Shallow Semantic Analysis of Interactive Learner Sentences

Levi King & Markus Dickinson
Indiana University

Workshop on Innovative use of NLP for Building Educational Applications; Atlanta, GA; June 13, 2013

Motivation

Issue:

- Intelligent Computer-Assisted Language Learning (ICALL) / Intelligent Language Tutor (ILT) systems tend to focus on grammatical errors & feedback.
- Second Language Acquisition (SLA) research has established:
  - correcting a learner's grammar is often ineffective
  - a dispreference for explicit grammar instruction

Overarching Goal:

- See ICALL/ILT focus on interaction, with learners producing *more* target language rather than perfect target language.

Where we're going

- This means shifting the task of an ICALL application from analyzing grammar to evaluating semantic accuracy and appropriateness.
  - We will focus on the extent to which we can reuse existing NLP resources.
- We approximate these goals by
  1. collecting data from a task which models some aspects of interaction, namely a picture description task (PDT),
  2. parsing it with an off-the-shelf parser,
  3. extracting semantic forms,
  4. evaluating these forms and the process, and noting the challenges throughout.

Related Work

- *Herr Komissar*: ILT/detective game for German learners, includes content analysis & sentence generation (DeSmedt, 1995), but uses many custom-built tools.
- Petersen (2010): ILT, provides feedback on questions in English, extracting meanings from an existing parser.
- Content assessment: (e.g., ETS's c-rater system (Leacock and Chodorow, 2003)); mostly focused on essay & short answer scoring.
  - Some focus on semantic analysis under restricted conditions, e.g., (Meurers et al., 2011).

Data Collection

We use a picture description task (PDT) because:

- Computer games/ILTs are visual.
- Visual prompts restrict response contents to image contents.
- Responses model real language use and are pure interlanguage– no influence of verbal prompts.

Our PDT:

- We chose 10 images depicting transitive events (unambiguous subject, verb, object) to restrict form in addition to content.
- Participants were instructed to view the image & describe the action in one sentence; past or present tense (and simple or progressive aspect) were accepted.

Data Collection

Example item and responses

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)

Data Collection

Participants

53 Participants: 14 NS, 39 NNS. The NNS consisted of:

- intermediate & upper-level adult learners enrolled in the IU Intensive English Program.
- 16 Arabic, 7 Chinese, 2 Japanese, 4 Korean, 1 Kurdish, 1 Polish, 2 Portuguese, 6 Spanish.

Method

1. Parse a sentence into a dependency representation
2. Extract a simple semantic form from this parse
   - to compare to gold standard semantic forms

Obtaining a syntactic form

Dependency parsing:

- labels dependency relations, not phrase structure;
- easily finds a sentence's subject, verb and object;

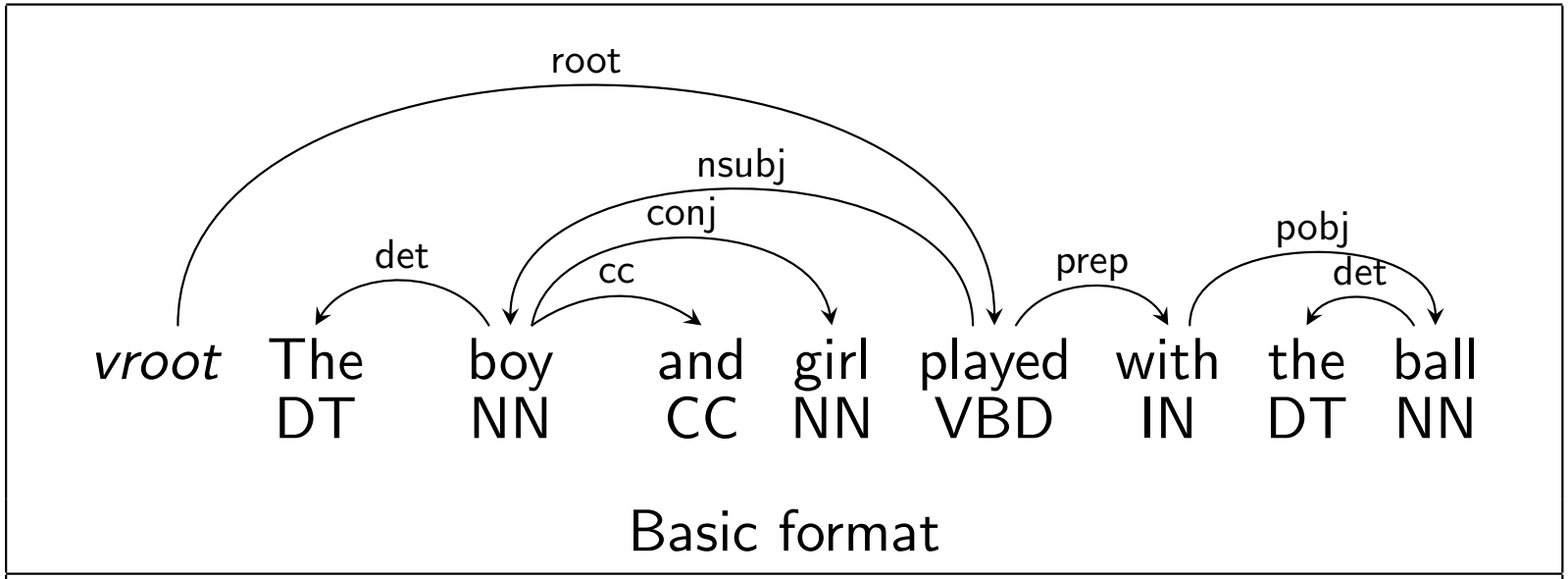
For transitive sentences, we consider S,V,O as adequate (for now) for evaluating whether sentence describes image.

We use the Stanford Parser for this task:

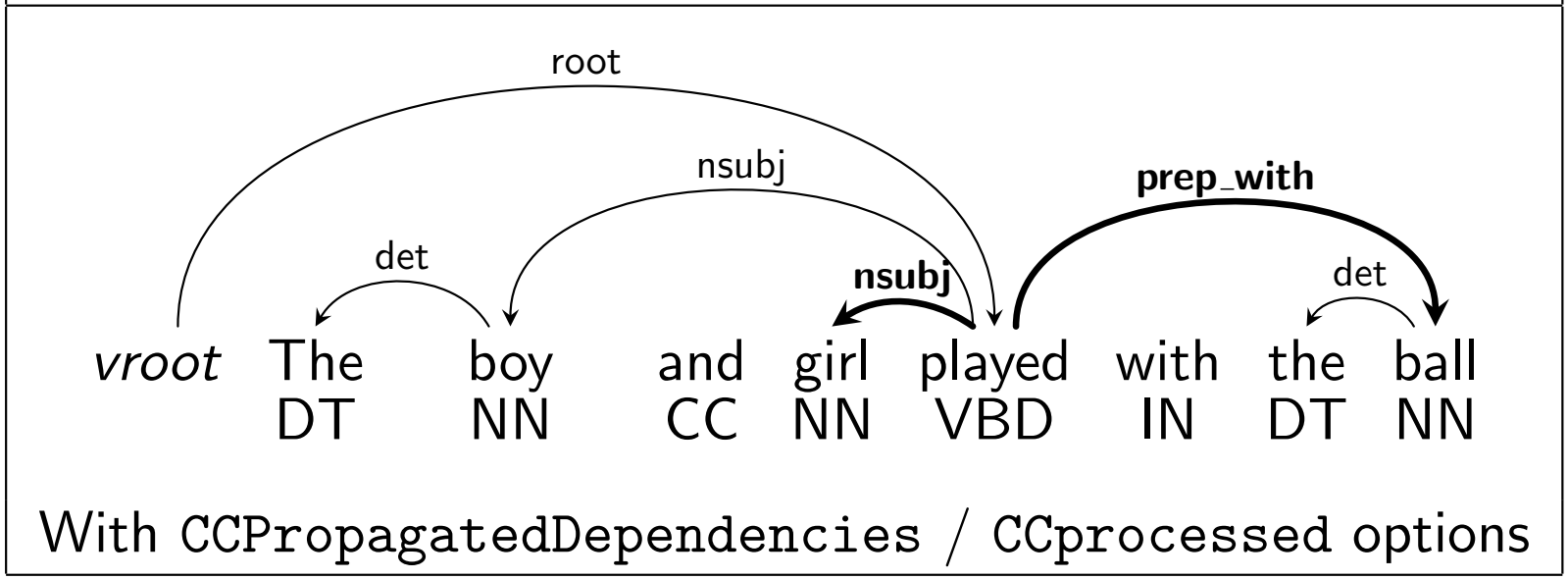
- trained on the Penn Treebank;
- use Stanford typed dependency labels;
- `CCPropagatedDependencies` / `CCprocessed` options:

1. propagate dependencies across conjunctions;
2. omit prepositions & conjunctions from sentence text; add them to dependency label between content words

Stanford Parser settings



Basic format



With `CCPropagatedDependencies` / `CCprocessed` options

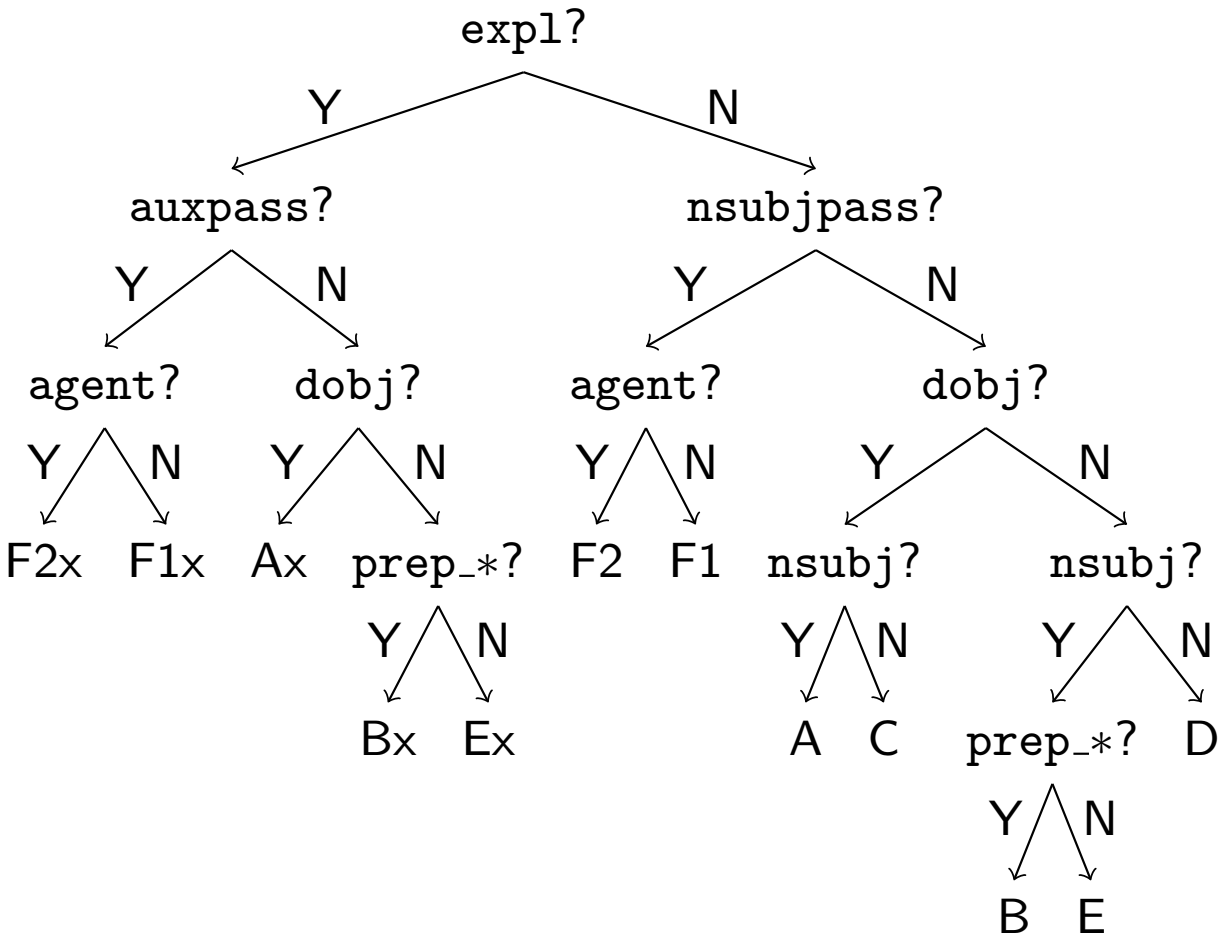
Obtaining a semantic form

- We categorized sentences into 12 types, each corresponding to a basic sentence structure.
- The type indicates that the logical S,V,O are found under particular labels, indices or POS tags.
- Distributions for the most common types are shown below; expletive types are omitted here.

Type	Description	Example	NS	NNS
A	Simple declar. trans.	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 V + prep	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
Z	All other forms	The man is trying to...	2	6

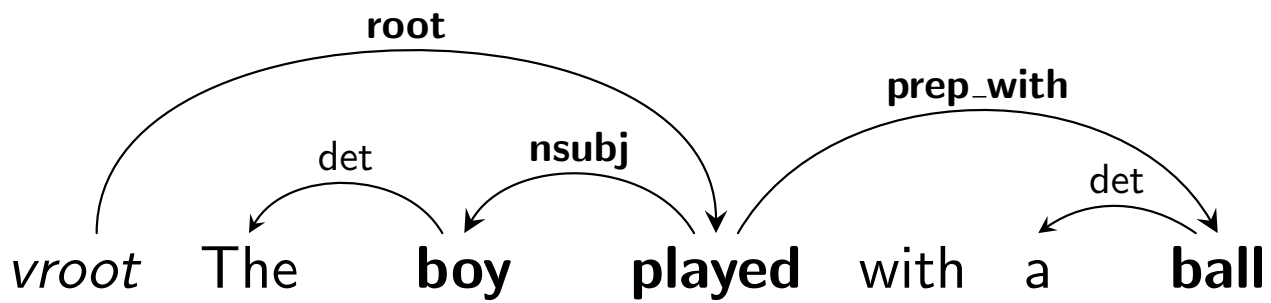
Sentence types

We use type features to construct a binary decision tree for determining type.



Rules for sentence types

- Each type has a set of rules for extracting semantic triples in the form *verb(subj,obj)*.
- For type B, for example, we extract the `root` as verb and `nsubj` as subject. The object is taken from `prep_*`, provided it is a dependent of the root.
- For the example below, we extract *played(boy,ball)*.



Evaluation

Two major questions for evaluation:

1. How accurately do we extract semantic information from potentially innovative sentences?
2. How many semantic forms do we need in order to capture the variability in learner sentences?
   - How well does the set of native speaker forms model a gold standard?

Semantic extraction

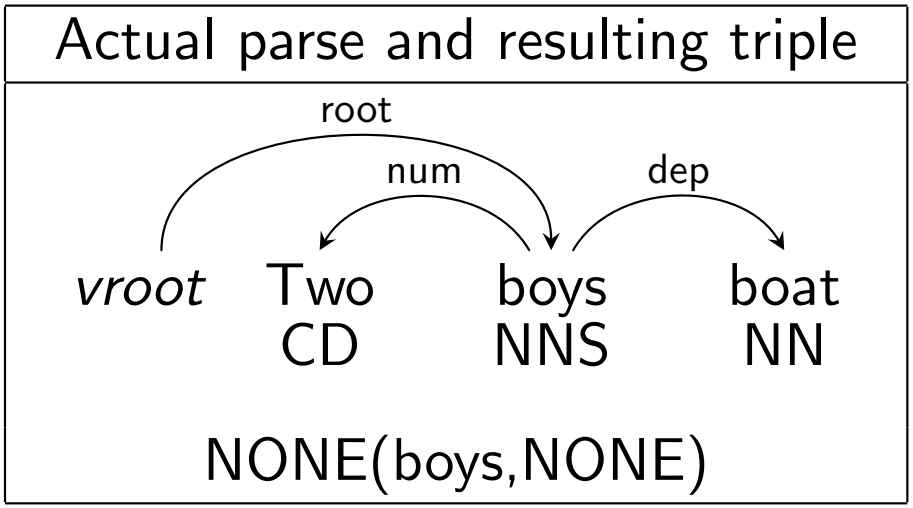
To evaluate our extraction system, we define two classes of errors:

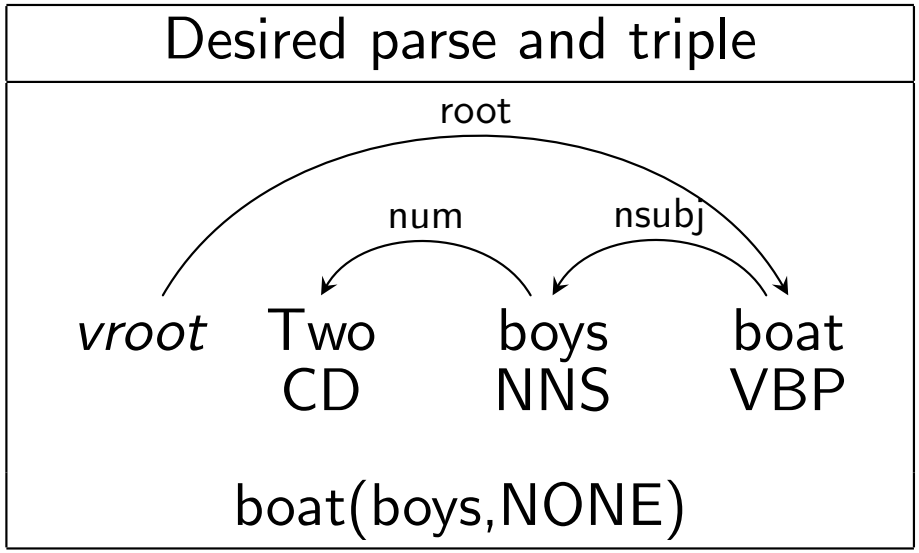
1. *triple errors*: system fails to extract one or more of the desired subject, verb, or object
   - No regard to target content
2. *content errors*: system extracts the desired triple, but the triple does not accurately describe the image

Triple errors

Error type	Example Sentence	Triple	Count (%)
	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%)**

Parser error example





Content errors

Error type	Example Sentence	Triple	Count (%)
	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%)**

Semantic coverage

Idea: Treat NS set as gold standard.

Pre-processing:

- Manually removed *triple* errors from set of NNS triples
- Manually removed *all* errors from set of NS triples
- Lemmatized: *rowed(boys,boat)* $\Rightarrow$ *row(boy,boat)*

Evaluation:

- Coverage: Measure of how many "good" NNS responses are found in NS data
- Accuracy: Measure of how many "good" NNS responses are found in NS data + how many "bad" NNS responses are *not* found in NS data

Semantic triple matching

Item	Coverage		Accuracy	
	Type	Token	Type	Token
1	3/12	23/38	5/14	25/39
2	3/9	15/28	8/14	20/32
3	5/12	23/30	12/19	30/36
4	2/6	32/37	4/8	34/39
5	1/16	3/25	9/24	11/33
6	3/17	16/31	8/22	21/36
7	5/19	14/35	9/23	18/39
8	5/16	10/30	11/22	17/36
9	3/21	3/23	15/33	15/35
10	2/8	14/24	15/21	27/35
Total	**32/136** **23.5%**	**153/301** **50.8%**	**96/200** **48.0%**	**218/360** **60.6%**

Variability of forms: single PDT item

Italics = not in NSs, but could be inferred

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)*

Gold standard difficulties

Recombination ⇒ unwanted triples in the gold standard set?

- Gold (NS): *wash(woman,shirt)*
- Gold (NS): *do(woman,laundry)*
- Recombined: *do(woman,shirt)*?

Matching semantics ≠ Matching nativeness?

- NNSs produce a wider range of forms to describe the prompts than NSs, e.g.,
  - NSs: overwhelmingly described *raking* action
  - NNSs: often described *cleaning* an area
- Related to issues of lexical gaps (Agustín Llach, 2010) & attaining native-like pragmatic usage (Bardovi-Harlig and Dörnyei, 1998)
- What counts as a correct meaning is application-specific

Summary & Outlook

Summary:

- Began process of examining ways to analyze semantics of learner constructions for interactive situations (PDT)
- Used existing parser & small set of extraction rules to obtain 92-93% extraction accuracy
- Learned that NS responses are probably not a good gold standard for evaluating NNS responses

Outlook:

- Implement automatic spelling correction
- Expand:
  - Beyond transitives
  - Handle type Z sentences (embedding, etc.)
  - More complex visual prompts (story retell, video description)
- Investigate ways to obtain a better gold standard

Acknowledgements

We would like to thank everyone who helped with this work:

- PDT help: David Stringer
- Recruiting help: Kathleen Bardovi-Harlig, Marlin Howard, Jayson Deese
- Feedback: Ross Israel, three anonymous reviewers, attendees of the IU Graduate Student Conference, CLingDing attendees

References

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

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.

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.

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.

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

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.