Modeling Collaborative Discourse with ENA using a Probabilistic Function

Yeyu Wang[1[0000–0003–1978–5453]], Andrew R. Ruis[1[0000-0003-1382-4677]] and David Williamson Shaffer[1[0000–0001–9613–5740]]

[1] University of Wisconsin – Madison, Madison, WI, USA
ywang2466@wisc.edu

Abstract. Models of collaborative learning need to account for *interdependence*, the ways in which collaborating individuals construct shared understanding by making connections to one another's contributions to the collaborative discourse. To operationalize these connections, researchers have proposed two approaches: (1) counting connections based on the presence or absence of events within a temporal window of fixed length, and (2) weighting connections using the probability of one event referring to another. Although most QE researchers use fixed-length windows to model collaborative interdependence, this may result in miscounting connections due to the variability of the appropriate relational context for a given event. To address this issue, we compared epistemic network analysis (ENA) models using both a window function (ENA-W) and a probabilistic function (ENA-P) to model collaborative discourse in an educational simulation of engineering design practice. We conducted a pilot study to compare ENA-W and ENA-P based on (1) interpretive alignment, (2) goodness of fit, and (3) explanatory power, and found that while ENA-P performs slightly better than ENA-W, both ENA-W and ENA-P are feasible approaches for modeling collaborative learning.

Keywords: Collaborative Learning, Learning Analytics, Epistemic Network Analysis, Modeling Recent Temporal Context, Engineering Education.

1 Introduction

A critical element of collaborative learning is that learners co-construct knowledge and make cognitive connections both intrapersonally and interpersonally [1]. That is, a learner forms links (a) between concepts that they themselves contribute to collaborative interactions and (b) between their own contributions and those of their collaborators. These links are operationalized in models of collaborative learning as connections between concepts. To accomplish this, Suthers et al. [2] suggest that in collaborative discourse, common ground can be represented in terms of the *recent temporal context* for an utterance: that is, the common ground for the current utterance in a conversation is composed of the utterances that precede it back to some prior point in time. Because both manual construction and natural language processing techniques face challenges in determining recent temporal context for each utterance at scale [3][4], models of collaborative learning approximate the appropriate recent temporal context for utterances using a *fixed-length moving window* or using a *probabilistic function*. Each of these approaches has advantages and disadvantages in approximating recent temporal context. Fixed windows are easy to compute but may over- or under-count connections in the model. Probabilistic models may, in some settings, be more accurate, but they can be more difficult to implement.

In this study, I examine one technique for modeling cognitive connections in collaborative contexts, Epistemic Network Analysis (ENA), which has been implemented primarily with a fixed-length moving window to operationalize recent temporal context. Using one dataset, I examine whether a novel approach, ENA with a probabilistic model, better models the process of collaborative learning.

2 Theory

2.1 Modeling Collaborative Learning

One important component of learning is the process by which individual learners work together to develop cognitive connections between concepts [5][6]. For example, Suthers et al. [2] refer to a particular type of connection in collaborative learning, *uptake,* as the process of one student contributing to the conversation based contributions of another. Clark [7] in turn argues that critical to the notion of uptake is the concept of *common ground*: the shared knowledge and assumptions across individuals, groups, and communities that are relevant to a specific turn of talk. As Suthers and Desiato [8] suggest, in collaborative discourse, common ground can be operationalized as *recent temporal context*: that is, the common ground for some current utterance in a conversation is composed of the utterances that precede it back to some prior point in time.

Thus, Swiecki [1] argues that *interactivity* and *interdependence* are fundamental to collaborative learning. Interactivity refers to the process through which learners co-construct knowledge by responding to others' opinions or actions. This interactivity results in interdependence—that is, one learner's utterances or actions influence others. In other words, all learning can be characterized, at least in part, as a process of making connections between ideas. In collaborative learning, those connections are made from a learner's ideas to some collaborative recent temporal context.

2.2 Quantifying Interdependent Connections

In cognitive science, scholars make two claims about how humans understand information and make connections within the common ground. Each of these claims leads to a different approach to modeling connections.

Counting cognitive connections based on presence or absence of events within the window. The first claim is that people have limitations on their capacity for processing information. [9] argues that a speaker engages a listener's attention during a conversation by dividing big chunks of information into smaller, logically connected units, which is termed as *intonation units*. To construct appropriate intonation units, a speaker needs to make an assumption at each moment about how much understanding they share with others in a conversation. Thus, there is an underlying assumption about *what each person thinks the others can remember* in the process of conversational uptake. Based on this framework, researchers operationalize cognitive connections within the recent temporal context using a *fixed-length moving window*. The window is *fixed* on the assumption that all of the participants make a similar assumption about other participants' shared understanding. The window is *moving* in the sense that each line in the dataset has a window that represents its recent temporal context. Within each window, researchers develop *indicators* to described learning patterns, such as whether or not two events co-occur within the window.

Weighting cognitive connections using the probability of one utterance referring to another. In addition to *how much* information a person can hold within their short-term memory, previous research has investigated *how likely* it is that a person can retain some piece of information in short-term memory as time passes. For example, Ebbinghaus [10] studied rates of retention and forgetting based on a test of vocabulary recall, resulting in an exponential decay function. Other research on information retention models forgetting based on power functions [11][12]. Regardless of what function we use to model information retention, this perspective suggests that the *probability* of recalling information decays as time passes. Rather than claiming the connection strength between two codes is either 1 or 0, I propose to quantify the connection strength as a function of *distance* between the two lines where codes occur. To operationalize connection strength, I use a probabilistic function to model recent temporal context in collaborative discourse. The probabilistic function estimates *the probability of one utterance referring to another* based on the distance between two utterances.

2.3 Epistemic Network Analysis

Epistemic Network Analysis (ENA) is an approach to quantifying connections between concepts, behaviors and other elements to model collaborative learning [13]. ENA takes coded data generated by individuals during interactivity and represents connections among those codes as a network structure. Specifically, ENA computes connections based on the *presence* of the codes in each line of data and the codes in the previous lines of data that constitute its recent temporal context. That is, ENA currently is operationalized based on the first approach in 2.2. Using this approach to model a dataset, we need to construct a window for each utterance. However, the challenge is that not every line has the same window size. Ruis et al. [14] proposed to resolve this issue by choosing a *fixed* length for the window as a best *approximation*. They argue that the window size needs to be sufficient to capture the recent temporal context for 95% of utterances in a dataset and minimize improperly-included connections within the fixed size of the window. As Shaffer [13] argues, a fixed window is a good approximation because even though some responses are not direct responses to preceeding lines, they are part of the common ground. This kind of response is called a *dispreferred response* [15], a contribution that is “not an expected and direct reply to prior referents” ([13], p. 159). Therefore, even if covered by the fixed-length window, such dispreferred responses are still in the common ground and should be included in the recent temporal context.

In what follows, I refer to these two requirements for a fixed window model as the *maximum window postulate* and the *dispreferred response postulate*. The first says that we should choose a large fixed-length window to cover the recent temporal context for majority of utterances; and the second says that in choosing a large fixed window, we believe that dispreferred responses should be included in the recent temporal context.

Problems with Fixed-length Window. The fixed-length window method thus depends on the dispreferred response postulate. That is, dispreferred responses should always be included in the recent temporal context for an utterance. However, it is not clear that this is always true. If we ignore the maximum window postulate and choose a shorter window, we will exclude lines which should be in the recent temporal context. That is, we produce a *Type II error* or a *false negative*, where we do not count connections which are relevant. If we follow the maximum window postulate, then we may include irrelevant responses within the window. That is, we produce a *Type I error* or a *false positive*, where we count connections which are actually irrelevant. This happens because, in this case, the dispreferred response postulate is not always valid: we cannot consider all dispreferred responses as relevant context for future utterances. Thus, situations where the dispreferred response postulate fails necessarily result in either Type I or Type II errors: either *overcounting* (Type I) or *undercounting* (Type II) connections.

Miscounting connections in an ENA model can lead to interpretive misalignment: the ENA model may include irrelevant connections or exclude relevant connections, which means the model is not aligned with a qualitative understanding of the data. Overcounting or undercounting connections also introduces error when constructing an ENA model, which may result in lower goodness of fit or lower the amount of variance explained.

2.4 Research Question

To address the issue of fixed window approach, I test whether ENA with a probabilistic approach provides a better model of collaborative discourse than a fixed-length window. In this study, I apply both ENA with a fixed-length *window* model (ENA-W) and ENA with a *probabilistic* model (ENA-P) to analyze the collaborative problem-solving processes in an engineering design training program, which consists of two primary learning activities: in the first half, student project teams explore a design space using a single material component, and in the second half, they attempt to create an optimal design using any available material. I compare these two models to answer the following research questions:

(1) Does ENA-P exhibit better interpretative alignment between qualitative and quantitative results than ENA-W?

(2) Does ENA-P have a better goodness of fit than ENA-W?

(3) Does ENA-P explain more variance between the two learning activities than ENA-W?

3 Methods

3.1 Dataset and Codebook

The dataset was collected from the virtual internship *Nephrotex* [16]. Nineteen engineering students participated in an online training simulation in which they designed a nanotechnology-based membrane for kidney dialysis machines at a fictitious company. The training program was divided into two activities. The goal for the first half was to help students explore the design space and learn the functional characteristics of a single material by analyzing graphs and data and conducting tests. The goal for the second half was to help students optimize the performance of a design across multiple parameters using a range of materials and other components. During these two activities, students communicated with their peers and a mentor through a persistent online chat tool that was a part of the simulation environment. To analyze the collaborative processes in this learning environment, researchers collected all 1443 chat posts across the 10 different groups in the simulation (5 in the first half, and 5 jigsawed groups in the second half). The chat posts were labeled by username and group number and arranged in a chronological order within each group.

To analyze the collaborative discourse, Siebert-Evenstone et al. [17] developed and validated a coding scheme with six codes: (1) PERFORMANCE PARAMETERS: criteria used to assess the design prototype including cost, marketability, reliability, flux, and blood cell reactivity; (2) DESIGN-BASED DECISION MAKING: processes of making design decisions, including prioritization and tradeoffs; (3) CLIENT AND CONSULTANT REQUESTS: concerns or needs of stakeholders in the simulation including suggestions and requirements for the final product; (4) DATA: specific technical or numeric information; (5) COLLABORATION: teamwork during decision making, including discussion of a team's collective action (e.g., "we need to…"); (6) TECHNICAL SPECIFICATIONS: characteristics of design prototypes, including selected materials, transformation processes, surfactant, and carbon nanotube precentage. All 6 codes were validated by two trained human raters (for each code, kappa > 0.83, $\rho(0.65) < 0.05$).

3.2 ENA

ENA takes binary-coded data as input and then constructs a fixed-length moving window to calculate connection counts between codes for each utterance. For each unit, ENA aggregates connection counts by summing across all windows for that unit's utterances. The aggregated connection counts are represented as an *adjacency vector*. ENA normalizes and centers the adjacency vectors, and the terms are used as *line weights* between nodes in the network representation. ENA performs a dimensional reduction technique to reduce the high-dimensional adjacency vectors to a low-dimensional space. In this study, the first dimension was constructed using a means rotation that maximizes the variance between the two primary activities in the simulation, and the second dimension was constructed using singular value decomposition, which maximizes variance among all units. ENA optimizes node positions in the resulting low-dimensional space to align the network centroids (based on line weights) with ENA scores (based on the dimensional reduction). To measure how aligned centroids and ENA scores are, ENA calculates the *goodness of fit* using Pearson's *r* correlation between these two values on all dimensions.

In the network representation, codes are represented as nodes, while connection strengths are represented by edge thickness and saturation. To compare patterns of connection-making in the first-half and second-half of activities, ENA creates a visualization called *difference plot*. That is, ENA calculates the mean line weights for units in each simulation activity separately and subtracts one group of mean line weights from the other, visualizing the differences with the color and thickness of the edges.

While network visualizations provide insights about different patterns of making connections between groups, ENA scores can be used to test whether these differences are *statistically* significant. In this analysis, I regressed the ENA scores from two different ENA models on a grouping variable of two activities. To test whether the variances explained by ENA-W and ENA-P are significantly different, I bootstrapped units and computed both regressions repeatedly, which created an empirical distribution of R^2 for both models. I applied Fischer's Z transformation and used a Monte Carlo rejection method to determine whether the difference in variance explained by the two ENA models was significant.

As a unified approach to data analysis, ENA integrates both qualitative interpretation and quantitative representation of data. Researchers establish *interpretive alignment* by showing that the conclusions derived from an ENA model is aligned with some qualitative interpretation of the original data. In my study, I checked the interpretive alignment in two ways:

- *Individual-Level*: I identified two segments of discussion and manually evaluated whether ENA-P addresses the potential over- and under-counting problems introduced by ENA-W.
- *Site-Level*: I evaluate whether the connection strengths in ENA-W or ENA-P provides a better representation of the expected outcomes based on the learning objectives for two activities.

3.3 Construction of ENA-W and ENA-P

Determining an Appropriate Window Size. To determine the window size for ENA-W, I adopted the method proposed by [14]. Two researchers randomly sampled 177 utterances from *Nephrotex* chat logs and determined the *furthest referent* for each utterance using social moderation. As proposed by [14], I identified the window size to be 7 utterances, accounted for recent temporal context in more than 95% of the sampled lines.

Determining an Appropriate Probabilistic Function. We derived the probabilistic function based on the same 177 samples. We define the sampled lines as an ordered set of lines $(l_1, l_2, \dots, l_{177})$. Each line l_i has its furthest referent l_{x_i}. Based on our definition of the recent temporal context, each line is related to its referents and itself. Thus, we operationalized the window to represent the recent temporal context is an ordered set of lines, $W_i = (l_{x_i}, l_{x_i+1}, \dots, l_i)$, where $|W_i| = i - x_i + 1$ (that is, the number of lines from l_{x_i} to l_i, inclusive). For each line, l_i, we identified its furthest referent, l_{x_i}. We then constructed a histogram of window sizes, $|W_i|$, as shown in Fig. 1.

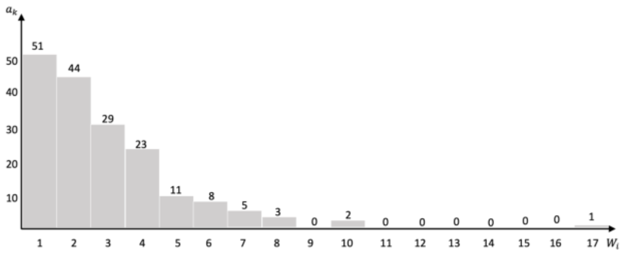


Fig. 1. Histogram for $|\boldsymbol{W_i}|$ based on 177 Sampled Lines

The height of bars in the histogram indicates the total *counts* of a referring line l_r with the window length of W_i. We define the counts of referring lines given a window length k as $a_k = |\{i \mid w_i = k\}|$. Let e be the maximum window length, $e = \max(\mathrm{W_i})$. The frequency distribution of window sizes lets us estimate the probability of a referring line (l_i) is related to any proceeding lines (l_λ), $\lambda < i$. We can also define the probability function $\pi(i - \lambda) = P(|W_i| > i - \lambda)$, which estimates the probability of the prior line l_λ is related to the referring line l_i. Thus, the frequency distribution of window sizes for all referring lines can be written as:

$$\pi(i - \lambda) = P(|W_i| > i - \lambda) = 1 - \frac{\sum_{j=0}^{i-\lambda} a_j}{\sum_{k=1}^{e} a_k}, a_0 \equiv 0.$$

4 Results

4.1 Research Question 1: Interpretive alignment

Individual-Level Interpretive Alignment. In this section, I examined two examples from one student, Lily, who participated in the training program: one example illustrates that ENA-W may overcount connections, and the other illustrates that ENA-W may undercount connections. In both examples, I conducted a qualitative analysis on the recent temporal context of one utterance and manually derived the adjacency matrix of connection strengths using ENA-W and ENA-P.

Overcounting Problem. In the following example, a group of students is discussing different prototypes and evaluating their performance in preparation for choosing a final design:

Line	Team Member	Utterance
1	Abby	sounds good
2	Jina	Well the only other prototype i would consider is the one that was comprised of PMMA, vapor, using a hydrophilic sufractant, with a nanotube % of 4.0%
3	Jina	this cost $100 dollars per unit, sold 500,000, had a reliability of 12 hours with a flux of 15 but a blood cell reactvity of 54.44

In line 1, Abby comments on a previous design, indicating that it “sounds good” as a candidate for the final prototype. Then Jina (line 2) proposes another candidate design “comprised of PMMA, vapor, using a hydrophilic sufractant, with a nanotube % of 4.0%.” That is, she lists TECHNICAL SPECIFICATIONS as inputs for the design. In line 3, Jina continues by describing the PERFORMANCE PARAMETERS of the prototype, including cost, marketability, reliability, flux and blood cell reactivity. She further adds that the performance on the first four parameters is great, “but … blood cell reactivity” is low at “54.44”.

After summarizing the TECHNICAL SPECIFICATIONS and PERFORMANCE PARAMETERS of her prototype, Jina suggests that her teammates type the values of the PERFORMANCE PARAMETERS for their prototypes in the chat (line 4), which will be used to justify their design choices:

Line	Team Member	Utterance
4	Jina	If you guys could, can you type out the the information from the prototypes on chat. We need it for the justifications
5	Bob	Flux: 29 BCR: 65.56 Reliability: 9 Marketability: 900,000
6	Abby	This resulted in a reliability of 8 hours, marketability of 600,000 units, a flux rate of 13m^3/m^2-day, and a low Bloodcell reactivity of 21.11. In total this prototype costs $130 dollars per unit
7	Lily	The BCR Type: Reliability-5, Market-800,000. Flux - 23. BCR- 10. Cost -$150

In response to Jina’s proposal, Bob (line 5), Abby (line 6), and Lily (line 7) all enter the numerical values for the PERFORMANCE PARAMETERS of their prototypes. That is, they summarize how each of their prototypes performed on the metrics that the stakeholders care about.

Jina’s response in line 4 is thus a dispreferred response. The previous discussion (lines 1-3) was focused on the TECHNICAL SPECIFICATIONS of the prototypes that each team member designed and tested. Jina (line 4) then abruptly shifted the discussion to the PERFORMANCE PARAMETERS of the prototypes, effectively beginning a new discussion.

How, then, should we model Lily’s utterance in line 7? Based on the fixed window with 7 utterances, all lines in this segment, including any dispreferred responses, are relevant context because they are within the window, which in turn quantifies the connection between TECHNICAL SPECIFICATIONS and PERFORMANCE PARAMETERS as **1**.

In other words, the window models a connection between PERFORMANCE PARAMETERS and TECHNICAL SPECIFICATIONS for line 7 even though qualitatively, Lily was not making that connection.

If we use a probabilistic function to quantify connection strength, line 7 comes 5 lines after line 2. Thus, the connection between Lily’s reference to PERFORMANCE PARAMETERS in line 7 and Jina’s reference to TECHNICAL SPECIFICATIONS is weighted by $\pi(7-2) = \pi(5) = \mathbf{0.107}$. Thus, the probabilistic model also shows a connection, but now with a weakened strength of only 0.107. This adjustment of connection strength suggests that the probabilistic model is a better representation—or perhaps in this case, a less imperfect representation—of Lily’s response.

Undercounting Problem. In the following example, which comes at the beginning of the second half of the training program, students have switched groups. In their new group they introduce themselves, and then Abby describes (line 1) the conclusion by their team in the first half, suggesting that it was not particularly useful for designing a final prototype:

Line	Team Member	Utterance
1	Abby	basically the only thing i was able to conclude from my surfactant was that the BCR was constant in all of the prototypes. was 43.33%
2	Jina	The group i had previously worked with came up with a prototype that gave us an all around great dialyzer. It was comprised of PMMA for the material, Used the process of Vapor, and used a biological sufractant, and had a nanotube percentage of 1.5%.
3	Jina	This resulted in a reliability of 8 hours, marketability of 600,000 units, a flux rate of 13m^3/m^2-day, and a low Bloodcell reactivity of 21.11. In total this prototype costs $130 dollars per unit
4	Abby	submit that protoype label in Team1 Batch1
5	Abby	do we want to stick to a specific material
6	Lily	I think we should include one prototype of each material.
7	Abby	okay so each of us creates one from our material
8	Jina	Well what was the best one out of the previous prototypes for each material? It makes sense to do the best ones overall for each.
9	Lily	well, change it up a bit. You can optimize your best result.

Jina replies to Abby by reporting (line 2) the TECHNICAL SPECIFICATIONS for “an all around great dialyzer” that their group tested in the first half of the training program. Then, she provides (line 3) DATA about the PERFORMANCE PARAMETERS for her prototype.

Abby replies by suggesting (line 4) that the team use one of their prototypes from the first half of the training program in their next submission (“submit that prototype label in Team1 Batch1”). She asks (line 5) whether the whole team should use one material for the final submission. Lily replies (line 6) to Abby's question, saying that each person on the team should test a prototype for a different material. Abby confirms (line 7) that she understands what Lily had said: each student on the new team should design a prototype using the material studied by their old team. In response to Lily and Abby, Jina suggests (line 8) that they should use the "best ones overall" from their previous team. However, Lily disagrees and suggests (line 9) that they should consider changing the design from the previous team to achieve the best result possible.

This final comment about DATA (the "best result" of a design using one material) is thus a response to the previous 8 lines where students were deciding how to move to the next phase of their design process. More specifically, it relates to Jina's description (line 2) of the TECHNICAL SPECIFICATIONS for one specific device and its PERFORMANCE PARAMETERS (line 3).

But notice that with a window size of 7, DATA (line 9) is connected to PERFORMANCE PARAMETERS (line 3)—which is aligned with this qualitative analysis of the example. However, it is not connected to TECHNICAL SPECIFICATIONS (line 2), even though they are part of what would have been read as a single continuous comment by the same student (Jina). That is, the connection calculated by ENA-W is **0**. In other words, the window excludes a connection between DATA and TECHNICAL SPECIFICATIONS for line 9 even though qualitatively, Lily was making that connection.

If we use a probabilistic function to quantify connection strength (see 3.2.1), line 9 comes 7 lines after line 2. Thus, the connection between Lily's DATA in line 9 and Jina's TECHNICAL SPECIFICATIONS is weighted by $\pi(9-2) = \pi(7) = 0.034$. In other words, in this example, a qualitative analysis shows that Lily was making a non-zeroconnection between DATA and TECHNICAL SPECIFICATIONS.

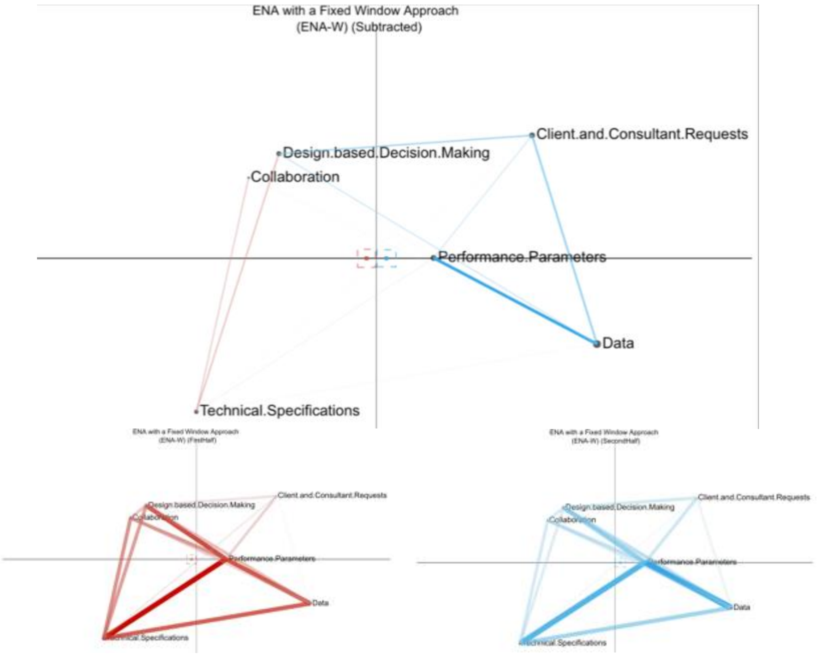
In summary, the probabilistic model (1) reduces the type I error by decreasing the connection strength between PERFORMANCE PARAMETERS and TECHNICAL SPECIFICATIONS, which is overcounted by the fixed-length window model and (2) reduces the type II error by increasing the connection strength between DATA and TECHNICAL SPECIFICATIONS, which is undercounted by the fixed-length window model. These patterns persist throughout Lily's network throughout the training program: The connection strength between PERFORMANCE PARAMETERS and TECHNICAL SPECIFICATIONS is 0.70 in the ENA-W model, which decreases to 0.67 in the ENA-P model; the connection strength between DATA and TECHNICAL. SPECIFICATIONS is 0.42 in the ENA-W model, which increases to 0.51 in the ENA-P model.

Site-Level Interpretive Alignment. To assess interpretive alignment at the site level, I constructed an ENA-W and ENA-P model based on the chats from the whole class during the training program. Recall that the training program was designed to help students learn two abilities: the goal of the first half is to explore the performance of a single material based on different data sources (e.g., technical reports and graphs) and experimentation, while the goal of the second half is to optimize the performance (i.e., cost, safety, reliability, etc.) of a design prototype using any available material.

Thus, we would anticipate that students in the first half of the training program would make more connections between DATA and TECHNICAL SPECIFICATIONS because they are spending more time reading and discussing technical reports, collecting preliminary data, and constructing graphs to understand various design attributes for one single material. Students in the second half are more likely to make connections between DATA and PERFORMANCE PARAMETERS, as they are designing and testing prototypes to better understand the design space and maximize device performance across a range of parameters.

According to Figure 5, the subtracted plot of ENA-P better aligns with expected difference between the first-half and the second-half of the training, based on the learning objectives of two halves. For example, students are expected to make more connections between DATA and TECHNICAL SPECIFICATIONS in the first-half of training. However, the edge between DATA and TECHNICAL SPECIFICATIONS in the subtracted plot for ENA-W is very weak, indicating little difference in the overall strength of that connection between the two halves: the edge weights differ by only 0.48. However, the ENA-P model shows that students made relatively more connections between DATA and TECHNICAL SPECIFICATIONS in the first half of training: the edge weights differ by 3.08. In other words, the ENA-W model does not reflect an expected difference in student discourse between the two halves of the simulation, while the ENA-P model does.

Similarly, students are expected to make more connections between DATA and PERFORMANCE PARAMETERS in their second-half, which is reflected in the subtracted plot of ENA-W: the edge weights differ by 1.60. However, the edge weight of this connection in ENA-P model shows even more salient difference, according to the thicker and darker blue edge. That is, the difference of this connections between two halves is larger in ENA-P: the edge weights differ by 4.01. In other words, the ENA-P model manifest and shows a more salient difference in network representation, compared to ENA-W.


ENA with a Fixed Window Approach
(ENA-W) (Subtracted)
Client.and.Consultant.Requests
Design.based.Decision.Making
Collaboration
Performance.Parameters
Data
Technical.Specifications
ENA with a Fixed Window Approach
(ENA-W) (FirstHalf)
Client.and.Consultant.Requests
Design.based.Decision.Making
Collaboration
Performance.Parameters
Data
Technical.Specifications
ENA with a Fixed Window Approach
(ENA-W) (SecondHalf)
Client.and.Consultant.Requests
Design.based.Decision.Making
Collaboration
Performance.Parameters
Data
Technical.Specifications

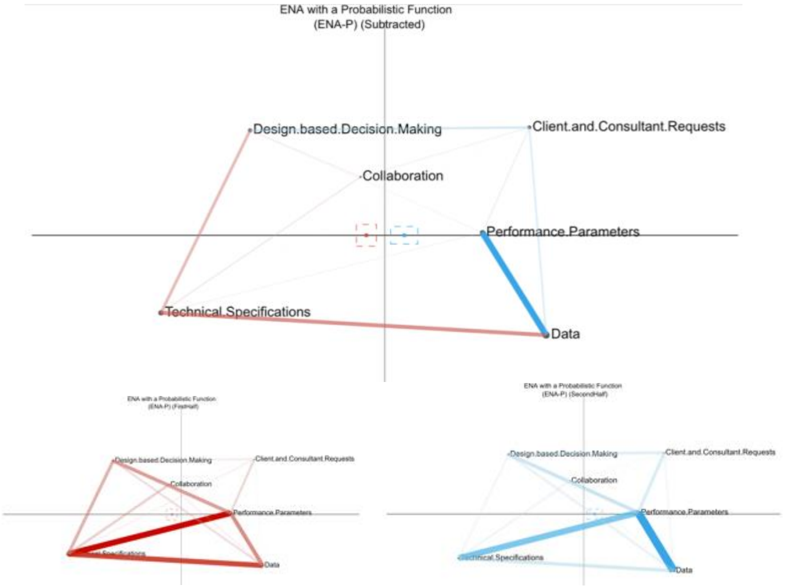

ENA with a Probabilistic Function
(ENA-P) (Subtracted)
Design.based.Decision.Making
Client.and.Consultant.Requests
Collaboration
Performance.Parameters
Technical.Specifications
Data
ENA with a Probabilistic Function
(ENA-P) (FirstHalf)
Design.based.Decision.Making
Client.and.Consultant.Requests
Collaboration
Performance.Parameters
Specifications
Data
ENA with a Probabilistic Function
(ENA-P) (SecondHalf)
Design.based.Decision.Making
Client.and.Consultant.Requests
Collaboration
Performance.Parameters
Technical.Specifications
Data

Fig. 5. Subtracted ENA Plots and Group ENA Plots Using Fixed Window Approach and Probabilistic Function

Thus, in the individual-level, the individual network of Lily using ENA-P is more aligned with the qualitative evidence; in the site-level, the subtracted plot using ENA-P is more aligned with the expected difference based on the design and intervention of the. Thus, ENA-P models collaborative learning process and achieves a better interpretive alignment, compared to ENA-W.

4.2 Research Question 2: Evaluation of ENA Models Using Goodness of fit

As described in 3.2., goodness of fit is a measure of discrepancy between dual representations for units. A higher goodness of fit provides a stronger warrant for the interpretation of the ENA scores based on the individual network. While goodness of fit for ENA-W is 0.93, goodness of fit for ENA-P is higher at 0.96. Thus, ENA-P has a better co-registration between dual representations than the ENA-W model.

4.3 Research Question 3: Evaluation for ENA Models using Regression Analysis and Variance Explained

As described in 3.2., I applied a two bivariate regression models to predict ENA scores on the primary axis for the ENA-W and ENA-P and based on the condition of first-half and second-half of the game. Condition of first versus second half significantly predicts ENA scores for both models. However, the variance explained by ENA-W ($R^2 = 0.22$) is lower than the variance explained by ENA-P ($R^2 = 0.38$).

To explore whether the variance explained is significantly different between two models, as described in section 3.2., I bootstrapped units from the whole set and ran the regression models repeatedly. With 1,000 iterations of bootstrapping, I calculated the 95% confidence interval (CI) for the R^2 of both models. The results show that the ENA-W model (95% CI $\in$ [0.20, 0.24]) has significantly lower variance explained than the ENA-P model (95% CI $\in$ [0.39, 0.43]). Thus, ENA-P has more explanatory power in accounting for differences between students in the first-half and second-half of the training program

5 Discussion

This study explored two approaches to modeling collaborative learning in which the unit of analysis is individuals-in-a-group. Specifically, it compared ENA models constructed using two different methods for quantifying the strength of connections in collaborative discourse: (a) a fixed-length window approach (ENA-W), which quantifies connections as either present or absent within a set number of turns of talk; and (b) a novel probabilistic function approach (ENA-P), which estimates the likelihood that a connection is present. I hypothesized that ENA-P would better address the problem of over- or undercounting connections—that is, incorrectly quantifying connection strength—when a fixed-length window is used. To test this hypothesis, I conducted a pilot study to test the feasibility of ENA-P relative to ENA-W using data from 19 students who participated in a collaborative engineering design training program.

I compared ENA-P with ENA-W using three criteria: interpretive alignment, variance explained between groups, and model goodness of fit. Both models performed well, but ENA-P achieved slightly higher goodness of fit, explained significantly more variance, and was better aligned with both qualitative interpretation and expected learning processes based on the design of the training program. At the individual-level, given two discourse segments involving one particular student, ENA-P better quantified the connections overcounted or undercounted by ENA-W. Furthermore, this pattern persisted when all connections were aggregated for this student. At the site-level, the ENA-P model better reflected expected differences in student discourse between the first and second halves of the training.

This pilot study suggests that in at least some collaborative learning contexts, ENA-P may perform better than ENA-W; thus, ENA-P is a feasible method for quantifying connections in ENA models. While ENA-P models may perform better in some circumstances, they are also more difficult to construct. For example, in this study, we manually identified a probabilistic function based on an empirical distribution, which takes more time and effort. There are also other possible probabilistic models, such as exponential functions or power functions; thus, the findings of this study suggest that such approaches should be tested in future work .

In summary, while ENA-P performs slightly better than ENA-W based on the pilot test, both ENA-W and ENA-P are feasible approaches to model collaborative learning.

References

1. Swiecki, Z.: Measuring the Impact of Interdependence on Individuals During Collaborative Problem-Solving. JLA. 8, 75–94 (2021).
2. Suthers, D.D., Dwyer, N., Medina, R., Vatrapu, R.: A framework for conceptualizing, representing, and analyzing distributed interaction. Computer Supported Learning. 5, 5–42 (2010).
3. Rose, C., Wang, Y.C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F.: Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. International Society of the Learning Sciences. 3, 237–271 (2008).
4. Espinoza, C., Lämsä, J., Araya, R., Hämäläinen, R., Gormaz, R., Viiri, J.: Automatic content analysis in collaborative inquiry-based learning. In European Science Education Research Association Conference. University of Bologna (2019).
5. DiSessa, A.A.: Knowledge in pieces. In: Forman, G. and Pufall, P. (eds.) Constructivism in the computer age. pp. 47–70. Erlbaum, Hillsdale, NJ (1988).
6. Shaffer, D.W.: Models of Situated Action. In: Steinkuehler, C., Squire, K., and Barab, S. (eds.) Games, Learning, and Society. pp. 403–432. Cambridge University Press, Cambridge (2012).
7. Clark, H.H. ed: Common ground. In: Using Language. pp. 92–122. Cambridge University Press, Cambridge (1996).
8. Suthers, D.D., Desiato, C.: Exposing Chat Features through Analysis of Uptake between Contributions. In: 2012 45th Hawaii International Conference on System Sciences. pp. 3368–3377. IEEE, Maui, HI, USA (2012).
9. Chafe, W.: Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing. University of Chicago Press (1994).
10. Ebbinghaus, H.: Memory: A contribution to experimental psychology. Annals of neurosciences. 20, 155 (2013).
11. Rubin, D.C., Wenzel, A.E.: One Hundred Years of Forgetting: A Quantitative Description of Retention. Psychological review, 103(4), 734 (1996).
12. Wixted, J.T., Ebbesen, E.B.: Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. Memory & Cognition. 25, 731–739 (1997).
13. Shaffer, D.W.: Quantitative Ethnography. Lulu.com (2017).
14. Ruis, A.R., Siebert-Evenstone, A.L., Pozen, R., Eagan, B.R., Shaffer, D.W.: Finding Common Ground: A Method for Measuring Recent Temporal Context in Analyses of Complex, Collaborative Thinking. In: 13th International Conference on Computer Supported Collaborative Learning (CSCL), pp.136-143 (2019).
15. Pomerantz, A.: Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shaped. Structures of Social Action: Studies in Conversation (1984).
16. Chesler, N.C., Ruis, A.R., Collier, W., Swiecki, Z., Arastoopour, G., Williamson Shaffer, D.: A Novel Paradigm for Engineering Education: Virtual Internships With Individualized Mentoring and Assessment of Engineering Thinking. Journal of Biomechanical Engineering. 137, 024701 (2015).
17. Siebert-Evenstone, A.L., Arastoopour Irgens, G., Collier, W., Swiecki, Z., Ruis, A.R., Williamson Shaffer, D.: In Search of Conversational Grain Size: Modeling Semantic Structure using Moving Stanza Windows. Learning Analytics. 4, (2017).