The Effect of Mutual Gaze Perception on Students' Verbal Coordination

Bertrand Schneider
Stanford University
schneibe@stanford.edu

Roy Pea
Stanford University
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 students' collaboration and learning [8]. The goals of this paper are to: 1) explore a variety of computational techniques for analyzing the transcripts of students' discussions; 2) examine whether any of those measures sheds new light on our previous results; and 3) 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 the *coherence* of students' discourse was significantly different across our experimental conditions; this measure was positively correlated with their learning gains. Finally, using various language metrics, we were able to roughly (i.e., using a median-split) predict 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.

1. INTRODUCTION

Despite recent efforts in developing automated ways to analyze students' discourse, most educational researchers still rely on traditional tools to analyze transcripts from students. Traditional methods include time-consuming qualitative analyses and the development of manual coding schemes. The field of Natural Language Processing (NLP) has significantly grown and gained in maturity over the past decades, and computational techniques can now be advantageously applied to educational datasets. Recent efforts in topic modeling, for instance, seem to be especially promising in terms of gaining insights into students' discourse and cognitive processes [9]. Unfortunately, social scientists willing to learn those tools are a rare breed, and multi-disciplinary work is slow to appear between educational researchers and computer scientists. In this paper, we describe our attempt at applying NLP techniques to educational transcripts.

2. THE CURRENT DATASET

In a previous work [8], we conducted a study on the effect of *mutual visual gaze perception* on students' collaborative problem-solving processes. In this experiment, student dyads were asked to remotely collaborate 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 an audio channel. The information on the screen was similar for both participants (i.e., the brain diagrams shown in Fig. 1). The structure of the activity was as follows: in the first step, students analyzed brain diagrams (12 minutes); in a second step, they were asked to read a textbook chapter about human vision and discuss their understanding of this topic (12 minutes). Finally, before the analysis activity and after the reading task, students were asked to complete a learning test (pre/post-questionnaires).

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 which recorded students' gaze. In a control group ("no-gaze"), the other half of our participants did not have access to this visualization. This intervention helped students in the first group achieve higher learning gains (Fig. 2) and a higher quality of collaboration (as measured by [4]).

We also recorded students' gaze movements and their collaborative discourse. Interestingly, 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 be looking at the same diagram at the same time on the screen), and this measure was significantly correlated with positive learning gains. This result reinforced the assumption that joint visual attention is a crucial mechanism for coordinating social interactions [10].

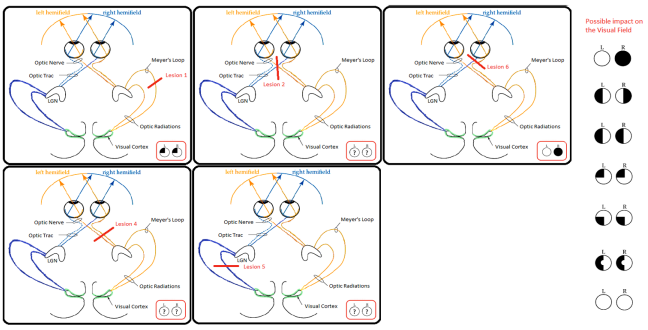

Figure 1: Diagrams students had to analyze. Five contrasting cases show the visual pathways of the human brain; students had to identify the effect of each lesion on the visual field.

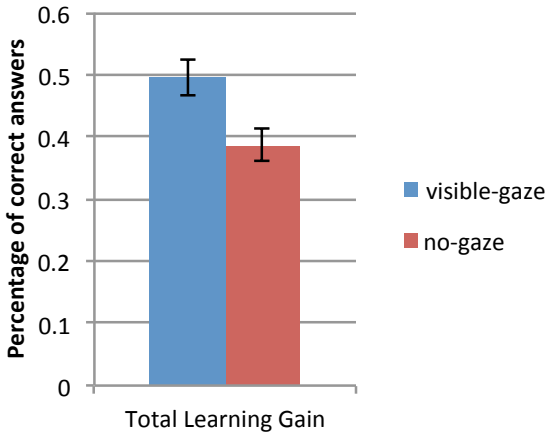

Figure 2: Learning gains for the two experimental groups of the study ($p < .01$).

In a subsequent analysis, we also 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*, and thus remove the need to explicitly mention locations on the diagrams; and finally, 3) they could monitor the visual activity of their partner at all times, providing an aid to establishing a common ground.

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

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

Technically, we can answer the first question by designing linguistic metrics and running statistical tests (i.e., 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.

3. NATURAL LANGUAGE PROCESSING AND MUTUAL GAZE PERCEPTION

In the next sections, we describe the measures used to provide a preliminary answer to those questions. First, we looked at unigrams, bigrams and trigrams counts to build categories of interest using a bag of words model. Next, we looked at the coordination of linguistic styles among students: are students more likely to mimic the grammatical structure of their peers in a good collaboration (as suggested by [2])? We then assessed the *coherence* of students' discourse, by comparing the similarity of consecutive sub-sections of the transcripts; our goal was to evaluate the extent to which students were building on each other's ideas during the task. Finally, we gathered all the previous measures and ran a machine-learning algorithm (Support Vector Machine) to roughly predict students' learning gains.

3.1 N-GRAMS

To get a sense of our dataset, we first computed unigram, bigram and trigram probabilities. This helped us 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 showed 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: unigrams are often ambiguous (see the example above), and bigrams or trigrams are usually so rare that they don't provide strong evidence for any type of hypothesis. This is why we decided to group them by categories instead of 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.

For instance, the category '*anaphora'* contained the words "it", "some", "that", "which", "each", "few" and so on; the category '*conceptual discussion*' contained "think", "cause", "because", "suppose", "impact", and so on. Table 1 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 '*anaphoras'* 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 a common ground between two participants: since anaphoras are ambiguous by nature, they have to be correctly interpreted by the interlocutor and thus indicate a stronger coordination between students. Herbert Clark has developed a considerable body of work investigating this topic [1].

Table 1: Categories built on common unigrams.

Category	Unigrams
Jargon	hemi, field, hemifield, brain, eye, lesion, optic, vision, meyers, 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, one, another, other, others, several, some, something, that, these, this, those, what, which

Participants in the experimental group used more anaphoras 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 dyads in 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 (Fig. 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.

Even with these limitations, it is interesting to see that categories built on n-grams frequencies can offer a new window into students' collaborative learning processes. In the next section, we employ algorithms from the field of information retrieval to further explore the differences between our experimental groups.

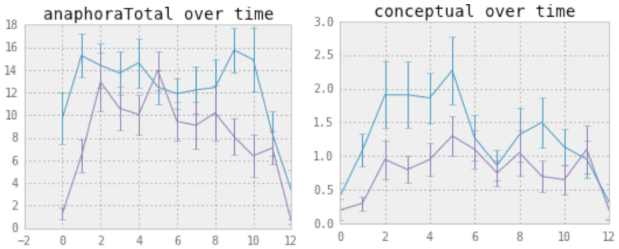

Figure 3: Evolution of words related to conceptual discussion and anaphoras over time. Blue line corresponds to the "visible-gaze" group; purple line to the "no-gaze" group.

3.2 COORDINATION OF LINGUISTIC STYLES (CONVERGENCE)

Computing n-grams counts and probabilities is an interesting way to look at students' discussions. However it doesn't contribute to our understanding of the linguistic patterns used in collaborative learning discussions. To address this issue, we propose studying the ways in which students build a discourse around the instructional material. More specifically, we looked at a specific phenomenon in social interactions called the *chameleon effect*. In a previous study, Danescu [2] showed how in a social setting people tend to mimic their interlocutor's grammatical structure. Here is an example:

Doc: At least you were outside.

Carol: It doesn't make much difference where you are [...]

From Danescu: "Note that "Carol" used a quantifier, one that is different than the one "Doc" employed. Also, notice that "Carol" could just as well have replied in a way that doesn't include a quantifier, for example, "It doesn't really matter where you are...".

In two large datasets (movie dialogues and twitter), Danescu importantly shows that this effect (called *convergence*) is relatively robust and pervasive. That is, people tend to consistently mimic the grammatical structure used by their interlocutor. Previous research suggests that this convergence is associated with enhanced communication in organizational contexts and in psychotherapy (cited in [2]). Our goals are to 1) replicate Danescu's results on our dataset, and 2) test whether *mutual visual gaze perception* supports convergence.

Concretely, Danescu used 9 categories from the LIWC corpus (Linguistic Inquiry and Word Counts [7]) to compute converge measures. Those categories are: articles, auxiliary verbs, conjunctions, high-frequency adverbs, impersonal pronouns, negations, personal pronouns, prepositions, and quantifiers. The way convergence is computed is relatively trivial:

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

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

Figure 4 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 previous utterance. Figure 5 shows our replication of Danescu's results. We can see 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).

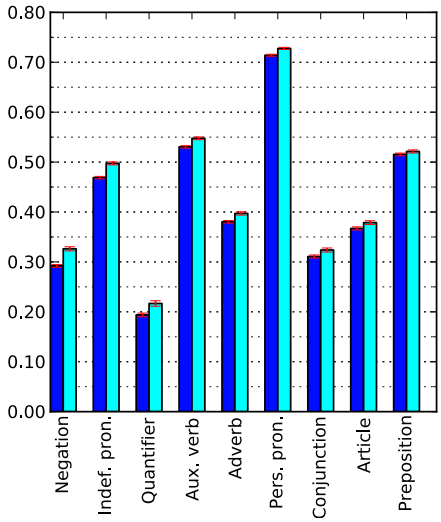

Figure 4: From Danescu [2], this graph shows how people tend to mimic the grammatical structure of their interlocutor. Light blue bars show the conditional probability of using a particular word type, given that an interlocutor used it in the previous utterance. Dark blue bars show the probability of using a particular word type in the entire corpus.

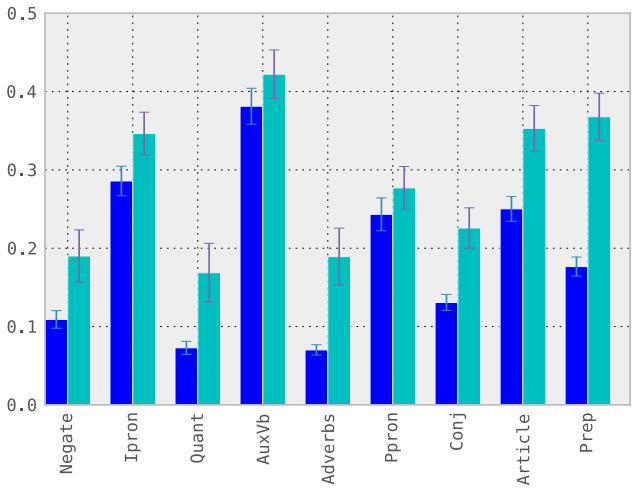

Figure 5: A replication of Danescu's results on the current dataset. Errors bars show standard errors. Non-overlapping error bars show statistically significant differences.

Most importantly, there was special potential in using this measure to discriminate between the two experimental groups (e.g. "visible-gaze" vs "no-gaze"; productive vs poor collaborators; good vs poor learners). Unfortunately, there wasn't any significant difference between those groups on our convergence measure ($F < 1$). This means that, at least in our corpus, coordination of linguistic styles is not predictive of positive learning gains. It also shows that *mutual gaze perception* doesn't influence this effect: students are not more likely to imitate each others' 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). It would be much more interesting to look at the words themselves. If one could show that productive students are more likely to mimic the *content* mentioned by their partner, this would be a more interesting result.

3.3 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 students' coherence. A cosine similarity score indicates how similar two text documents (or subsections of a transcript) are. Our approach was to segment students' transcripts into smaller texts and compute similarity measures between those segments. By iteratively repeating this procedure, we can evaluate the *coherence* of a discussion [6]. The idea behind coherence is that interlocutors tend to adapt to the patterns in each other's utterances; this alignment, in turn, is believed to be indicative of shared understanding (or common ground). Ward and Litman, for instance, showed that coherence was predictive of learning in tutoring dialogues [11]. There has been a significant amount of additional work on this topic, in various domains. We won't summarize the literature on coherence, but the interested reader can look at the work done around Coh-Metrix [3] for more information.

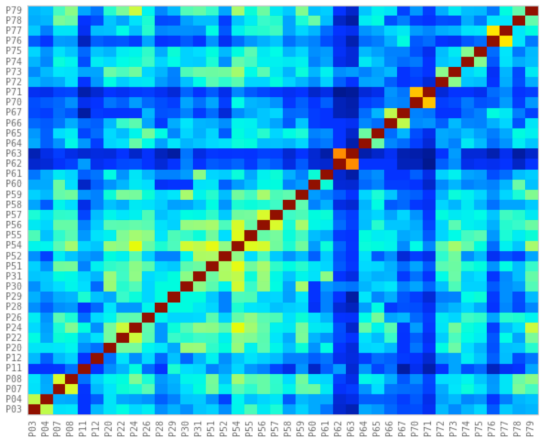

Figure 6: cosine similarity between each participant of the experiment. The diagonal is red because it represents each students' perfect similarity with herself / himself.

The first step of the process 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 information retrieval to score documents' relevance to a query. We then compared each student's discourse similarity with other participants by using a cosine similarity measure over the entire transcripts. A cosine 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 6: dark blue lines show students who are very dissimilar to everyone else; hot colors 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, and 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.

Additionally, we tried to reorganize students on each axis based on their learning scores (Fig.7, left side) and their quality of collaboration (Fig.7, right side). The first approach did not cluster students in any meaningful way; however, the second one showed 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.

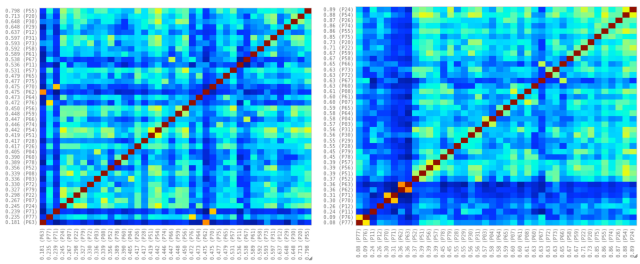

Figure 7: cosine similarity matrix, reorganized with students' learning scores (left) and quality of collaboration (right).

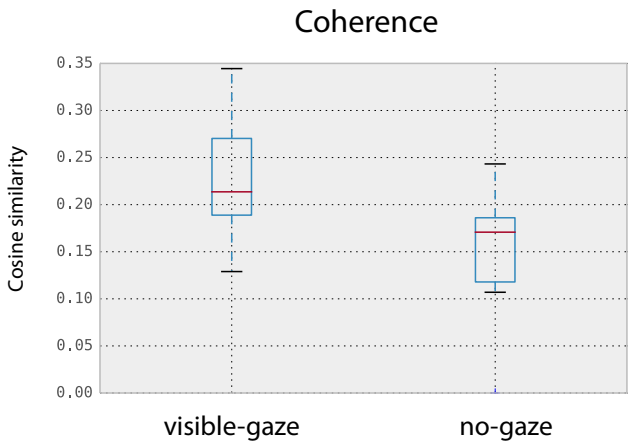

Figure 8: Students' coherence when discussing the task. Students in the "visible-gaze" group were significantly more coherent (p < .05); higher coherence was also significantly correlated with higher learning gains (p < .05).

We then computed a first measure of students' *coherence*: while our approach was simplistic (more complicated measures of coherence do exist [3]), it provided an approach relatively easy to understand and to apply. We built on our previous results using tf-idf and cosine similarity to assess whether students were re-using 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-exchanges window through the transcript and averaged the similarity across all exchanges to compute our measure of coherence. Using this measure, we found that students in the "visible-gaze" condition were more coherent than students in the "no-gaze" condition (Fig. 8): $F(1,20) = 7.45$, $p = 0.01$, 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). This measure was positively correlated with students' learning gain: $r(19) = 0.540$, $p = 0.011$ (Fig. 9). *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*.

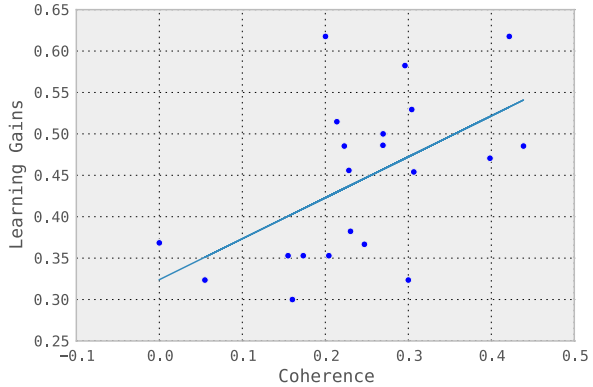

Figure 9: Correlation between dyads' dialogue coherence and learning gain: r(19) = 0.540, p = 0.011.

On a side note, we tried various values for *n* and *m*. Some of those results were not significant, but we always found that students in the "visible-gaze" group were more coherent than students in the "no-gaze" group. At the end, we observed that comparing 5 exchanges with the 5 previous utterances produced the results that were the clearest and easier to interpret.

Here we provide an example of a highly coherent exchange (cosine similarity of 0.5). We highlighted similar words between the two sets of utterances in bold:

--- *Exchange 1* ---

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, < mumbling to self >, the **number** for that side is gonna be, um, this is tricky business.

A: Yeah it is. < Laughs >.

--- *Exchange 2 (same discussion, continued)* ---

B: **Kind of**? < Laughs >.

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* ---

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, is in line with the results of [5], who showed that convergence is not only associated with conceptual understanding but also with affective components such as frustration, engagement and confusion.

3.4 ADDITIONAL RESULTS

In a subsequent step, we sought baselines to use for comparing students' utterance corpora. For instance, we can imagine that comparing the transcripts of students with a baseline of an expert discussion on this topic would be predictive of their learning gains. To this end, we used two corpora as references: first, we used the best student (in terms of her learning score) of our dataset (P55). 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 2nd step of the experiment into our dataset. This text is highly technical and is likely to pick up students' use of the particular terminology associated with this domain.

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).

Interestingly, this measure was positively correlated with students' quality of collaboration: $r(38) = 0.545$, $p < 0.001$. There wasn't any 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 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, since students' utterances reflect the way novices discuss and learn about a new topic; a scientific text, on the other hand, is produced by experts who have mastered the concepts and terminology of a domain. In sum, those two features could be advantageously used to further explore students' discussion, as well as to feed machine learning algorithms trying to predict students' learning.

3.5 PUTTING OUR MEASURES TOGETHER: PREDICTING STUDENTS' QUALITY OF COLLABORATION AND LEARNING GAINS USING LINGUISTIC FEATURES

Our final contribution is to test whether the measures described above have any predictive value. More 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 students' 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-labeled categories from section one (n-grams), the cosine similarity scores, the convergence measures and the coherence metrics as features. The complete dataframe contained 60 features and 40 rows. 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). For the learning scores, we found that SVM with a multilayer perceptron kernel and 8 features could correctly classify 94.44% of our participants. We also used a *validation set* (4 participants, which constitutes 10% of our sample). 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).

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 datapoints). Finally, those results should be contrasted with other baselines, such as decision trees or naïve bayes.

Table 2: Rough classification of students (using a median-split) in terms of their learning gains.

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

In sum, these analyses indicate noteworthy promise in using linguistic features to predict students' learning and ability to collaborate with their peers, but those results need to be replicated on larger datasets to be truly convincing.

Interestingly, SVM selected some of the correlations we found above between students' learning gains and particular features of our transcripts: number of anaphoras used and keyword showing students' uncertainty. However, other measures such as coherence, cosine similarity with a textbook chapter were not included in our final model. Instead, it favored low-level measures, such as the number of words used by students, the length of their sentence and particular grammatical forms (negations, auxiliary verbs, prepositions). This shows that some variables may be good predictors in isolation, but lose their predictive power when associated with other measures.

4. DISCUSSION

The goal of this project was to explore various NLP techniques to make sense of educational datasets; we favored 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 to examine not only the cosine similarity to the best student of the other students' transcripts but to more aggregate exemplars of 'better or worse students', such as the upper and lower quartile of the students in terms of learning score.

To recap our results, we have found that: 1) n-grams probabilities can help characterize groups of students in terms of building a common ground with their partners (anaphoras); 2) 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) coordination of linguistic style has little predictive power in terms of explaining dyads' collaborative learning processes; 4) coherence measures, on the other hand, are positively associated with students' learning 5) using SVM and the features mentioned above, we can roughly predict students' learning outcomes with an accuracy higher than 90% (which dropped to 75% for our 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 students'

discourse, and we found interesting predictors for learning gains and students' collaboration quality. However, we also argue that those techniques could be used in other domains. For instance, comparing the similarity between a reference text and students' utterances has already been used for assessing essays. Coherence can be used in similar contexts. More interestingly, those metrics could be advantageously used on multi-modal datasets. Eye-tracking data, for instance, could be converted in a series of word tokens representing the location of students' 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 into the ways that students construct their understanding of a particular concept, and to 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). Replicating those results on larger datasets would make a more convincing argument for using NLP measures in education.

5. CONCLUSION

This paper showed 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 and deepening our sense of their predictive value; replicating those results on other datasets; and exploring additional topics in NLP (e.g., topic modeling with Latent Semantic Analsyis or Latent Dirichlet Allocation).

6. ACKNOWLEDGMENTS

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

7. REFERENCES

1. Clark, H.H. and Brennan, S.E. "Grounding in communication." In L.B. Resnick, J.M. Levine and S.D. Teasley, eds., *Perspectives on socially shared cognition*. American Psychological Association, Washington, DC. 1991, 127–149.
2. Danescu-Niculescu-Mizil, C. and Lee, L. 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*, Association for Computational Linguistics (2011), 76–87.
3. Graesser, A.C., McNamara, D.S., Louwerse, M.M., and Cai, Z. "Coh-Metrix: Analysis of text on cohesion and language." *Behavior Research Methods, Instruments, & Computers 36*, 2 (2004), 193–202.
4. Meier, Anne, Hans Spada, and Nikol Rummel. "A rating scheme for assessing the quality of computer-supported collaboration processes." *International Journal of Computer-Supported Collaborative Learning, 2*, 1 (2007), 63-86.
5. Mitchell, C.M., Boyer, K.E., and Lester, J.C. 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 Dialogue*, Association for Computational Linguistics (2012), 94–98.
6. Pickering, M.J. and Garrod, S. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences 27*, 02 (2004), 169–190.
7. Pennebaker, J W., M. E. Francis, and R. J. Booth. "Linguistic inquiry and word count: LIWC 2001." *Mahway, NJ: Lawrence Erlbaum Associates,* 2001, 71.
8. Schneider, B. and Pea, R. Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning 8*, 4 (2013), 375–397.
9. Sherin, B. "Using computational methods to discover student science conceptions in interview data." *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. New York: ACM, (2012).
10. Tomasello, M. "Joint attention as social cognition." In C. Moore and P.J. Dunham, eds., *Joint attention: Its origins and role in development*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995, 103–130.
11. Ward, A. and D. Litman. Dialog convergence and learning. In *International Conference on Artificial Intelligence in Education* (AIED). 2007. Los Angeles, CA.