

Cohesion, Entrainment and Task Success in Educational Dialog

Diane Litman

*Computer Science Department
Learning Research & Development Center*

University of Pittsburgh

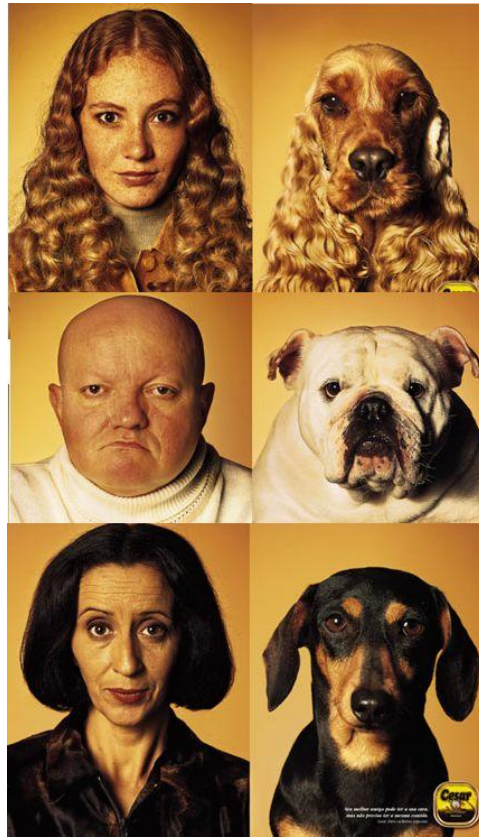


SIGdial 2012

Outline

- ◆ Cohesion and entrainment
- ◆ Measuring / relationships with task success
 - One-on-one tutoring dialogues
 - Multi-party student team dialogues
- ◆ From measuring to manipulating cohesion
- ◆ Summing up

*Joint work with Arthur Ward (tutoring, manipulating),
and with Heather Friedberg / Susannah Paletz (teams)*



Cohesion

- ◆ Related to SIGDIAL 2012 special theme:
characterizing dialogue *coherence*
 - general overall interrelatedness, continuity in meaning and context (Louwerse and Graesser, 2005)
- ◆ Measuring and using *cohesion*
 - Word-count similarity (Hearst, 1994)
 - Lexical (semantic) chains (Morris & Hirst, 1991; Barzilay & Elhadad, 1997)
 - Applications such as discourse segmentation

Entrainment in Human Dialogue

- ◆ **Existence:** Dialogue partners tend to become more similar in a variety of ways, e.g. prosody, speaking rate, word choice, syntax (Brennan & Clark, 1996; Reitter et al., 2006; Levitan and Hirschberg, 2011)
 - Entrainment, accommodation, adaptation, alignment, convergence, coordination, priming
- ◆ **Role:** Subconscious way that speakers work toward successful conversations (Pickering and Garrod, 2004)
 - Entrainment often correlates with task success (Reitter & Moore, 2007; Nenkova et al., 2008; Lee et al. 2010; Levitan et al., 2011)

Entrainment in Dialogue Systems

- ◆ Users also entrain to dialogue systems
 - Speaking rate (Bell et al., 2003) Influence speakers to slow down to improve speech recognition
 - Loudness, response latency, lexical choice (Brennan, 1996; Coulston et al. 2002; Darves & Oviatt, 2002; Gustafson et al., 1997; Levow, 2003; Parent & Eskenazi, 2010)
- ◆ And, benefits of systems entraining to users
 - Phrasal & lexical choice (Porzel et al., 2006)
 - Lexical /syntax (Stoyanchev & Stent, 2009; Lopes et al., 2011)
- ◆ Online communities (Huffaker et al., 2006)

This Talk

- ◆ Corpus-based measures of (multi-party) dialogue cohesion and entrainment
 - Computationally tractable
- ◆ Cohesion, Entrainment and...
 - Learning gains in one-on-one human and computer tutoring dialogues
 - Team success in multi-party student dialogues
- ◆ Towards manipulating cohesion in a tutorial dialogue system

Outline

- ◆ Cohesion and entrainment
- ◆ Measuring / relationships with task success
 - One-on-one tutoring dialogues
 - Multi-party student team dialogues
- ◆ From measuring to manipulating cohesion
- ◆ Summing up

Spoken Tutorial Dialogue Systems

◆ Why is one-on-one tutoring so effective?

*“...there is **something about discourse and natural language** (as opposed to sophisticated pedagogical strategies) that explains the effectiveness of unaccomplished human [tutors].” (Graesser et al. 2001)*

- What is the “something” ?
- Can it be computed automatically and in real-time?

◆ Our approach: Cohesion and entrainment

Theoretical Motivation

- ◆ Learning from text (McNamara & Kintsch, 1996)
 - Students read high and low coherence versions of a text
 - Relationship found between coherence and task success
- ◆ Interactive alignment (Pickering & Garrod, 2004)
 - Relationship between entrainment (e.g. lexical, syntactic, semantic) and shared mental models

Theoretical Motivation

- ◆ Learning from text (McNamara & Kintsch, 1996)
 - Students read high and low coherence versions of a text
 - Relationship found between coherence and task success

From text coherence to dialogue cohesion?

- ◆ Interactive alignment (Pickering & Garrod, 2004)
 - Relationship between entrainment (e.g. lexical, syntactic, semantic) and shared mental models

From mental models to student learning?

ITSpoke - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://rockhopper/itspoke/

ITSpoke

→

55. An airplane flying horizontally drops a packet when it is directly above the center of a swimming pool. Does the packet hit that spot? Explain.

Dialogue History:

Tutor said: Well... If an object has a non-zero constant velocity, is it moving or staying still?

You said: moving

Tutor said: Yep. If it's moving, then its position is changing. So then what will happen to the packet's horizontal displacement from the point of its release?

You said: it will change

Enter your essay here:

No, because the packet has a horizontal component to its velocity it will not drop perfectly vertical. the velocity in the horizontal direction will cause the packet to have a displacement horizontally from the point at which it was dropped.

- Back-end is [Why2-Atlas](#) system [VanLehn et al. 2002]
- [Sphinx](#) speech recognition and [Cepstral](#) text-to-speech

ITSpoke - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://rockhopper/itspoke/

ITSpoke

55. An airplane flying horizontally drops a packet when it is directly above the center of a swimming pool. Does the packet hit that spot? Explain.

Dialogue History:

Tutor said: Well... If an object has a non-zero constant velocity, is it moving or staying still?

You said: moving

Tutor said: Yep. If it's moving, then its position is changing. So then what will happen to the packet's horizontal displacement from the point of its release?

You said: it will change

Enter your essay here:

No, because the packet has a horizontal component to its velocity it will not drop perfectly vertical. the velocity in the horizontal direction will cause the packet to have a displacement horizontally from the point at which it was dropped.


- Back-end is [Why2-Atlas](#) system [VanLehn et al. 2002]
- [Sphinx](#) speech recognition and [Cepstral](#) text-to-speech

ITSpoke - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://rockhopper/itspoke/

ITSpoke



Dialogue History:

Tutor said: Well... If an object has a non-zero constant velocity, is it moving or staying still?

You said: moving

Tutor said: Yep. If it's moving, then its position is changing. So then what will happen to the packet's horizontal displacement from the point of its release?

You said: it will change

55. An airplane flying horizontally drops a packet when it is directly above the center of a swimming pool. Does the packet hit that spot? Explain.

Enter your essay here:

No, because the packet has a horizontal component to its velocity it will not drop perfectly vertical. the velocity in the horizontal direction will cause the packet to have a displacement horizontally from the point at which it was dropped.

- Back-end is [Why2-Atlas](#) system [VanLehn et al. 2002]
- [Sphinx](#) speech recognition and [Cepstral](#) text-to-speech

Two Types of Tutoring Corpora

◆ *Human Tutoring*

- 14 students / 128 dialogues (physics problems)
- 5948 student turns, 5505 tutor turns

◆ *Computer Tutoring*

- ITSPOKE v1
 - » 20 students / 100 dialogues
 - » 2445 student turns, 2967 tutor turns
- ITSPOKE v2
 - » 57 students / 285 dialogues
 - » both synthesized and pre-recorded tutor voices

ITSPOKE Experimental Procedure

- ◆ College students without physics
 - Read a small background document
 - Took a multiple-choice Pretest
 - Worked 5 problems (dialogues) with ITSPOKE
 - Took a multiple-choice Posttest
- ◆ Goal was to optimize Learning Gain
 - e.g., Posttest – Pretest

Measure I: Lexical Cohesion

(Ward & Litman, 2006)

- ◆ Lexical co-occurrence between speaker turns
- ◆ Motivated by Halliday and Hassan (1976)
 - Cohesion measured by counting “cohesive ties”
 - » Two words joined by a cohesive device (i.e. repetition)
 - Cohesive devices
 - » Exact word repetition
 - » Synonym repetition
 - » Near Synonym repetition
 - » Super-ordinate class
 - » General referring noun

Example

Speaker	Utterance	
Student Essay	No. The airplane and the packet have the same horizontal velocity. When the packet is dropped, the only force acting on it is g, and the net force is zero. The packet accelerates vertically down, but does not accelerate horizontally. The packet keeps moving at the same velocity while it is falling as it had when it was on the airplane. There will be displacement because the packet still moves horizontally after it is dropped. The packet will keep moving past the center of the swimming pool because of its horizontal velocity.	
ITSPOKE	Uh huh. There is more still that your essay should cover. Maybe this will help you remember some of the details need in the explanation. After the packet is released, the only force acting on it is gravitational force, which acts in the vertical direction. What is the magnitude of the acceleration of the packet in the horizontal direction?	
Cohesive Ties	Matches	Score
Stem, no stop	packet, horizont, onli, forc, act, acceler , vertic , there, will, still, after	11

Lexical Cohesion and Task Success

- ◆ Partial correlation performed between posttest & cohesion scores, controlling for pretest

	ITSPOKE v1 (20 students)		ITSPOKE v2a (34 students)	
	R	P-Value	R	P-Value
All Students	0.4	0.06	0.3	0.14
Low Pretest	0.7	0.01	0.5	0.05
High Pretest	0.6	0.13	0.1	0.66

- ◆ Students were split into Low and High Pretest groups using mean pretest score

Lexical Cohesion and Task Success

- ◆ Partial correlation performed between posttest & cohesion scores, controlling for pretest

	ITSPOKE v1 (20 students)		ITSPOKE v2a (34 students)	
	R	P-Value	R	P-Value
All Students	0.4	0.06	0.3	0.14
Low Pretest	0.7	0.01	0.5	0.05
High Pretest	0.6	0.13	0.1	0.66

- Cohesion is correlated with learning, but only for low pretest students
- Similar aptitude-treatment interaction in text comprehension (McNamara and Kintsch, 1996)

Measure II: Semantic Cohesion

(Ward & Litman, 2008)

- ◆ Semantic similarity between words across speaker turns
 - Extends cohesive devices
 - » Exact word repetition
 - » **Synonym repetition**
 - » **Near Synonym repetition**
 - » **Super-ordinate class**
 - » General referring noun
- ◆ Implemented using NLTK's Path Distance Similarity for WordNet
 - Allows new cohesive ties such as *man/person*, *acceleration/change*

Semantic Cohesion and Task Success

- ◆ Partial correlations again performed between posttest & cohesion scores, controlling for pretest

	ITSPOKE v1 (lexical cohesion)		ITSPOKE v1 (semantic cohesion)	
	R	P-Value	R	P-Value
All Students	0.5	0.04	0.5	0.05
Low Pretest	0.7	0.01	0.7	0.01
High Pretest	0.8	0.11	1.0	0.01

- New results for high pretesters
- Correlation strength varies with similarity thresholds

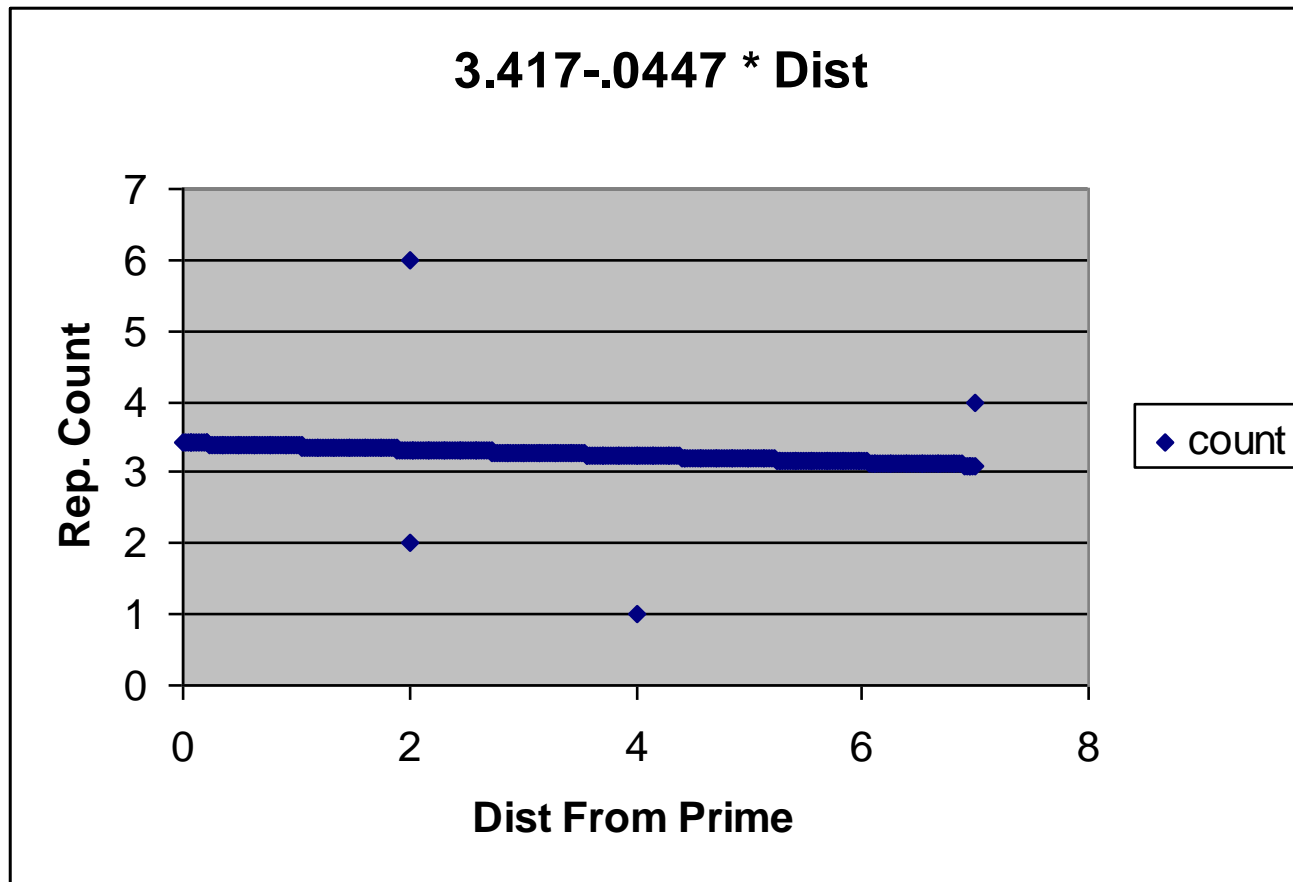
Measure III: From Cohesion to Entrainment (Ward & Litman, 2007a,b)

- ◆ Entrainment considers more temporal aspects of a dialogue than cohesion
- ◆ Many entrainment results are from controlled studies (e.g., Brennan & Clark, 1996)
- ◆ Recent work has also focused on the use of corpus-based measures
 - Syntactic priming (Reitter et al., 2006)
 - » Identify a “prime” in speaker’s utterance
 - » Look for response to the prime

Adapting Reitter's Measure to Lexical Entrainment

- ◆ Count word repetitions in response window following a prime
 - Prime: word occurrence in tutor utterance
 - Response window: next N student utterances
 - Data: up to N points per response window
 - » [distance from prime, repetition count]
- ◆ Linear regression
 - Fit line to [dist, count] points
 - Negative slope is evidence for convergence

Lexical Example – Fitted Line



The slope of the lines fitted to each student will be used to predict learning

Lexical Entrainment and Task Success

				Pretest Only		Full Model	
		Selected		Model	Adj	Model	Adj
N	Corpus	Features		pVal	R^2	pVal	R^2
14	Train HH	PreTest *	lex.w20	0.012	0.368	0.017	0.434
20	Test HC	PreTest **	lex.w20 **	0.044	0.162	0.003	0.431
		Significance code: p<.05:*; p<.01:**; p<.001:***					

- ◆ Tune lexical window size (to 20) using HH corpus, then evaluate slope feature on ITSPOKE corpus
- ◆ Adding lexical entrainment explains more variance (adjusted R^2 higher)

Lexical Cohesion vs. Entrainment

- ◆ Partial correlations again performed between posttest & cohesion scores, controlling for pretest

	ITSPOKE (lexical cohesion)		ITSPOKE (lexical entrainment)	
	R	P-Value	R	P-Value
All Students	0.4	0.06	-0.6	0.01
Low Pretest	0.7	0.01	-0.8	0.00
High Pretest	0.6	0.15	-0.4	0.33

- Use of entrainment provides new result for all students
- Larger absolute magnitudes for entrainment

Extending Reitter's Measure to Acoustic/Prosodic Entrainment

- ◆ Collect acoustic/prosodic values in response window following a prime.
 - Prime: tutor acoustic/prosodic *value above threshold*
- ◆ Linear regression
 - Fit line to [dist, *value*] points
 - Negative slope is evidence for convergence
- ◆ Result
 - Mean RMS slope (window size = 20) predicts learning

Outline

- ◆ Cohesion and entrainment
- ◆ Measuring / relationships with task success
 - One-on-one tutoring dialogues
 - Multi-party student team dialogues
- ◆ From measuring to manipulating cohesion
- ◆ Summing up

Theoretical Motivation

- ◆ Shared mental models (Johnson-Laird, 1980)
 - Model sharing can improve team effectiveness (Kozlowski & Ilgen, 2006)
 - Distinction between models about the task versus about the teamwork (Lim & Klein, 2006)

Theoretical Motivation

- ◆ Shared mental models (Johnson-Laird, 1980)
 - Model sharing can improve team effectiveness (Kozlowski & Ilgen, 2006)

Entrainment to operationalize sharing of models?

- Distinction between models about the task versus about the teamwork (Lin & Klein, 2006)

What kind of words are entrained on?

Corpus: Student Engineering Teams (Chan, Paletz & Schunn, LRDC)

- ◆ Pitt student teams working on engineering projects
 - Variety of group sizes and projects
 - “In vivo” dialogues
 - » Meetings over the semester were recorded in a specially prepared room in exchange for payment
- ◆ 10 high and 10 low-performing teams
- ◆ Sampled ~1 hour of dialogue / team (~43000 turns)



Corpus Excerpt (5 student team)

S2: These walls right here were only built for that 20 psi and that is not the same as what we got now so

S4: Who knows about, anything about materials?
Anybody?

S3: Like what?

S4: Just in general.

S1: At least the uh, the specialty materials.

Corpus Pre-processing

- ◆ Combined consecutive utterances from the same speaker to form turns
- ◆ Removed “questionable” transcriptions
 - Missing words
 - Speaker unknown or questionable

Measure I: Lexical Entrainment

◆ High-Frequency Words (Nenkova et al., 2008)

$$entr(w) = - \left| \frac{count_{S_1}(w)}{ALL_{S_1}} - \frac{count_{S_2}(w)}{ALL_{S_2}} \right|$$

- Overlap of most common words over the entire dialogue
 - » more of a similarity rather than an entrainment measure
- Significant correlation with task success

◆ Note that since $entr(w)$ doesn't depend on individual turns, appropriate for our noisy corpus

Computing Pair Entrainment

- ◆ Calculated high frequency words
 - Over entire corpus or over just a session
 - Top 25 or Top 100
- ◆ For each pair of speakers, calculated entrainment for these words, then summed

$$entr(w) = - \left| \frac{count_{S_1}(w)}{ALL_{S_1}} - \frac{count_{S_2}(w)}{ALL_{S_2}} \right|$$

$$ENTR_1(c) = \sum_{w \in c} entr(w)$$

From Pair to Groups

- ◆ Combined multiple pair scores into one group score
 - Average
 - Weighted Average (by proportion of turns by speakers)
 - Max (one good pair helps the whole team?)
 - Min (one bad pair hurts the whole team?)
 - Range

Group Entrainment and Task Success

- ◆ Use success scores ($0 \leq \text{score} \leq 1$) directly
 - Pearson correlation between entrainment and success
- ◆ Split teams into two groups - “high” (score $> .7$) and “low” (score $< .5$)
 - T-test to test for significant difference between average group entrainment of teams in the two groups

Group Entrainment and Task Success

- ◆ Use success scores ($0 \leq \text{score} \leq 1$) directly
 - Pearson correlation between entrainment and task
- ◆ Split teams into two groups - “high” (score $> .7$) and “low” (score $< .5$)
 - T-test to test for significant difference between average group entrainment of teams in the two groups
- ◆ *No significant results*

Measure II: Change in Lexical Entrainment; Project Words

- ◆ Project Words (Litman et al., 2009)
 - Computed from the project descriptions for each team
- ◆ Change in entrainment over a single session
 - Second half score - first half score
 - » Note: raw score, not normalized by number of project words
 - Similar to convergence tests (Levitan and Hirschberg, 2011)

Entrainment Change & Task Success

- ◆ Use task success scores of student teams directly

Method for Combining Student Pairs	Pearson Correlation between Team Success and Group Entrainment Change	P-value
Average	.207	.083
Weighted Average	.198	.099

- The trending positive correlations support the hypothesis that entrainment happens more in successful teams

Entrainment Change & Task Success

- ◆ Split teams into “low” and “high” success group

Method for Combining Student Pairs	Group Entrainment Change for Low Teams	Group Entrainment Change for High Teams	P-value
Average	-.030	+.022	.034
Weighted Average	-.028	+.020	.044

- The significant differences again support the hypothesis that entrainment happens more in successful teams
 - » High success teams increase their project word entrainment
 - » Less successful teams diverge in their use of project words

Outline

- ◆ Cohesion and entrainment
- ◆ Measuring / relationships with task success
 - One-on-one tutoring dialogues
 - Multi-party student team dialogues
- ◆ From measuring to manipulating cohesion
- ◆ Summing up

Theoretical Motivation

- ◆ *Reflection*: a “process by which students derive abstractions by comparing multiple performances simultaneously” (Collins and Brown, 1986)
- ◆ Reflection after tutoring
 - Improved learning from a quantitative physics computer tutor (Katz et al. 2003, 2007)
 - Reading a text was as effective as engaging in a (typed) dialogue

Theoretical Motivation

- ◆ *Reflection*: a “process by which students derive abstractions by comparing multiple performances simultaneously” (Collins and Brown, 1986)
- ◆ Reflection after tutoring
 - Improved learning from a quantitative physics computer tutor (Katz et al. 2003, 2007)

Same results for a qualitative physics dialogue tutor?

- Reading a text was as effective as engaging in a (typed) dialogue

Does the cohesion of the reflective text matter?

Manipulating Cohesion in Reflection after Dialogue (Ward & Litman, 2011a,b)

- ◆ Revised ITSPOKE experimental procedure
 - Read a small background document
 - Take a multiple-choice Pretest
 - Work 5 problems (dialogues) with ITSPOKE
 - Read a post-dialogue text, either:
 - » the background document again
 - » a high cohesion reflective text
 - » a low cohesion reflective text
 - Take a multiple-choice Posttest

Authoring the Reflective Texts

High vs. Low cohesion versions were constructed following (McNamara & Kintsch, 1996)

- ◆ **High Cohesion:** In these two **problems** we used **Newton's Third Law** to show that the forces involved in an action/reaction pair had the same magnitude but acted in opposite directions **to each other**. An action-reaction pair is formed whenever one object exerts a force on a second object. **Newton's Third Law** says that when **one object exerts a force on a second object**, there is an equal and opposite reaction force from the second object back onto the first object. In addition, the type of force is always the same for both objects in the action/reaction pair. **For example** it was gravitational force on both Earth and Sun, and impact force on both car and truck.
- ◆ **Low Cohesion:** In both **situations** we used **the Third Law** to show that the forces involved in an action/reaction pair had the same magnitude but acted in opposite directions **[]**. An action-reaction pair is formed whenever one object exerts a force on another object. **Newton's Third Law** says **this force** will have an equal and opposite reaction force. The type of force is always the same for both objects in the pair. **[]** It was gravitational on both Earth and Sun, and impact on both car and truck.

Results: Reflective cohesion affects learning

- ◆ Reflection is a useful addition to dialogue tutors
 - Reading a reflective text improves learning
 - High cohesion is better, in general
 - Low cohesion is better for certain subgroups
 - » who aren't likely to perform inference anyway, for lack of high knowledge or motivation
- ◆ Current work (Rimac system): dynamically generate dialogue that adapts the level of cohesion (and coherence) to the needs of the user (Katz et al., 2011)

Outline

- ◆ Cohesion and entrainment
- ◆ Measuring / relationships with task success
 - One-on-one tutoring dialogues
 - Multi-party student team dialogues
- ◆ From measuring to manipulating cohesion
- ◆ Summing up

Summing Up

- ◆ Participants in educational dialogues (involving either humans or computers) show similarity in both what they say, and how they say it
 - Lexical and semantic cohesion in one-on-one tutoring
 - Lexical and acoustic-prosodic entrainment in tutoring
 - Lexical entrainment during team conversations
- ◆ Increased dialogue similarity is associated with dialogue success
 - Learning gains in tutoring
 - Team success in design

Current Directions

- ◆ Monitoring and dynamically manipulating dialogue similarity
 - Rimaac dialogue system
- ◆ Multi-party dialogue entrainment
 - Weighting pairs via other speaker attributes (e.g. gender, domain expertise)
 - Entrainment on other conversational features (e.g. acoustic-prosodic, syntactic, on versus off-task)

Acknowledgements

- ◆ ITSPOKE group
- ◆ NLP@Pitt (*R. Hwa, J. Wiebe et al.*)
- ◆ Why2-Atlas and Rimac Tutoring groups
(*K. VanLehn, P. Jordan, S. Katz et al.*)
- ◆ Engineering Design group (*C. Schunn et al.*)



Thank You!

- ◆ Questions?

- ◆ Further Information

 - <http://www.cs.pitt.edu/~litman/itspoke.html>