(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

Does Seeing One Another's Gaze Affect Group Dialogue? A Computational Approach

Bertrand Schneider and Roy Pea
Stanford University, USA
schneibe@stanford.edu, roypea@stanford.edu

ABSTRACT: In a previous study, we found that real-time mutual gaze perception (i.e., being able to see the gaze of your partner in real time on a computer screen while solving a learning task) had a positive effect on student collaboration and learning (Schneider & Pea, 2013). The goals of this paper are 1) to explore a variety of computational techniques for analyzing the transcripts of student discussions; 2) to examine whether any of those measures sheds new light on our previous results; and 3) to test whether those metrics have any predictive power regarding learning outcomes. Using various natural language processing algorithms, we found that linguistic coordination (i.e., the extent to which students mimic each other in terms of their grammatical structure) did not predict the quality of student collaboration or learning gains. However, we found that a simple computational measure of student verbal coherence (i.e., the extent to which students build on each other's ideas) was positively correlated with their learning gains. Additionally, this measure was significantly different across our experimental conditions: students who could see the gaze of their partner in real time were more likely to develop a coherent discussion. Finally, using various language metrics, we were able to roughly predict (i.e., using a median-split) learning gains with a 94.4% accuracy using Support Vector Machine. The accuracy dropped to 75% when we used our model on a validation set. We conclude by discussing the benefits of using computational techniques on educational datasets.

Keywords: Natural language processing, eye-tracking, learning analytics, computer-supported collaborative learning

Editor's Note: As part of the Special Section on Learning Analytics & Learning Theory this article is followed by a short commentary on pp. 134-137 that discusses the challenges it faced and successes it achieved in drawing on and contributing to theory use in learning analytics.

1 INTRODUCTION

Despite recent efforts in developing automated ways to analyze student discourse, educational research primarily relies on time-consuming traditional methods to analyze student transcripts, including qualitative analyses and the development of manual coding schemes. Yet the field of Natural Language Processing (NLP) has significantly gained in maturity over the past decades, and computational techniques can be advantageously applied to educational datasets. Recent efforts in topic modelling, for instance, seem especially promising for gaining insights into student discourse and cognitive processes (Sherin, 2012). Unfortunately, social scientists willing to learn those tools are a rare breed, and collaborative multidisciplinary work among educational researchers and computer scientists is slow to appear. In this paper, we describe our attempt at applying NLP techniques to educational transcripts. In a previous study, we investigated the effect of *mutual visual gaze perception* on student collaborative learning; we found that students who could see the gaze of their partner in real time on a computer

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

screen were more likely to develop a high-quality collaboration and to learn more from the task. In this paper, we investigate the effect of this gaze awareness tool on student discourse. More specifically, we looked at students' linguistic coordination (i.e., the extent to which students mimic each other in terms of their grammatical structure) and coherence (i.e., the extent to which students build on each other's ideas), and investigated their predictive value for estimating learning gains.

In the next section, we briefly review the literature on joint visual attention (with a heavier focus on eye-tracking studies) and its relationship to student discourse. We then describe our dataset, introduce each of our measures, apply them to our transcripts, and present our findings from using this approach. Additionally, we trained a machine-learning algorithm using all those features and attempted to predict student learning. We conclude by discussing our findings, and sketch several applications where these results could be exploited in a classroom setting.

2 LITERATURE REVIEW: WHY IS JOINT VISUAL ATTENTION AN IMPORTANT EDUCATION CONSTRUCT?

Joint Visual Attention (JVA) has been extensively studied by developmental psychologists, learning scientists and social psychologists. This foundational line of work highlights the crucial role of JVA in any kind of social interaction for learning. Babies learn to speak a language in part by establishing joint visual attention with their caregivers and associating a particular sound with the focus of visual regard (Stern, 1977). Parents signal important features of the environment to their children by pointing at them using deictic gestures (Bates, Thal, Whitesell, Fenson, & Oakes, 1989). Students' ability to achieve joint attention is shown to be associated with better coordination among members of collaborative problem-solving groups (Barron, 2003). This last example is central to the analyses presented in this paper. We know from socio-constructivist perspectives that powerful learning can happen when students build on each other's ideas (i.e., so-called "transactive discourse" [Stahl, 2013] or "coherence"); but we also know that coordination between students is difficult to achieve (Salomon & Globerson, 1989). In a foundational study in the Learning Sciences, Roschelle (1992) provides a thick description of how a student dyad goes through *convergent conceptual change*: he demonstrates that the management of joint attention is a pre-requisite for the establishment of a common problem space, and how ambiguous utterances laden with anaphoras (such as "It's like the line. Fat arrow is the line of where it pulls that down," reproduced from Roschelle's transcript) can actually be important stepping-stones toward conceptual convergence when joint attention is present. Without joint visual attention, there is little chance that groups of students would be able to build a common ground to co-construct knowledge and make sure that they are actually discussing the same concepts or entities. Thus, our work is based on the observation that successful collaborative learning episodes are the results of the complex interplay between coordinated visual attention and transactive discourse. Based on this observation, our goal is to capture visual and verbal synchronization using computerized means and to analyze how those two types of coordination interact with each other. In the paragraph below, we describe some of the previous work that attempted to capture those two measures of convergence using computational

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

methods.

Traditional work on JVA has been qualitative in nature, requiring researchers to painstakingly annotate scores to hundreds of hours of videos (e.g., Barron, 2003; Roschelle, 1992). Over the past decade, however, this work has experienced a rebirth with the advent of eye-tracking technology, as researchers have started to use synchronized eye-trackers to capture subjects' attentional processes and develop computational measures of JVA (sometimes called *gaze recurrence*). Richardson, Dale, and Kirkham (2007), for instance, found that the degree of gaze recurrence between individual speaker–listener dyads (i.e., the proportion of times that their gazes were aligned) was correlated with the listeners' accuracy on comprehension questions; they also found that humans usually need +/- 2 seconds to disengage from their current thought process before being able to attend to their partner's offer to establish joint attention (Richardson & Dale, 2005). Jermann, Mullins, Nüssli and Dillenbourg (2011) studied pairs of programmers in a dual eye-tracking setting and observed that productive collaborations were associated with a higher level of gaze recurrence: more JVA was associated with successfully building a common ground and with students' ability to sustain mutual understanding. Cherubini, Nüssli, and Dillenbourg (2008) designed an algorithm to detect misunderstanding in a remote collaboration by computing the distance between the gaze of the speaker and the listener from eye-tracking log files. They found that the likelihood of misunderstandings increased with gaze dispersion. Brennan, Chen, Dickinson, Neider, & Zelinsky (2008) investigated the effect of shared gaze and speech during a spatial search task; the shared gaze condition was the best of all, twice as fast and efficient as solitary search, and significantly faster than other collaborative conditions. Finally, researchers have also started to investigate the relationship between student gaze coordination and discourse. Jermann, Gergle, Bednarik & Brennan (2012) for instance, describe a taxonomy for four types of attentional similarity between participants of a collaborative task computed from eye-tracking logs (broad versus narrow; stable versus changing over time), and show how they relate to different types of dialogue. They found, for instance, that conceptual discussions with reference to instructional material are more likely to be found in episodes of stable focus, and that meta-cognitive episodes are more frequent in unstable broad focus. Thus, there is some evidence suggesting that types of visual coordination are associated with different types of verbal interactions. This type of study provides us with preliminary evidence that combining dual eye-tracking data and student utterances offers a useful methodology for investigating transactive discourse.

In summary, we can confidently say that studies on JVA have proven their productivity for investigating collaborative processes among students. Recent work is advancing this idea by adopting a computational approach and by using eye-tracking technology; additionally, researchers have started investigating the relationship between subjects' gaze and utterances. We contribute to this nascent line of research by exploring the effect of a gaze-awareness tool (i.e., being able to see the gaze of your partner on a computer screen in real time) on student dialogue, and by exploiting common techniques used in Natural Language Processing (NLP). More specifically, previous work on student multimodal convergence informed the exploration of our dataset: we expect to see positive relationships between

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

students' visual/verbal coordination and outcomes of interests (e.g., students' quality of collaboration and learning gains). In the study we conducted, we also found that the ability to see the gaze of one's partner had a beneficial impact on student collaboration, most notably by increasing their levels of joint visual attention; thus, our exploration of this dataset will also include comparisons with a control group that did the exact same task but could not see the gaze of their partner. Based on prior work, we expect to see higher levels of convergence in the treatment group.

In the next section, we discuss our experiment and dataset and describe our approach and research questions in detail.

3 THE CURRENT DATASET

In previous work (Schneider & Pea, 2013), we conducted a study on the effect of *mutual gaze perception* on student collaborative problem-solving processes; student dyads were asked to collaborate remotely on a set of diagrams to discover how the human brain processes visual information. Each student was located in a different room, and could communicate with his/her partner via audio. The information on the screen was similar for both participants (i.e., the brain diagrams shown in Figure 1). The stages of the activity were as follows: 1) students analyzed brain diagrams (12 minutes) and tried to associate visual impairments with particular lesions (two answers out of five were given, i.e., the top left and top right diagrams); 2) they were asked to read a textbook chapter about human vision and discuss their understanding of this topic (12 minutes). Before the first activity and after the reading task, students were also asked to complete a learning test (pre- and post-questionnaires).

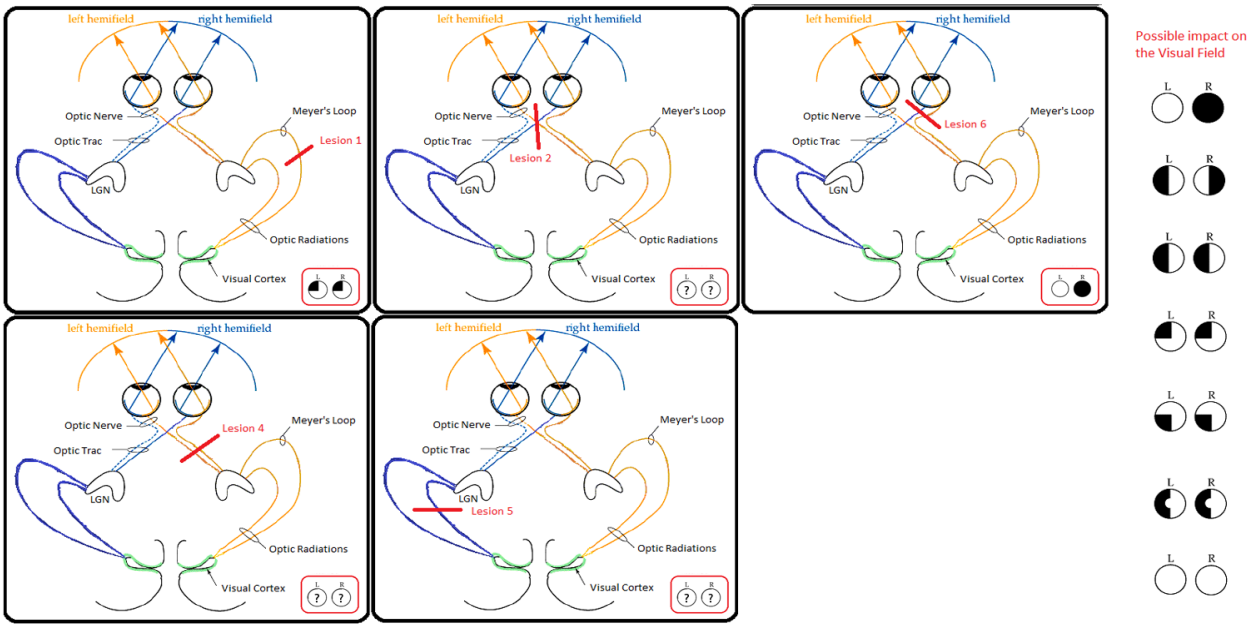

Figure 1. Diagrams that students had to analyze. Five contrasting cases show the visual pathways of the human brain; students had to identify the effect of each lesion on the visual field.

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

The pre- and post-test contained 15 questions: five terminology questions (participants were asked to provide the name of a specific brain region or pathway), five conceptual questions (participants had to predict the effect of a specific lesion), and five transfer questions (subjects were asked to use their new knowledge to solve a vignette; e.g., "patient X is likely to have a lesion in region Y of the brain; should he be allowed to drive?"). The learning tests were administered on a computer and were coded automatically for the first two categories (terminology and conceptual questions). Transfer questions were open-ended and coded by two different researchers.

Half of our participants were assigned to an experimental group ("visible-gaze") where they could see the gaze of their partner displayed in real time on a screen. To achieve this, we used two Tobii X1 eye-trackers running at 30Hz to record student gaze (shown as a transparent blue dot on the screen). In a control group ("no-gaze"), the other half of our participants did not have access to this visualization. This intervention helped dyads in the first group achieve higher learning gains (Figure 2, left side): $F(1,19)=9.19$, $p<0.001$ (for the "visible-gaze" group, mean=0.49, SD=0.08; for the "no-gaze" group, mean=0.38, SD=0.08). They also achieved a higher quality of collaboration (as measured by the coding scheme developed by Meier, Spada, & Rummel, 2007): $F(1,19)=11.73$, $p<0.01$, Cohen's $d=1.24$ (mean for the treatment group=0.89, SD=0.48; mean for the control group=-0.08, SD=0.79). The right side of Figure 2 shows an overall score for student collaboration by summing their score on the rating scheme's nine dimensions: sustaining mutual understanding, dialogue management, information pooling, reaching consensus, task division, time management, technical coordination, reciprocal interaction, and individual task orientation.

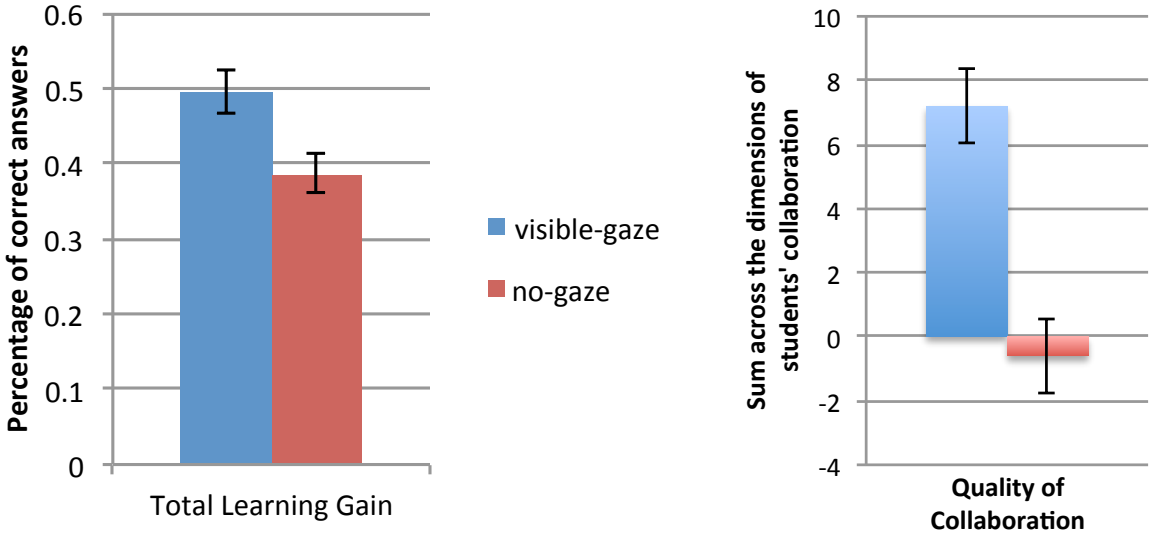

Figure 2. Learning Gains (left) and Collaboration Quality (right) for the two experimental groups (p<.01)

We also recorded student gaze and discourse during the task. By analyzing the eye-tracking data, we found that participants in the experimental condition had more moments of joint attention (i.e., they were more likely to look at the same diagram at the same time on the screen), and this measure was significantly correlated with positive learning gains ($r(37)=0.39$, $p<0.05$). This result reinforced the

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

conjecture that joint visual attention is a crucial mechanism for coordinating social interactions (Tomasello, 1995). A subsequent qualitative analysis of the videos suggested that our intervention helped students because 1) they were able to anticipate what their partner was about to say, because *they could already see the location of their partner's gaze on the screen*; 2) *they could use gaze as a pointer to complement their discourse*, obviating the need to mention locations explicitly on the diagrams; and finally, 3) they could monitor the visual activity of their partner at all times, providing an aid to establishing common ground.

We propose to use computational techniques to further illuminate this dataset. More specifically, we are interested in exploring three aspects of student dialogue:

1. Are there ways to characterize the effect of our intervention on student discourse?
2. Is it possible to find markers of productive learning trajectories?
3. Is it possible to find markers of constructive collaborations?

More specifically, those questions are in line with the micro-genetic method developed by Siegler and Crowley (1991): we seek to isolate micro-behaviours associated with positive collaborative learning outcomes. Tactically, we can answer the first question by designing linguistic metrics and running statistical tests (i.e., an ANOVA) between our two experimental conditions. The second and third questions can be answered by running correlations between our measures of interest, learning gains and collaboration scores. Finally, in the last section of this paper, we explore whether we can exploit the predictive value of those measures to train a supervised machine-learning algorithm. More specifically, we performed a median-split on student learning gains and tested the accuracy of our algorithm on this classification task (i.e., whether a particular student would be above or below the median split) using linguistic features.

4 NATURAL LANGUAGE PROCESSING AND MUTUAL GAZE PERCEPTION

In the next sections, we describe the measures we developed to provide a preliminary answer to those questions. First, we looked at unigram, bigram, and trigram counts to build categories of interest using a "bag of words" model. Next, we looked at the coordination of linguistic styles among dyads: are students more likely to mimic the grammatical structures of their peers in a good collaboration (as suggested by Danescu-Niculescu-Mizil & Lee, 2011)? We then assessed the *coherence* of student discourse (also called transactivity; see Stahl, 2013), by comparing the similarity of consecutive sub-sections of the transcripts; our goal was to evaluate the extent to which students build upon each other's ideas during the analysis task. Finally, we gathered all the previous measures and ran a machine-learning algorithm (Support Vector Machine) to roughly predict student learning gains (i.e., was a particular student above or below the median, in terms of their learning gains?).

4.1 Description of the Dataset

It should be mentioned that all the analyses below are performed on the brain diagram discussions, since students talked very little when they were reading the textbook chapter. We provide in Table 1

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

some descriptive statistics on the transcripts we analyzed.

Table 1. Descriptive statistics of transcripts (mean with standard deviation in parentheses). No statistical significance between experimental groups (p-value>.05).

Condition	Number of sentences	Number of words	Number of turns	Average sentence length
"visible-gaze"	108.23 (35.29)	823.18 (317.50)	490.29 (182.16)	65.32 (21.91)
"no-gaze"	88.55 (36.92)	636.75 (332.37)	398.13 (193.25)	57.00 (23.35)

From Table 1, we can see that students in the "visible-gaze" group seemed to talk more, build longer sentences, and have more turn-taking. None of those differences were statistically different across our two experimental groups for all those measures, however (p-value>.05).

4.2 N-Grams

To get a preliminary sense of our dataset, we first computed unigram, bigram, and trigram probabilities. This helped us to understand which words were frequently used in our two experimental groups, and allowed us to build relevant categories for grouping our n-grams. For instance, we observed that the word "look" was positively correlated with learning gains ($r(37)=0.42$, $p=0.008$), which can be associated with either the content to be learned (i.e., the brain diagrams showing how visual information is processed by the human brain) or a verbal indication to share visual information (e.g., "look at my gaze!"). However, we did not conduct in-depth analyses of the unigrams alone, because they were difficult to interpret: as the example indicates, unigrams are often ambiguous, and bigrams or trigrams were so rare in our dataset that they did not provide strong evidence for any type of hypothesis (the most common bi-gram found in our transcripts was "the cut" and was only mentioned 9 times in all the transcripts; similarly, the most common trigram was "cut in the" and was only mentioned 7 times across all groups). This is why we decided to group unigrams by categories instead of looking at bigrams/trigrams or analyzing them in isolation. As a first pass, we decided to create those categories based on common sense: a researcher looked at the 200 most common words and manually created groups of words that seemed to relate to a common topic.

As a side note, we also experimented with automatic approaches before manually coding unigrams. We followed Sherin's methodology for isolating topics in our transcripts. Unfortunately, no apparent pattern was revealed in the topics by using this approach, suggesting that either our dataset was too small, or that topics in student utterances were not stable enough to be detected. Either way, we were convinced to adopt a manual approach instead and grouped unigrams by looking at the most common words used in our transcripts.

For instance, the category "*anaphora"* subsumed the words "it," "some," "that," "which," "each," "few" and so on; the category "*conceptual discussion*" included "think," "cause," "because," "suppose,"

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

"impact," and so on. Table 2 shows the final 8 categories constructed from our dataset. We agree that those groups were built in an arbitrary manner, and that some words could belong to several categories. Nonetheless, our approach was data-driven — in the sense that we used the most common words from our dataset — and theory-driven, in that we designed potential indicators for collaborative learning. For instance, the category "*conceptual discussion"* is likely to be associated with higher learning gains, and the category "*anaphora"* is likely to be associated with a higher quality of collaboration. Why? Because this measure can serve as a proxy for measuring the quality of common ground between two participants: since anaphoras are ambiguous by nature, to be correctly interpreted by the interlocutor will require a stronger coordination between students. Herbert Clark has developed a considerable body of work investigating this topic (e.g., Clark & Brennan, 1991).

Table 2. Categories built on common unigrams.

Category	Unigrams
Jargon	hemi, field, hemifield, brain, eye, lesion, optic, vision, track, gaze, nerve, hemisphere, loop, information, blind, radiation, meyer, LGN
Diagram	blue, orange, case, circle, box, yellow, line, arrow, white, black, circle, number, half
Location	right, middle, left, top, bottom, diagram, opposite, corner, side, down, underneath, back, inner, outer, between, toward, lower, here, there, first, second, third, fourth, fifth, one, two, three, four, five
Conceptual discussion	think, cause, because, since, change, figure, would, wouldn't, impact, affect, explain, suppose, interpret
Uncertainty	maybe, possible, though, but, know, could, guess
Anaphora (person)	anybody, anyone, both, each, each other, everybody, everyone, he, her, hers, herself, him, himself, his, I, it, its, itself, me, mine, myself, neither, nobody, others, ours, ourselves, several, she, somebody, someone, their, theirs, them, themselves, they, us, we, who, whoever, whom, whomever, whose, you, your, yours, yourself, yourselves
Anaphora (thing)	all, another, anything, both, each, each other, everything, few, it, its, itself, most, much, neither, one, none, nothing, one another, other, others, several, some, something, that, these, this, those, what, which

Participants in the experimental group used more anaphoras (referring to a "thing") compared to participants in the control group: $F(1,41)=4.88$, $p=0.03$. Our results suggest that *real-time mutual gaze perception* may be a way to support establishing common ground. The findings indicate that participants in the *real-time mutual gaze perception condition* were able to exploit this information to the extent that they could employ ambiguous anaphora, realizing that the pointing manifested by their partner's gaze would disambiguate the referent of their speech act. Additionally, there appears to be a trend showing that more conceptual discussion occurred in the "visible-gaze" group (Figure 3, right side): $F(1,41)=5.52$, $p=0.02$. One limitation of this measure is that the number of words representing this construct is relatively small (between 0 and three words used every minute). The other categories did not yield any significant effect.

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

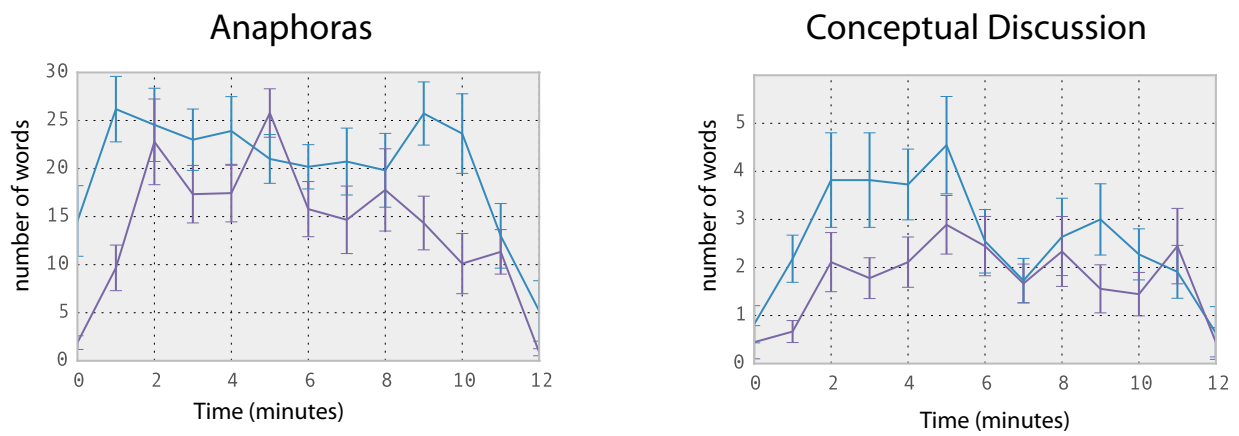

Figure 3. Evolution of word frequency related to Anaphora and Conceptual Discussions over time. The blue line corresponds to the "visible-gaze" group; the purple line corresponds to the "no-gaze" group.

Even with these limitations, it is interesting to see that categories built on n-gram frequencies can offer a new window into student collaborative learning processes. Figure 3, for instance, shows the aggregate values of those two groups for each minute of the activity; one can imagine similar graphs for individuals, as a way to guide qualitative analysis. In the next section, we employ algorithms from the field of information retrieval to further explore the differences between our experimental groups.

4.3 Coordination of Linguistic Styles (Convergence)

Computing n-gram counts and probabilities is an interesting way to look at student discussions. But it does not contribute to our understanding of the linguistic patterns used in collaborative learning discussions; it merely describes word counts for a particular time slice and experimental group. To address this limitation, we examined the ways in which students progressively build a discourse around the instructional material. Specifically, we looked at a particular phenomenon in social interactions called the *chameleon effect*, according to which, in a social setting, people tend to mimic their interlocutor's grammatical structure (Danescu-Niculescu-Mizil & Lee, 2011), as in this example:

Doc: At least you were outside.
Carol: It doesn't make much difference where you are [...]

Note that "Carol" used a quantifier different than the one "Doc" employed. Also, notice that "Carol" could just as well have replied without including a quantifier, for example: "It doesn't really matter where you are..." (p. 1, right column)

Using two large datasets (movie dialogues and Twitter), Danescu importantly shows that this effect (called *convergence*) is relatively robust and pervasive — people tend to mimic their interlocutor's grammatical structure consistently. Previous research suggests that this convergence is associated with

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

enhanced communication in organizational contexts and in psychotherapy (cited in Danescu-Niculescu-Mizil & Lee, 2011). Our goals are 1) to replicate Danescu's results on our dataset, and 2) to test whether *mutual visual gaze perception* supports convergence. Concretely, Danescu used 9 categories from the LIWC corpus (Linguistic Inquiry and Word Counts; Pennebaker, Francis, & Booth, 2001) to compute convergence measures. Those categories are as follows: articles, auxiliary verbs, conjunctions, high-frequency adverbs, impersonal pronouns, negations, personal pronouns, prepositions, and quantifiers. The way convergence is computed is straightforward:

$$P(b^t_{\hookrightarrow a} = 1 | a^t = 1) - P(b^t_{\hookrightarrow a} = 1).$$

The first expression is the conditional probability of seeing word type *t* expressed by person *b* in answer to person *a*, given that person *a* used this word type in the prior utterance. The second expression is just the probability of seeing a particular word type expressed by person *b* in answer to person *a* in the entire corpus. Subtracting the second expression from the first one gives us a measure of *convergence*.

Figure 4 (left side) shows Danescu's results for his dataset. Error bars are flat and barely visible (shown in red) because his dataset is relatively large; dark blue bars show the probability of using a particular word type (e.g., articles, pronouns) and light blue bars show the conditional probability of using a particular word type, given that an interlocutor used the same word type in the prior utterance. Figure 5 shows our replication of Danescu's results. Observe the same pattern emerging: light blue bars (conditional probability that a certain category of words is mirrored by the same word type in the interlocutor's response) are always higher than the probabilities of this type of word in the corpus. Due to our smaller corpus, not all differences are statistically significant, but most of them are (i.e., where the standard errors do not overlap).

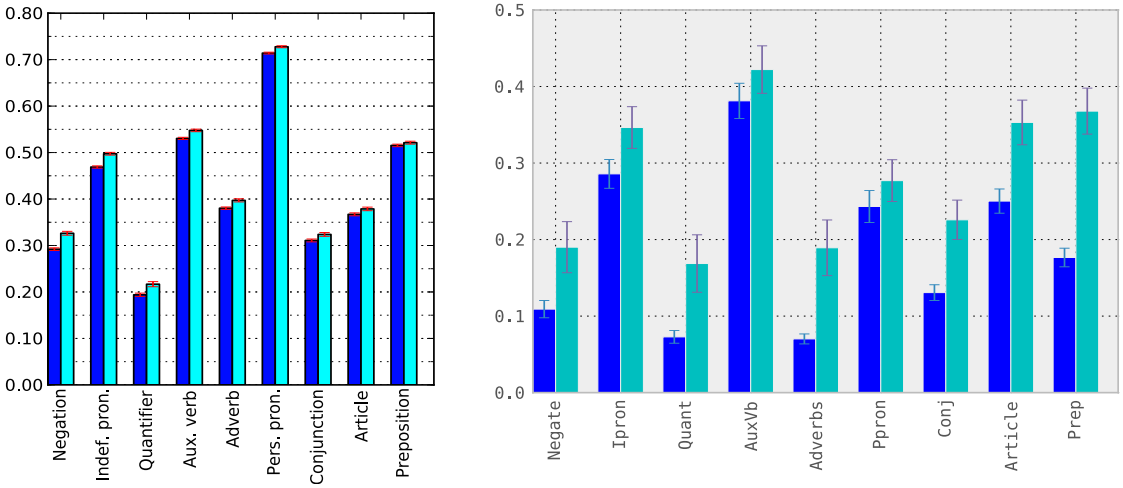

Figure 4. *Left side:* Danescu-Niculescu-Mizil & Lee's (2011) graph shows how people tend to mimic the grammatical structure of their interlocutor. Light blue bars show the conditional probability of using a particular word type, given that an interlocutor used it in the prior utterance. Dark blue bars show the probability of using a particular word type in the entire corpus. *Right side:* Danescu's results replicated on our dataset. Whiskers show standard errors; non-overlapping bars show significant differences.

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

Most importantly, there was special potential in using this measure to discriminate between the two experimental groups (e.g., "visible-gaze" versus "no-gaze"; productive versus poor collaborators; good versus poor learners). Contrary to our prediction, there was no significant difference between those groups on our convergence measure ($F<1$), meaning that coordination of linguistic styles is not predictive of positive learning gains, at least for our corpus. This result also indicates that *mutual gaze perception* doesn't influence this effect: students are not more likely to imitate each other's grammatical patterns if they can see the gaze of their partner in real time.

This convergence measure, however, only looks at superficial features of collaborative dialogues (i.e., word types). As it would be more compelling to look at the words themselves, in the next section, we explore whether productive students are more likely to mimic the *content* mentioned by their partner.

4.4 Building on Your Partner's Ideas (Coherence)

In this section, we describe how we summarized our data in a very high dimensional space, separated the transcripts in several consecutive segments, and applied cosine similarity metrics to measure student coherence. A cosine similarity score characterizes the similarity of two text documents (or transcript subsections). Our approach was to segment student transcripts into smaller texts and compute similarity measures between those subsequent segments. By iteratively repeating this procedure, we can evaluate the *coherence* of a discussion (for more details about linguistic alignment, see Pickering & Garrod, 2004). The idea behind *coherence* is that interlocutors tend to adapt to the patterns expressed in each other's utterances. For instance, a group of students discussing the inner working of the human brain will tend to reuse the same kinds of words when they build on each other's ideas. This alignment, in turn, is believed to indicate shared understanding (or common ground). Ward and Litman (2007), for instance, showed that coherence was predictive of learning in tutoring dialogues. There has been a significant amount of additional work on this topic, in various domains. We will not summarize the literature on coherence, but the interested reader can review work done around Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004). Since this is an exploratory study, we compute here only one aspect of a discussion's coherence (i.e., word repetition). Thus our measure does not fully reflect the concept of coherence, as linguists describe it. We also acknowledge that there are several ways of capturing word repetition (e.g., explicit semantic similarity) that are not used in the analyses below. We plan to compare those different methods in future work.

The first step was to apply tf-idf transformations (term frequency–inverse document frequency) to our dataset. Tf-idf is commonly used to summarize a text corpus. The value of highly frequent words is decreased, and is offset by their frequency in the corpus; this way, rare words gain a bigger weight and common words (e.g., "the," "it") gain a smaller weight. This technique is used in the field of information retrieval (Manning, Raghavan, & Schütze, 2008, p. 6) to score a document's relevance to a query. But before we computed a measure of student coherence, we first compared each student's discourse similarity with other participants by using a cosine similarity measure over the *entire* transcript. A cosine

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

similarity measure takes two vectors and computes the magnitude of the angle between them to represent their similarity. We show every pairwise comparison in Figure 5: dark blue lines show students who are very dissimilar to everyone else; hot colours represent similarity. As a sanity check, we can observe that students are identical to themselves (red diagonal). Students in the same group are next to each other on each axis; we can see that students belonging to the same group tend to resemble each other (2x2 squares along the diagonal). Finally, we can isolate students who are very different from everyone else (e.g., P62 and P63) and try to explain why they are very distinct from other participants: in our case, P63 achieved the lowest learning gain after the activity. P62 was within one standard deviation of the mean.

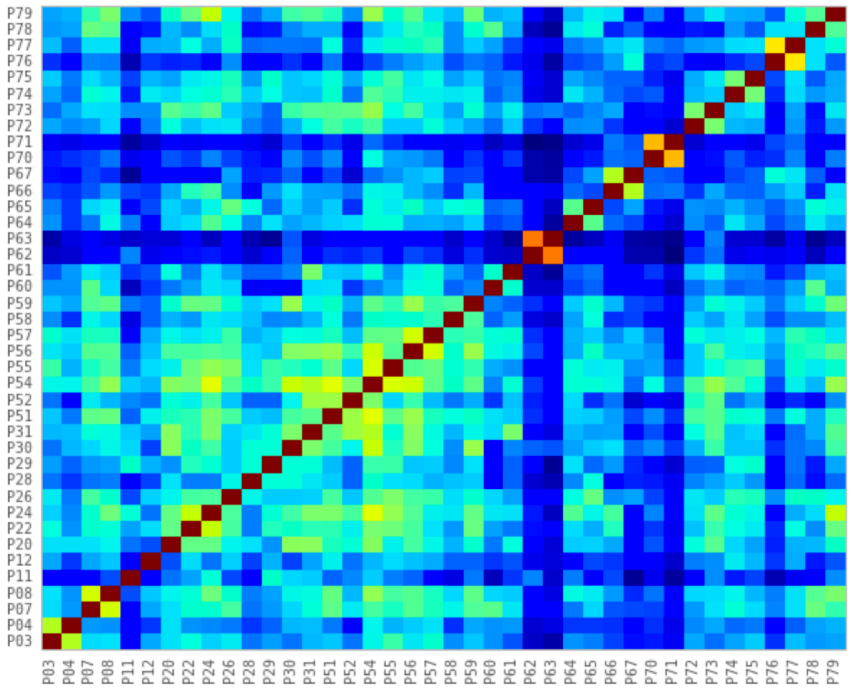

Figure 5. Cosine similarity between each participant of the experiment. The diagonal is red because it represents each student's perfect similarity with herself/himself.

Additionally, we tried to reorganize students on each axis based on their learning scores (Figure 6, left side) and their quality of collaboration (Figure 6, right side). The first approach did not cluster students in any meaningful way, but the second one revealed that students with a poor quality of collaboration (left and bottom rows) tend to look very dissimilar to everyone else (shown in dark blue). This result suggests that poor collaborative groups can potentially be detected using cosine similarity measures.

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

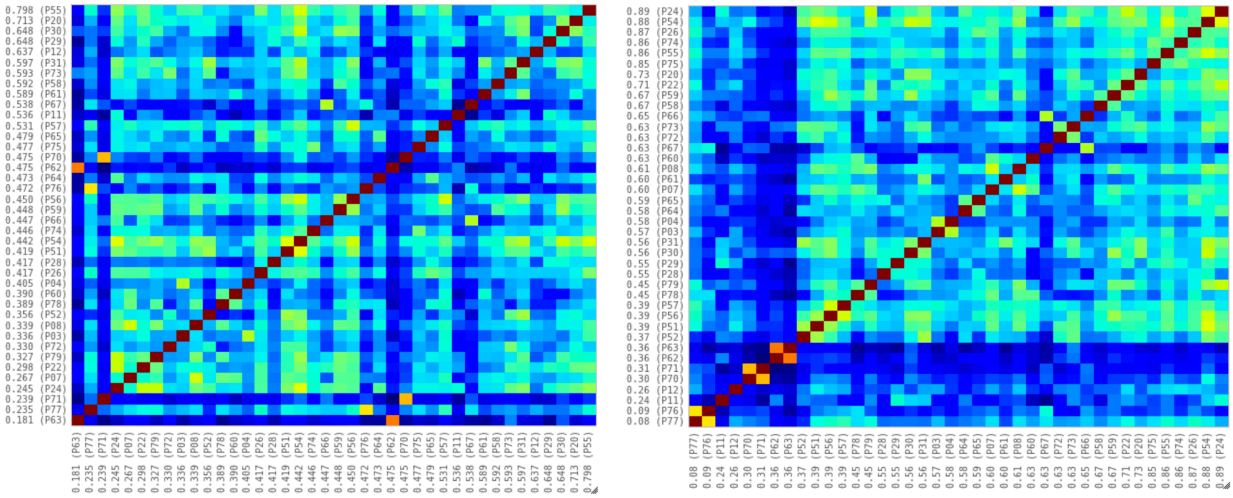

Figure 6. Cosine similarity matrix, reorganized with students' learning scores (left) and quality of collaboration (right).

We then computed a measure of student *coherence*: while our approach is extremely simplistic (more complicated measures exist; see Graesser et al., 2004), it provided an approach relatively easy to understand and apply. We built on our previous results using tf-idf and cosine similarity to assess whether students were reusing ideas mentioned earlier in their discussion. More specifically, we considered *n* exchanges and compared them to the *m* previous exchanges. For instance, where n=5 and m=5, we computed the similarity between utterances 15 to 20 (current discussion) with utterances 10 to 15 (ideas exchanged at the beginning of the experiment). We then iteratively moved this 5-exchange window through the transcript and averaged the similarity across all exchanges to compute our measure of coherence.

Here we provide an example of a highly coherent exchange, highlighting similar words between the two sets of utterances in bold:

— Exchange 1 (5 utterances) —

A: **I think that we did say the fifth one down**.

B: **OK**. **So** then it**'s lesion five**. **OK**.

A: And **you said for** your answer, **you said the** third one down whereas **I said the** sixth one down. **The** rest are **kind of** similar besides **for that kind of** like semi-circle in **the** middle being **kind of** white.

B: **Right, right**. Hold on. **Number** six, **the number for that** side **is** going **to** be, **um**, this **is** tricky business.

A: **Yeah** it is.

— Exchange 2 (5 utterances; same discussion, continued) —

B: **Kind of**?

A: **Yeah**. **So** what do **you** want **to do for lesion five**?

B: **For lesion five**? **Um**, **number… the fifth one down**, **is that what we said** originally? **I think that that's** still **the** correct way **to** go

A: **OK**.

B: **That's** what **we said** initially, **right**?

— End of Exchange 2 —

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

We can observe at least three common repetitions across those two segments. First, the reference to lesion 5 introduced by A in the first exchange and repeated by B in the second exchange. Secondly, both participants express uncertainty by saying "kind of" in the two segments. Finally, there is an abundance of acknowledgement in the form of keywords like "OK" and "right." All those elements point to a relatively solid common ground between the two participants, which is captured by our measure of coherence. Our results, illustrated by the exchange above, are in line with the results of Mitchell, Boyer, and Lester (2012), who showed that convergence is not only associated with conceptual understanding but also with affective components such as frustration, engagement, and confusion. One potential limitation of our measure is that repetitive disagreement (e.g., where students use an abundance of negations) would produce a high coherence score as computed by our method. We did not see any instance of this pattern in our transcript, but it is a limitation that readers should keep in mind.

We now turn to examine how coherence unfolds over time (Figure 7). Each score compares the similarity of 5 exchanges to the 5 following utterances. Observe that the group starts by discussing the diagrams and is focused on solving the problem at hand. But then, they suddenly disengage themselves from the task and start an off-topic discussion. The similarity scores drop at this point of the transcript, which seems to suggest that our measure of coherence can potentially capture topical shifts in the dialogues. *We hypothesize that this may be the main reason why coherence is correlated with learning in our corpus*: Discussing the diagrams is likely to *reduce* the size of the vocabulary used by the students (because they are most likely to use task-related terms, such as "eye," "lesion," "optic nerve," or "hemifield"), which increases the likelihood of repeating the same words. Discussing an off-topic subject is mostly likely to *increase* the size of the vocabulary used, because student focus is larger (e.g., the discussion is not just about human vision anymore, but about anything that comes to mind); additionally, it seems reasonable to expect that students will be likely to shift quickly between mini-topics when talking about off-topic subjects, which will decrease their coherence scores (as opposed to maintaining a 10-minute long discussion on one very specific topic, as required by the task).

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

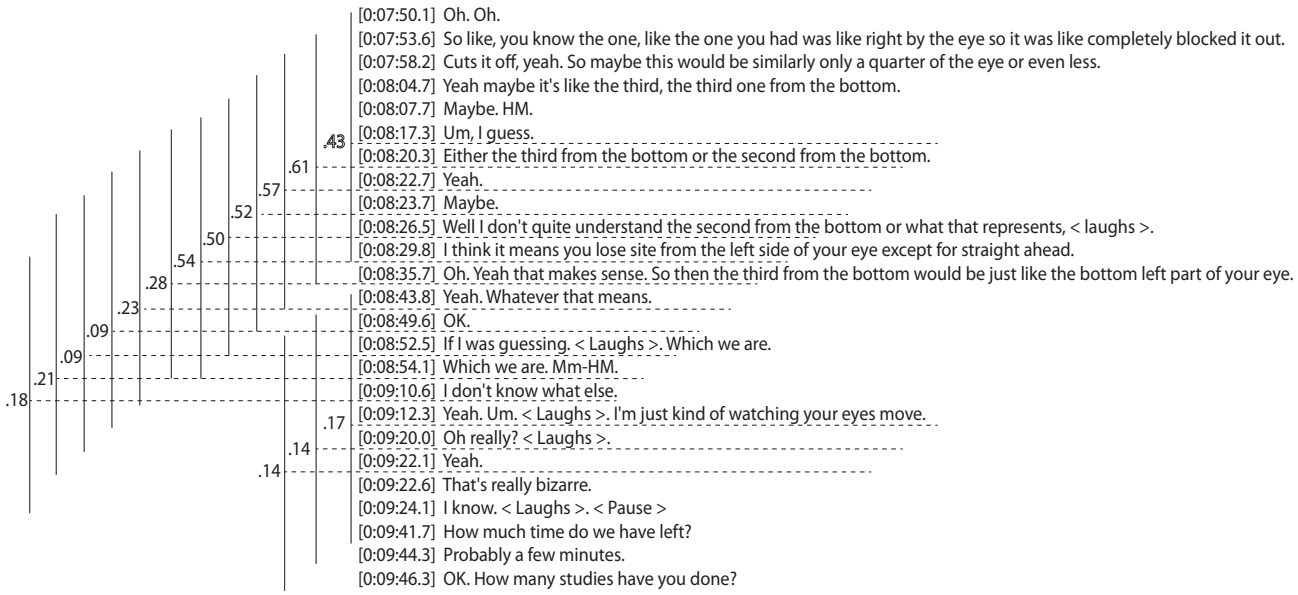

Figure 7. Similarity scores when using a sliding window over the transcript (here each similarity score compares 5 particular exchanges with the 5 following utterances). This example shows how similarity scores drop when students shift the topic of the discussion.

Those hypotheses seem to be supported by quantitative measures. We found aggregated measures of coherence to be positively correlated with student learning gains: r(19)=0.540, p=0.011 (Figure 8, right side). Additionally, we found that students in the "visible-gaze" condition were more coherent than students in the "no-gaze" condition (Figure 9): F(1,20)=7.45, p=0.01, Cohen's d=0.34 (for the visible-gaze group, mean=0.23, SD=0.07; for the no-gaze group, mean=0.15, SD=0.06). *Those results suggest that students who could see the gaze of their partner in real time on the screen were more likely to have a coherent discourse; additionally, a coherent discourse was more likely to lead to higher learning gains*.

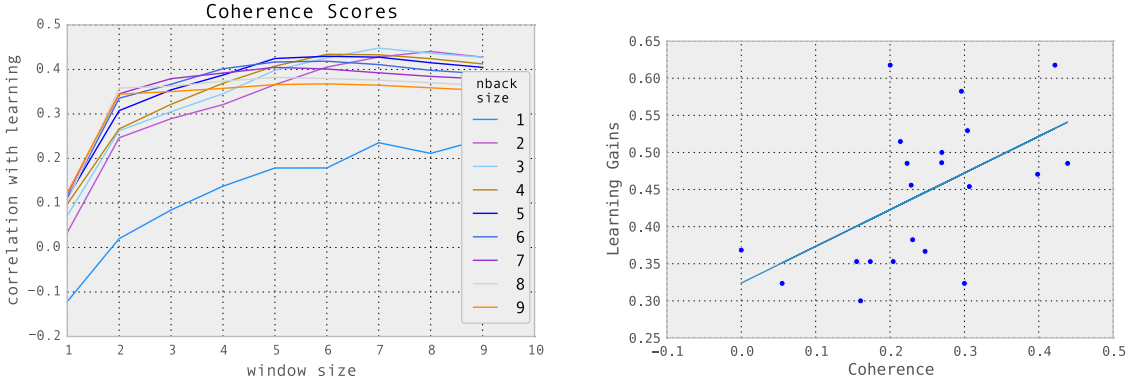

Figure 8. Left side: fine-tuning our coherence measure by comparing *m* utterances (window size) with *n* preceding utterances (nback) and its correlation with learning. Right side: Correlation between dialogue coherence and learning gain (window size=5, nback size=5): r(19)=0.43, p<.05.

On a side note, we tried various values for *n* and *m* (Figure, 8, left side). Some of those results were not significant, but we always found that students in the "visible-gaze" group were more coherent than

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

students in the "no-gaze" group. Finally, we chose to compare five exchanges with the five previous utterances for the following reasons: 1) having the same number of exchanges for the window size (i.e., the *m* current utterances) and the nback size (i.e., the *n* previous exchanges) seems to be more defensible, since we did not have any theoretical reasons for considering an asymmetrical relationship (where m≠n); 2) there do not seem to be large differences when considering a correlation coefficient above 0.4; and, in fact, most values >4 for the window size seems to produce a reasonable correlation coefficient with learning.

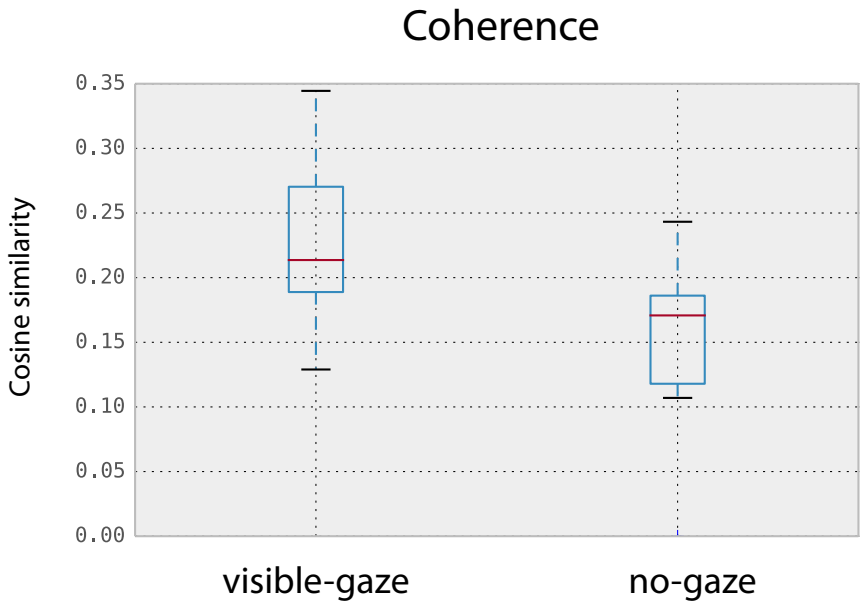

Figure 9. Student coherence when discussing the task. Students in the "visible-gaze" group were significantly more coherent (p<.05).

Finally, it is vital to specify what our measure of coherence is and is not. Our approach is extremely rudimentary compared to state-of-the-art coherence measures developed in educational NLP (Graesser et al., 2004). Approaches like the one used by Coh-Metrix can, for instance, detect topic similarity even when different words are used to refer to the same concept (e.g., sight, vision, field of view). We limited ourselves to comparing the similarity of two blocks of text using a cosine similarity measure, which did not connect synonyms, and a sliding window of five exchanges; thus, our scores do not exactly match what people mean by the word "coherence" in an everyday discussion. In our context, coherence means something akin to "word similarity" or "topical stability"; it does not necessarily mean that a low score indicates incoherence. In our context, a low score seems to indicate either an off-topic discussion, or a discussion with a larger focus (e.g., when students try to think out-of-the-box by connecting the problem they are trying to solve with their prior experiences). Thus, we do not believe that low coherence scores always mean unproductive discussion; similarly, students can potentially have a highly coherent discussion on an off-topic subject. What our measure seems to indicate, however, is that *on average* high coherence scores are related to focused discussion on the diagrams, which seems to be associated with better outcomes evidenced by student learning gains.

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

In the next section, we describe a first effort at refining this Coherence construct. Since we already have measures related to students' joint visual attention from our previous study (Schneider & Pea, 2013), we decided to discriminate between high/low coherence scores by using measures of high/low JVA. Our hypothesis is that moments of high JVA and high coherence are more likely to be reflective of productive discussions about the diagrams, whereas moments of low JVA and low coherence are more likely to be reflective of unproductive and unfocused discussions.

4.5 Coherence and Joint Visual Attention

Joint Visual Attention indicates whether two participants are looking at the same area of the screen at the same time. JVA is a useful construct because it reflects the quality of students' common ground when trying to understand a concept together. In this section, we computed JVA as follows: for each gaze, we defined it as joint visual attention if the gaze of the two participants converges within a time window of +/-2 seconds (as suggested by Richardson & Dale, 2005); additionally, we binned gazes in large Areas Of Interest (AOIs) represented by the diagrams in Figure 1. Because our areas of interest are so large, our measure of JVA is higher than the numbers traditionally found in the dual eye-tracking literature (i.e., on average two individuals look at the same area ~20% of the time when considering a radius of a few centimeters). We also computed JVA in this manner, by limiting joint visual attention to an area of 70 pixels (as used by Jermann, Mullins, Nuessli, & Dillenbourg, 2011). There was a strong correlation between the AOI method and the radius methods: $r(19)=0.913$, $p<0.001$, which indicates that those two ways of computing JVA are closely related. We opted to use the AOI method because it was significantly correlated with student learning gains ($r(19)=0.50$, $p=0.048$) whereas the radius method was not ($r(19)=0.35$, $p=0.18$).

For our experimental context, augmenting measures of coherence with JVA can help us discriminate between spatially locked and non-spatially locked discussions. To explore this question, we graphed the evolution of those two measures over time for each participant (Figure 10): green lines indicate levels of JVA while blue lines indicate levels of coherence. Bands show the sample mean and +/-1 standard deviation.

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

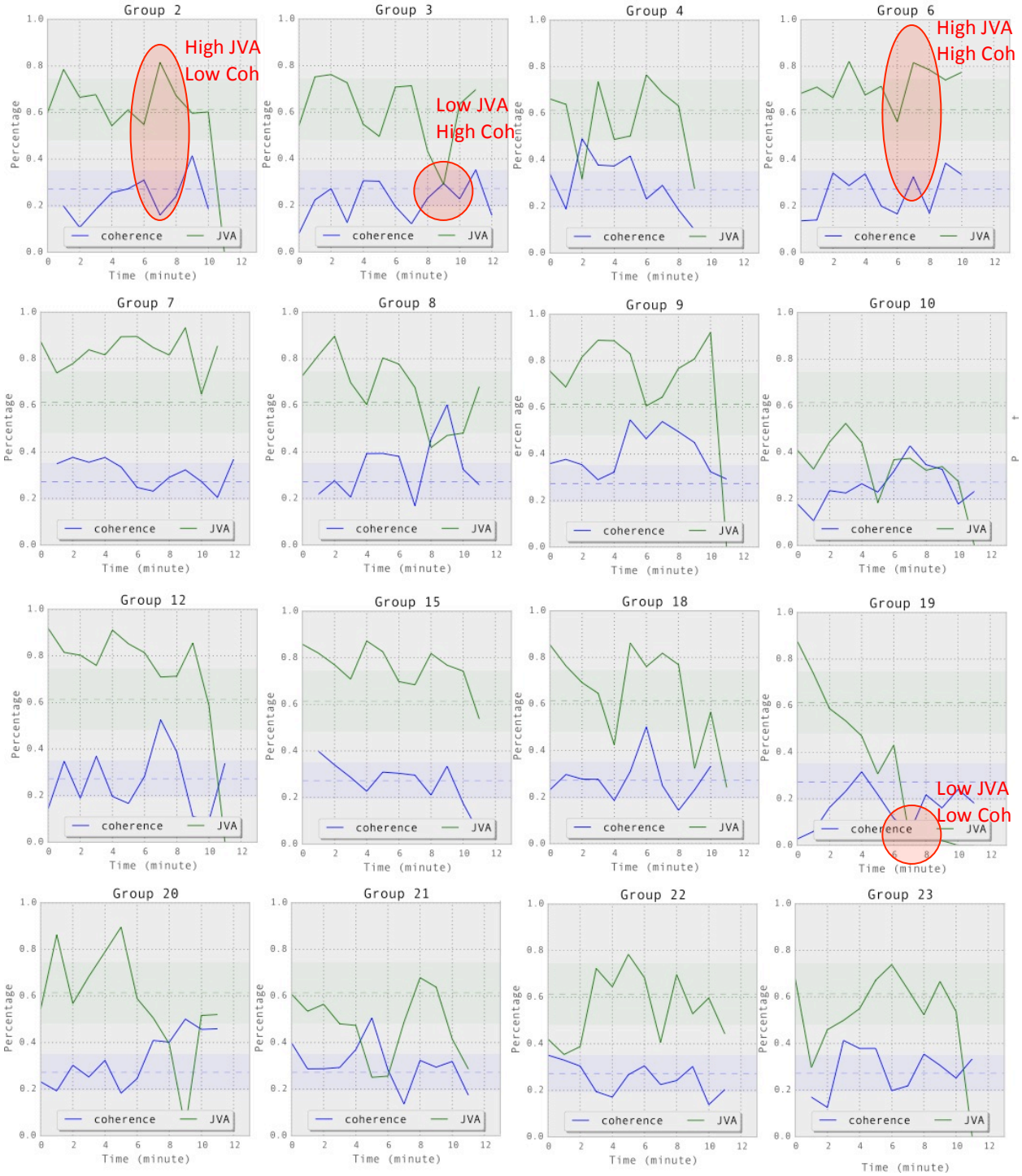

Figure 10. Evolution of JVA and Coherence for each experimental group. The dashed line shows the group mean; the transparent bands represent +/-1 standard deviation of the sample. Moments of interest are highlighted in red (High /Low JVA crossed with High /Low Coherence).

Here, our goal was to use those graphs to explore our transcripts qualitatively and find interesting

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

learning moments. We selected interesting examples by adopting a qualitative approach to describe variability in our graphs (Firestone, 1993). We visually inspected the graphs and located moments where both measures dramatically increased or decreased. This approach implies that lower or higher scores are *relative* to the dyad's average: the exchanges shown below can exhibit poor coherence with regard to the rest of the dyad's discussion (within-group comparison) but high coherence compared to other groups (between-groups comparison). In this context, "high" means "increasing," and "low" means "decreasing." The goal of this analysis is to identify moments where the conversation is improving or degrading, relative to prior exchanges. Another reasonable approach would have been to isolate those moments compared to all other dyads in this study, a type of analysis to be explored in future work. Finally, to facilitate visual exploration, we aggregated data for each minute. After conducting in-depth explorations of transcripts, we show representative examples of both High/Low JVA and High/Low Coherence:

High JVA, High Coherence (group 6)		**Low JVA, Low Coherence (group 19)**	
0:07:03.3	I think it's either this one or this one.	00:07:07.35	Okay.
0:07:09.2	Um, well it's not the one you're looking at right now.	00:07:11.30	we still have, uhh, 10 minutes.
0:07:12.0	Uh-huh.	00:07:15.20	10 minutes?
0:07:13.7	Because that's the top left.	00:07:16.20	No, actually 5.
0:07:15.2	Oh that's right, that's right.	00:07:17.70	So just keep talking about it?
0:07:16.5	Yeah. So, um. Yeah, I could say, I'd say that it's the one that you're looking at right now.	00:07:19.25	<Clears throat> Yeah.
0:07:25.8	OK.	00:07:21.30	Oh.
0:07:26.6	I think I'd agree on that one. <Pause>.	00:07:22.10	Ohh.
0:07:33.6	Or actually no. I think it's the one that I'm looking at now. Do you see it?	00:07:22.40	So how's your day been? <Laughing>
0:07:37.9	Yeah. Oh 'cause it's the opposite of, OK. Yeah.	00:07:25.40	Pretty good. I'm tired <chuckles>.
0:07:44.9	Yeah. Where the left sides of the circles are being cut off.	00:07:27.20	Me too. I had, I left class just to come over here <laughs>.
0:07:48.4	Yeah that makes, yeah, OK.	00:07:30.05	Really?
0:07:50.6	Yeah, I think.	00:07:30.80	<Laughing> Yeah.
		00:07:31.20	<Sniffles> I didn't miss any class. I got a burger to eat before this.
		00:07:37.00	Oh, lucky.

The first example suggests that exchanges with both high JVA and high Coherence are spatially locked discussions where both participants are exchanging ideas about the diagrams. There are a multitude of spatial referents ("this one," "the one you're looking at," "the top left," "where the circles are being cut off") that anchor the discussion. The excerpt on the right, on the other hand, has low Coherence (which means that a variety of words are being used, and that the exchanges are rather short), low JVA (which means that the two students are looking at different locations), and is clearly off-topic. *This contrast suggests that we can potentially detect productive discussions on a diagram and off-task behaviour by finding moments where both measures of JVA and Coherence either increase or decrease*.

The next two examples contrast the two remaining cases. Group 2 (high JVA, low Coherence) seems to be looking at the same location on the screen, but does not exhibit a high similarity in terms of the vocabulary used: here, low coherence indicates that students are not listening to each other; two monologues are taking place simultaneously about the same diagram. Group 3 (low JVA, high

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

Coherence) exhibits the opposite pattern: here, low JVA seems to suggest that students are trying to integrate information from all the diagrams and their prior knowledge, but high Coherence indicates that they are discussing the same concept. It is noteworthy that low JVA can signal productive moments of learning; usually, low aggregate measures of joint attention are perceived as a proxy for unproductive collaborations in the dual eye-tracking literature. The counter-example below suggests that this is not always the case.

High JVA, low coherence (group 2)		**Low JVA, high coherence (group 3)**	
0:07:05.4	Oh. I see what you're saying.	0:09:10.0	I mean isn't that kind of how your brain works? It works in opposites.
0:07:06.8	There really isn't a choice for that, is there?	0:09:13.2	Oh really? <laughs>. I've never taken any of these classes.
0:07:10.2	No. Let me think for a second.	0:09:17.4	I mean I could be totally wrong but I'm pretty sure that the right side of your brain affects the left side of your body.
0:07:13.3	Unless we can pick, unless we can pick the same one as six again. We can use it twice. Maybe it's also left and dark on the right.	0:09:28.1	Maybe. Oh. OK.
0:07:26.3	<Mumbling to self>, left hemi-field. Right. OK. So.	0:09:30.4	So if lesion one is blocking out the right side...
0:07:31.4	No. But that doesn't make sense because the right hemi-field isn't impacted. <Sighs>, HM.	0:09:33.9	Oh! And then OK.
0:07:39.3	This is... <mumbling to self>. Corner...	0:09:37.1	Then it would affect your left side.
		0:09:39.0	Left side — that makes sense.
		0:09:40.9	So then lesion five has to be wrong at some point.
		0:09:45.6	You should be looking at your right then, I mean for the vision thing.
		0:09:51.6	What if lesion five is the third one from the top? Where, no, never mind, that doesn't make sense. I don't know.
		0:09:59.7	It makes more sense, well it makes more sense if it be the second from the top. Because one of the right vision blocking, I don't know it doesn't make sense to me too.

Finally, we can also observe that, more often than not, JVA and Coherence trend in opposite directions. For instance, group 3 exhibits a "mirrored" pattern, where increasing Coherence is accompanied by decreasing JVA, and inversely, increasing JVA is accompanied by decreasing Coherence. This pattern suggests that dyads go through "cycles" of problem-solving, where they alternate between collecting and integrating information from a multitude of locations (low JVA) and moments of intense analysis of one or two particular diagrams (high JVA). Those examples show that combining very rudimentary measures on student dialogue and gaze movements can allow us to isolate relevant steps in the problem-solving process. They also suggest that researchers should move from a simple dichotomy where high JVA = good collaboration and low JVA = bad collaboration. On average, this seems to be true; but we believe that the bigger picture is more complicated than that, and that we can uncover new collaborative learning patterns by adopting a more nuanced perspective.

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

4.6 Additional Results

In a subsequent step, we explored a few additional ways to exploit text similarity to predict our variables of interest (i.e., learning gains and student collaboration quality). Our approach was to seek additional baselines for comparing student utterances to other documents. For instance, we can imagine comparing the transcripts of students with a baseline of an expert discussion on this topic. To this end, we used the following two corpora: first, we compared the best student (in terms of her learning score) of our dataset (P55) with every other participant. She was in the visible-gaze condition and got an impressive 80% gain on the post-test, where the average was around 50%. Second, we inserted the text that students had to read in the second step of the experiment into our dataset. This text is highly technical and is likely to pick up student use of the particular terminology associated with this domain. It should be noted that students were exposed to this text *after* discussing the diagrams.

We found that students in the "visible-gaze" group looked more like P55: $F(1,39)$, $p=0.04$, Cohen's $d=0.35$ (visible-gaze mean=0.97, SD=0.27; no-gaze mean=0.80, SD=0.20). This measure was positively correlated with students' quality of collaboration: $r(38)=0.545$, $p<0.001$. There was no significant difference between the two groups when looking at their similarity with the textbook chapter: $F(1,39)$, $p=0.17$, Cohen's $d=0.10$ (visible-gaze mean=0.11, SD=0.04; no-gaze mean=0.09, SD=0.04). However, this measure was significantly correlated with students' conceptual understanding of the topic taught: $r(38)=0.335$, $p=0.035$.

In summary, it appears that taking different baselines is helpful for finding relevant predictors of good learning groups. Taking a student's cosine similarity with a standard reference of domain knowledge (i.e., a textbook chapter) seems to be associated with higher learning gains on a test. Taking a student's cosine similarity with the "best" student of the dataset seems to be associated with productive patterns of collaboration. This makes sense because student utterances reflect the way novices discuss and learn about a new topic; a scientific text, by contrast, is produced by experts who have mastered the domain concepts and terminology. In sum, those two features could be advantageously used to further explore student discussions, as well as to feed machine learning algorithms trying to predict student learning.

4.7 Predicting Student Learning Gains Using Linguistic Features

Our final contribution is to test whether the measures described above have any predictive value. Specifically, can we roughly classify students in terms of their learning gains using machine learning algorithms? To answer this question, we separated our participants into two groups based on the median value of student learning gains. We then tried to predict in which group each student belonged, i.e., below or above the median split.

We then used our hand-labelled categories from section one (n-grams), the cosine similarity scores, the convergence measures and the coherence metrics as features. The complete mxn matrix contained

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

m=40 rows and n=60 features. We used the built-in version of Support Vector Machine (SVM) provided by Matlab with a forward search feature selection and tried various kernels (linear, quadratic, polynomial, Gaussian, multilayer perceptron). To minimize overfitting, we used a validation set (4 rows) and a Leave-One-Out Cross Validation procedure (LOOCV) on the remaining dataset (36 rows), where the model was repeatedly trained on m-1 rows and tested on the remaining row. The averaged accuracy of this model on the left-out row is reported in the "test set" column of Table 3. We found that SVM with a multilayer perceptron kernel and 8 features could correctly classify 94.44% of our participants on the median-split performed on the learning scores. As mentioned above, we also used a *validation set* (4 participants, which constitutes 10% of our sample). Considering our low sample, 10% is a commonly used ratio to divide our dataset between training and validation sets. Those 4 participants were randomly selected from our dataset and we predicted whether they were above or below the median split on the learning gains after we found our best model. On the validation set, our model correctly classified 75% of the participants (3/4).

Table 3. Rough classification of students (using a median-split) in terms of their learning gains. To minimize overfitting, we used a validation set (4 rows) and a Leave-One-Out Cross Validation procedure (LOOCV) on the remaining dataset (36 rows), where the model was repeatedly trained on m-1 rows and tested on the remaining row. The averaged accuracy of this model on the left-out row is reported in the "test set" column.

	Accuracy on the test set	Accuracy on the validation set	Features
SVM	94.44% (34/36)	75% (3/4)	Uncertainty Negations Aux. Verbs Sentence Length Prepositions Number of words used Number of Anaphoras Impersonal Pronouns

Those results are impressive, but they need to be hedged with healthy skepticism. First, many features were used to make this prediction. It is probable that the algorithm is cherry-picking the relevant features to improve its accuracy (which is also over-fitting the data). Secondly, the training set is rather small. There are only ~40 students to classify, which is another serious limitation. Finally, even though we are using a validation set, it should be kept in mind that this set is small (only four data points). It is imperative for future work to replicate those results with larger samples.

In sum, these analyses indicate noteworthy promise in using linguistic features to predict student learning and productive collaboration with their peers, but the results need to be replicated on larger datasets to be truly convincing.

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

Interestingly, SVM selected some of the measures we manually coded in the first section of this paper: number of anaphoras used and keywords showing student uncertainty. However, other measures such as coherence, and cosine similarity with a textbook chapter were not included in our final model. Instead, the SVM favoured low-level measures, such as the number of words used by students, the length of their sentences and particular grammatical forms (negations, auxiliary verbs, prepositions). This result shows that some variables may be good predictors in isolation, but lose their predictive power when associated with other measures.

5 DISCUSSION

The goal of this project was to explore various NLP techniques to make sense of educational datasets from collaborative learning interactions; we favoured a "breadth" approach where we tried promising techniques rather than exploring one specific measure in depth. In future work, we will go back to our most promising results (e.g., coherence and cosine similarity) and explore them in more detail, as well as examining not only the cosine similarity to the best student of the other student transcripts but to more aggregate exemplars of "better or worse students," such as the upper and lower quartile of students gauged by learning score.

To recap our results, we found 1) that n-gram probabilities can help characterize groups of students in terms of building a common ground with their learning partners (i.e., use of anaphora); 2) that cosine similarity measures are most useful when used with a "reference" corpus (e.g., textbook chapter; transcript of a very good student as measured by learning gains); 3) that coordination of linguistic style has little predictive power in terms of explaining dyads' collaborative learning processes; 4) that coherence measures, on the other hand, are positively associated with student learning; 5) that we were able to refine our measure of coherence by integrating data on joint visual attention; and 6) that using SVM and the features mentioned above, we found that we could roughly predict student learning outcomes with an accuracy higher than 90% (although accuracy dropped to 75% for our smallish validation set).

We argue that our approach is especially useful when analyzing the results of a controlled experiment. We were able to characterize the effects of *mutual gaze perception* on student discourse, and we found interesting predictors for learning gains and student collaboration quality. Yet we also argue that those techniques could be used in other domains. For instance, comparing the similarity between a reference text and student utterances has already been used for assessing essays. Coherence can be used in similar contexts. More interestingly, those metrics could be profitably used on multi-modal datasets. Eye-tracking data, for instance, could be converted into a series of word tokens representing the location of student gaze over time. Similarity measures could then be used as described above to characterize visual exploration of a problem space. We believe that NLP measures have been too rarely used on non-linguistic datasets (e.g., gestures, as measured by a Kinect sensor; gaze, as measured by eye-trackers; arousal, as measured by galvanic skin response devices) and could provide new insights

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

into the ways in which students construct their understanding of a particular concept, and establish a productive collaboration with one another.

Limitations of this work have been mentioned in previous sections (e.g., small dataset, limited amount of error analysis — in the sense that we did not analyze why our SVM algorithm misclassified some dyads). Replicating those results on larger datasets would make a more convincing argument for using NLP measures in education. Finally, it should be mentioned that the study described in this paper has low ecological validity as compared to studies in classrooms or in co-located collaborative settings. Students worked in a remote collaboration, where they had few ways to communicate. There were no shared cursors or representations that they could exploit to establish joint visual attention. We agree that a third condition, where students could create shared annotations or see the cursor of their partner, would enable teasing apart the effects of seeing the gaze of one's collaborator at all times, and the effects of having a common deictic pointer. It is an important limitation that readers should take into account when interpreting our findings. We hypothesize that the gaze-awareness condition nonetheless offers advantages compared to a "visible mouse cursor" condition, since students can persistently monitor the attention of their partner and anticipate their partner's contribution; a mouse cursor, on the other hand, needs to be consciously moved to an area of interest to attract a peer's attention, which can potentially increase the cognitive load of the group. Those issues and limitations are discussed in more detail in our previous publication (Schneider & Pea, 2013).

6 CONCLUSION

This paper demonstrates that NLP approaches offer substantial promise for understanding educational datasets and automating currently unwieldy and time-consuming hand analyses. The measures described above could easily be applied to other settings, such as forums or online discussions. Future work includes refining those measures, deepening our sense of their predictive value, and replicating those results on other datasets.

More generally, we see a focus on joint visual attention and coherence as two productive ways for the field of learning analytics to move forward. Those two constructs have been extensively studied in the learning sciences, and allow us to conduct promising theory-driven research. Sometimes we see work in educational data mining where researchers are overwhelmed with massive datasets and where they find themselves extracting semi-random indicators of learning due to a lack of a solid theoretical framing. By starting with accepted constructs, we limited the risks of finding an effect by chance (Type 1 error) and constrained the ways we could analyze our datasets. From our perspective, the most promising path to move forward is 1) to keep improving simple measures of student verbal coherence, and 2) to add more layers of multi-modal information to those analyses. Our findings above suggest that a simple construct such as joint visual attention can become increasingly rich and complex as we start to discriminate between moments of low or high coherence. Similarly, it is likely that silent moments of joint attention capture a different type of collaborative synchronization, and that moments of joint

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

attention convey different meaning depending on the kind of gestures that they accompany (e.g., deictic, iconic, metaphoric). By combining NLP measures with various kinds of sensors (eye-trackers, motion sensors, Galvanic Skin Response sensors) we can start to unpack a complex taxonomy of productive collaborative learning markers. The overarching goal is to then use those markers to construct "learning states" and map student trajectories using those states. One could then differentiate productive from unproductive trajectories and intervene to redirect students from the latter state to the former one.

In conclusion, we believe that the field of learning analytics (and more specifically our work) can contribute to education in several ways. First, finding positive predictors of learning can help us unpack student learning trajectories by detecting unproductive patterns; this information, in turn, can provide interesting feedback loops for both students and teachers to avoid dead-ends in their learning paths. Second, high-frequency sensors (such as eye-trackers) can provide an additional layer of complexity to nuance simple measures of learning: in this paper, we showed that combining JVA with coherence had the potential for uncovering productive and off-task behaviours. Finally, computational measures offer the prospect of speeding up the pace of educational research by automatically extracting constructs of interest: instead of painstakingly annotating hours of videos and reams of transcripts, we can start to graph the evolution of particular behaviours and use those graphs to isolate interesting learning moments. We believe that those three points are within the reach of the budding field of educational data mining, and have the potential to make consequential differences in the ways that students learn.

ACKNOWLEDGMENTS

We gratefully acknowledge grant support from the National Science Foundation (NSF) for this work from the LIFE Center (NSF #0835854).

REFERENCES

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences, 12*(3), 307–359. http://dx.doi.org/10.1207/S15327809JLS1203_1

Bates, E., Thal, D., Whitesell, K., Fenson, L., & Oakes, L. (1989). Integrating language and gesture in infancy. *Developmental Psychology, 25*(6), 1004–1019.

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition, 106*(3), 1465–1477. http://dx.doi.org/10.1016/j.cognition.2007.05.012

Cherubini, M., Nüssli, M., & Dillenbourg, P. (2008). Deixis and gaze in collaborative work at a distance (over a shared map): A computational model to detect misunderstandings. *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (ETRA '08), 26–28 March 2008, Savannah, Georgia (pp. 173–180). New York: ACM.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine, S. D.

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

Behrend (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: American Psychological Association.

Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 76–87). New York: Association for Computational Linguistics.

Firestone, W. A. (1993). Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher, 22*(4), 16–23. http:dx.doi.org/10.3102/0013189X022004016

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. http://dx.doi.org/10.3758/BF03195564

Jermann, P., Mullins, D., Nuessli, M.-A., & Dillenbourg, P. (2011). Collaborative gaze footprints: Correlates of interaction quality. In H. Spada, G. Stahl, N. Miyake, & N. Law (Eds.), *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL2011 Conference Proceedings* (Vol.1, pp. 184–191). Sydney, Australia: ISLS.

Jermann, P., Gergle, D., Bednarik, R., Brennan, S. (2012): Duet 2012: Dual eye tracking in CSCW. *Companion Proceedings of ACM CSCW12 Conference on Computer-Supported Cooperative Work* (pp. 23–24.) New York: ACM. http://dx.doi.org/10.1145/2141512.2141525

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge, UK: Cambridge University Press.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *International Journal of Computer-Supported Collaborative Learning*, *2*(1), 63–86. http://dx.doi.org/10.1007/s11412-006-9005-x

Mitchell, C. M., Boyer, K. E., & Lester, J. C. (2012). From strangers to partners: Examining convergence within a longitudinal study of task-oriented dialogue. *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogu*e, (pp. 94–98). New York: Association for Computational Linguistics. Retrieved from http://www.sigdial.org/workshops/sigdial2014/proceedings/W14-43-2014.pdf

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001, Mahwah, NJ: Lawrence Erlbaum Associates.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*(6), 1045–1060. http:dx.doi.org/10.1207/s15516709cog0000_29

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science, 18*(5), 407–413. http://dx.doi.org/10.1111/j.1467-9280.2007.01914.x

Roschelle, J. (1992). Learning by collaborating: Convergent conceptual change. *The Journal of the Learning Sciences*, *2*(3), 235–276. http://dx.doi.org/10.1207/s15327809jls0203_1

(2015). Does seeing one another's gaze affect group dialogue A computational approach. *Journal of Learning Analytics*, *2*(2), 107–133. http://dx.doi.org/10.18608/jla.2015.22.9

Schneider, B., & Pea, R. (2013). Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning*, *8*(4), 375–397. http://dx.doi.org/10.1007/s11412-013-9181-4

Salomon, G., & Globerson, T. (1989). When teams do not function the way they ought to. *International Journal of Educational Research*, *13*(1), 89–99.

Stahl, G. (2013). Transactive discourse in CSCL. *International Journal of Computer-Supported Collaborative Learning*, *8*(2), 145–147. http://dx.doi.org/10.1007/s11412-013-9171-6

Sherin, B. (2012). Using computational methods to discover student science conceptions in interview data. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)*, (pp. 188–197). http://dx.doi.org/10.1145/2330601.2330649

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist, 46*(6), 606–620.

Stern, D. (1977). *The first relationship: Infant and mother*. Cambridge, MA: Harvard University Press.

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore, & P. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). New York: Psychology Press.

Ward, A., & Litman, D. (2007). Dialog convergence and learning. In R. Luckin, K. R. Koedinger, J. Greer (Eds.), Artificial intelligence in education: Building technology rich learning contexts that work. (pp.262–269). Los Angeles, CA: IOS Press.