to rely on dependencies as the core representation. Currently, the system simply concatenates the dependency label with the lemmatized dependent and head using some delimiter, e.g., `subj#boy#kick` (label#dependent#head). For clarity, I refer to this as a *dependency string*. Experiments here will compare the use of the fully specified dependency with partially abstracted dependencies where either the label, head, or dependent is replaced

[1] Note that the decision to mention these tools here is somewhat arbitrary; it may prove more sensible to use them in deriving the GS or during the response evaluation; any such changes will reflected in the dissertation.

with a dummy word, as in `subj#(null)#kick`, `(null)#boy#kick`, etc. I expect the most effective representation to make use of some weighted combination of these variations. Following experimentation, the final representation will likely also incorporate information from the SRL and WordNet.

Response Evaluation. This part of the work will focus on the exact process by which the system takes in an NNS response to a PDT item and returns an evaluation of how well that response matches the meaning of NS responses. This degree of matching can be seen as an approximation of a rating for how accurate, appropriate, and/or nativelike a NNS response is. Such a measure is intended to be useful in the development of ICALL systems, but it would also be applicable in language testing and other subfields of NLP. I will present the various approaches used to compare the internal representation of an NNS response with the GS. The current approaches rely on dependency parsing (via the Stanford Parser), but experiments are planned to determine whether simpler word-based approaches can boost performance. In addition to comparing current NNS dependency representations against the GS, the system will need to make use of any new information in the response representation, which may include WordNet entries, SRL output, or information from similar tools.

Given the unexplored system variations like including SRL or WordNet information, new approaches will likely be added; as of now, four major approaches have been developed for the task of rating a response's similarity to the GS:

- **FA** (for *frequency average*): *This approach is the baseline, which relies only on frequency in the GS.* The system assigns the NNS terms scores equal to their relative frequencies in the GS, then calculates the average term score, which becomes the sentence score. Note that a *term* is the relevant unit of analysis; usually this a word, but here it may be a word or a concatenated dependency string (discussed above).

- **TA** (for *tf-idf average*): *Like **FA**, this approach assigns scores from the GS terms to terms in the NNS, but in this case, the scores come from a statistical analysis called tf-idf*[2]. The system runs tf-idf on the GS sentences (but not the NNS responses), and simply assigns each term in the NNS sentence the tf-idf score that term received in the GS (or 0 if it doesn't occur there), then calculates the

[2]term frequency–inverse document frequency; a method of comparing a term's frequency in a given document against its frequency in a general sample of the language; this is often used for auto-indexing or, more importantly to this work, to determine what content is important in a document.

average dependency score, which becomes the sentence score.

- **FC** (for *frequency comparison*): *Like **FA**, this approach relies on term frequencies, but adds a statistical comparison of the NNS and GS term frequencies.* The system calculates the relative frequency of each term in the GS sentences, then does the same for the NNS sentence. It then treats these lists of scores as vectors, and GS terms missing from the NNS are inserted into the NNS vector with a score of zero, and vice versa. The vectors are then compared using cosine similarity.

- **TC** (for *tf-idf comparison*): *This approach combines the use of tf-idf in **FA** with the vector comparison in **FC**.* This approach was the initial focus of attempts to automate comparison of the GS with NNS sentences and gave rise to approaches **FA**, **FC** and **TA** above. Here, the system calculates tf-idf scores for the GS terms and for the NNS terms. As in *FC*, these sets of scores are treated as vectors, and the missing terms in each vector are added with a score of zero. The vectors are then compared using cosine similarity.

The current approaches result in scores for each NNS response, allowing responses to be ranked. These scores indicate how closely each response matches the collection of NS responses, as represented in the GS. A major challenge in this part of the work will be determining how to use these scores to discriminate between acceptable and unacceptable responses. Experiments to address this challenge are in development; I am currently performing hierarchical clustering of the PDT items by grouping together all NNS responses (and separately, all NS responses) for a given item, extracting various features such as *type-to-token ratios* (the ratio of the number of unique wordforms in a text to the total number of words in the text), creating vectors from these features and using clustering software to identify any natural classes among PDT items. Next I plan to examine the output and see if acceptable and unacceptable responses follow any reliable patterns with regard to ranking in a given cluster, as these patterns may be exploited to improve the system. For example, I suspect that "relaxing" the GS by including dependencies extrapolated by combining elements from separate responses may be more appropriate in the case of ditransitive PDT items, where a dative alternation could lead to a wider range of responses being acceptable. The process of providing users with a response evaluation will ultimately involve more than this; an exploration of the data is expected to suggest additional techniques.

Experiments & Results. Minor experiments with individual components will most likely be discussed in other sections of the dissertation, as appropriate; for example,

experiments with spelling correction will be presented with the work on response processing. This section will focus on experiments involving the four major approaches to the system outlined above, and any additional approaches developed. This will include experiments varying the parameters of these approaches, such as the form of the dependency strings and the inclusion or exclusion of SRL output. Numerous minor variables will contribute to the fine tuning of the system, but the problem of optimizing performance essentially entails exploring a search space defined by two major dimensions: **1)** the specificity or abstraction of sentence representations, and **2)** the strictness or flexibility required when matching responses to the GS. I will report the results of experiments in this space and seek to explain why certain approaches and parameters perform better than others in particular cases. The work here will overlap significantly with the clustering experiments mentioned in the previous section; as I seek to identify statistical similarities among response sets to particular PDT items, I will search for the system settings that optimize performance for these clusters. Again, this will involve selecting various features from the response sets like type-to-token ratios and the distributions of part-of-speech tags, among others.

As a hypothetical example, among items where the NNS and NS type-to-token ratios are the most similar, one might expect the use of fully specified dependency strings (*label#dependent#head*) to perform relatively well. The similar type-to-token ratios here *might* indicate that NNSs are using roughly the same vocabulary as NSs for the item. As I have observed in previous work, NNSs sometimes lack the specific, optimal vocabulary for describing an image, resulting in a wider variety of responses (and thus a higher type-to-token ratio) than their NS counterparts. (See King and Dickinson [2013] for a discussion of how NSs converge on precise words like *rake*, while NNSs fill lexical gaps with more general words like *clean*, *collect* or *sweep*.) In such situations, fully specified dependency strings might be expected to perform relatively poorly. Better performance might come from some weighted combination of less specified dependency strings; in other words, in situations where learners have a lacking vocabulary, an approach that combines smaller, overlapping bits of information might outperform an approach that searches for larger, specific matches.

I anticipate taking these findings and the findings regarding clustering or other patterns in the PDT data and performing additional experiments in which I test methods for automatically selecting the approach and parameters based on the PDT item, the GS, the set of NS responses, any available previous NNS responses (to the same PDT item), and the NNS response at hand in order to get an optimal evaluation of the NNS

response. For example, the system should be able to check the type-to-token ratios (and other features) of the NS and NNS response sets, and in cases where the vocabularies of both groups appear similar in size and distribution, automatically select the optimal system settings for evaluating a new response (perhaps involving more fully-specified dependencies, as discussed above).

The results presented here will generally measure the performance of the system at evaluating NNS PDT responses as compared to the performance and consistency of human raters, and may provide insights for SLA and ICALL. More specifically, I will measure the rates at which the system correctly accepts "good" responses (true positives) and rejects "bad" responses (true negatives), as well as the rates at which it accepts bad responses (false positives) and rejects good responses (false negatives). The precise definition of good and bad (or acceptable and unacceptable) responses is currently being explored and is contingent on the details of the annotation scheme, as the annotation will be crucial in evaluating the system.

Feedback. I believe an in depth exploration of the utility and implementation of a full-fledged feedback module for a system such as this one could be worthy of a dissertation in its own right. Thus, instead of building such a module, I will explore what such a feedback component might look like and what kind of information this system can provide (in addition to response evaluation) toward making it possible. At present, I plan to provide much simpler (and admittedly less valuable) feedback to user responses. If a NNS response exactly matches a highly ranked NS response, feedback will most likely simply indicate the match. Excluding full matches, feedback will provide the user with one or more of the NS responses that are judged by the system to be most similar to the NNS response, including information about how frequent the NS response is and how it fits into the set of NS responses. A final version of the system including such a feedback module will be piloted with several intermediate and advanced English learners, with their written impressions of the system and its feedback to be included here.

4 Timeline

Task	**Start**	**Finish**
Finalize committee		February 2016
System development	January 2016	March 2017
Data collection materials, IRB forms	July 2016	October 2016
Collect data	October 2016	November 2016
Write introduction	January 2017	February 2017
Annotate data	January 2017	February 2017
System experiments with new data	January 2017	March 2017
Write SLA section (motivation, lit review)	January 2017	February 2017
Write data collection & description section	February 2017	March 2017
Write (NLP) literature review	February 2017	March 2017
Write method section (including training data)	February 2017	March 2017
Write experiments & results section	March 2017	April 2017
Write conclusion (finish 1st draft)	April 2017	April 2017
Revisions	April 2017	April 2017
Final draft complete / Defense		May 2017

References

Luiz Amaral and Detmar Meurers. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, 23(1):4–24, 2011. URL `http://purl.org/dm/papers/amaral-meurers-11.html`.

Stacey Bailey and Detmar Meurers. 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)*, pages 107–115, Columbus, OH, 2008. `http://aclweb.org/anthology-new/W/W08/W08-0913.pdf`.

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

Marianne Celce-Murcia. 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, 2002.

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

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

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

Levi King and Markus Dickinson. Shallow semantic analysis of interactive learner sentences. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–21, Atlanta, Georgia, June 2013. URL `http://www.aclweb.org/anthology/W13-1702`.

Levi King and Markus Dickinson. Leveraging known semantics for spelling correction. In *Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning*, pages 43–58, Uppsala, Sweden, November 2014. URL `http://www.aclweb.org/anthology/W14-3504`.

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

Swapna Somasundaran and Martin Chodorow. 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*, pages 1–11, Baltimore, Maryland, June 2014. URL `http://www.aclweb.org/anthology/W14-1801`.

Swapna Somasundaran, Chong Min Lee, Martin Chodorow, and Xinhao Wang. Automated scoring of picture-based story narration. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–48, 2015.