An Empirical Bridge for Recurrence-Aware Conversation Decision Support

Mariia Yakovleva
mariia.creater@gmail.com

ABSTRACT

We report three linked pilot studies that test whether computational methods can recover meaningful structure from relationship-relevant text. Study 1 examines adjacent-turn dialogue carryover in a large movie-dialogue corpus (Danescu-Niculescu-Mizil & Lee, 2011). Study 2 analyzes emotional and interactional signals in online relationship-help posts (Entwistle et al., 2021). Study 3 evaluates whether simple text features can distinguish rough episode-trajectory labels in update-style relationship narratives. Across all three levels, shallow lexical and n-gram methods detect interpretable signal, but the signal remains limited. We argue that these results constitute an empirical bridge between a prior conceptual framework on recurrence-aware conversation decision support and more theory-aligned empirical work.

1. INTRODUCTION

The conceptual paper that precedes our empirical work proposed recurrence-aware conversation decision support as a distinct analytical frame for emotionally loaded interpersonal conversations. Our core claim was that difficult conversations, especially in romantic contexts, should not be treated as one-shot exchanges or mere wording problems. Instead, we argue they should be understood as decision environments shaped by repeated episodes, strategic objectives, branching response paths, and adaptive learning over time.

Relationship help-seeking occurs through both formal and informal channels, including therapy, education, books, internet resources, clergy, and friends or family (Stewart et al., 2016). Anonymous online platforms now provide large-scale naturalistic data on relationship help-seeking behavior (Entwistle et al., 2021). Communication during romantic conflict is consequential, but its effects depend on contextual factors rather than any single universally 'good' style (Overall & McNulty, 2017; Domingue & Mollen, 2009). Computerized lexical text analysis can provide psychologically informative, though necessarily coarse, indicators of language use (Tausczik & Pennebaker, 2010).

That conceptual argument created a clear theoretical direction, but it also left an empirical gap. Before recurrence, branching, objective-strategy alignment, and post-conversation adaptation can be modeled directly, we must first provide evidence that computational methods can recover any meaningful structure from conversational and relationship-related text at all. Our paper addresses that gap through three linked pilot studies that move from a lower-level unit of analysis to a higher-level one.

The spine of our paper is progressive. In Study 1, we focus on adjacent-turn dialogue structure and ask whether simple lexical features capture turn-to-turn carryover in a large dialogue corpus. In Study 2, we move into the target domain of relationship-help narratives and ask whether similarly simple features recover local emotional and interactional signals within single posts. In Study 3, we move one level higher and ask whether update-based romantic relationship narratives contain enough textual regularity to distinguish rough episode-status labels such as concluded and ongoing.

Our paper is therefore best read as an empirical bridge rather than a direct test of the full conceptual framework. Our goal is not to claim that recurrence-aware decision support has already been validated. Our goal is to show that there is measurable structure at multiple levels of relationship-relevant text, while being explicit that shallow lexical and n-gram methods remain insufficient for modeling deeper relational processes.

2. CROSS-STUDY DESIGN AND LOGIC

Our three studies are connected by an escalating unit of analysis. Together they trace a narrow but defensible path from local dialogue dynamics to coarse episode-level relational trajectory.

In Study 1, we examine adjacent dialogue pairs and ask: do simple lexical features capture turn-to-turn carryover in dialogue? Our contribution here is establishing a minimal turn-level operationalization of local response dynamics.

In Study 2, we analyze single relationship-help posts and ask: do simple lexical and discourse-level features recover emotional and interactional signals in relationship narratives? Our contribution is moving the analysis into the target domain of relationship-related support text.

In Study 3, we examine update-style relationship episodes and ask: can simple text features distinguish rough narrative update statuses such as concluded versus ongoing? Our contribution is introducing a coarse trajectory proxy at the episode level.

This structure matters because it prevents our three studies from reading as unrelated pilots. Their empirical value lies less in any one model than in their combined progression: adjacent turns, then relationship narratives, then update-based narrative trajectory. That progression is our paper's central design logic.

3. STUDY 1 - LOCAL TURN-TO-TURN SIGNAL CARRYOVER IN DIALOGUE

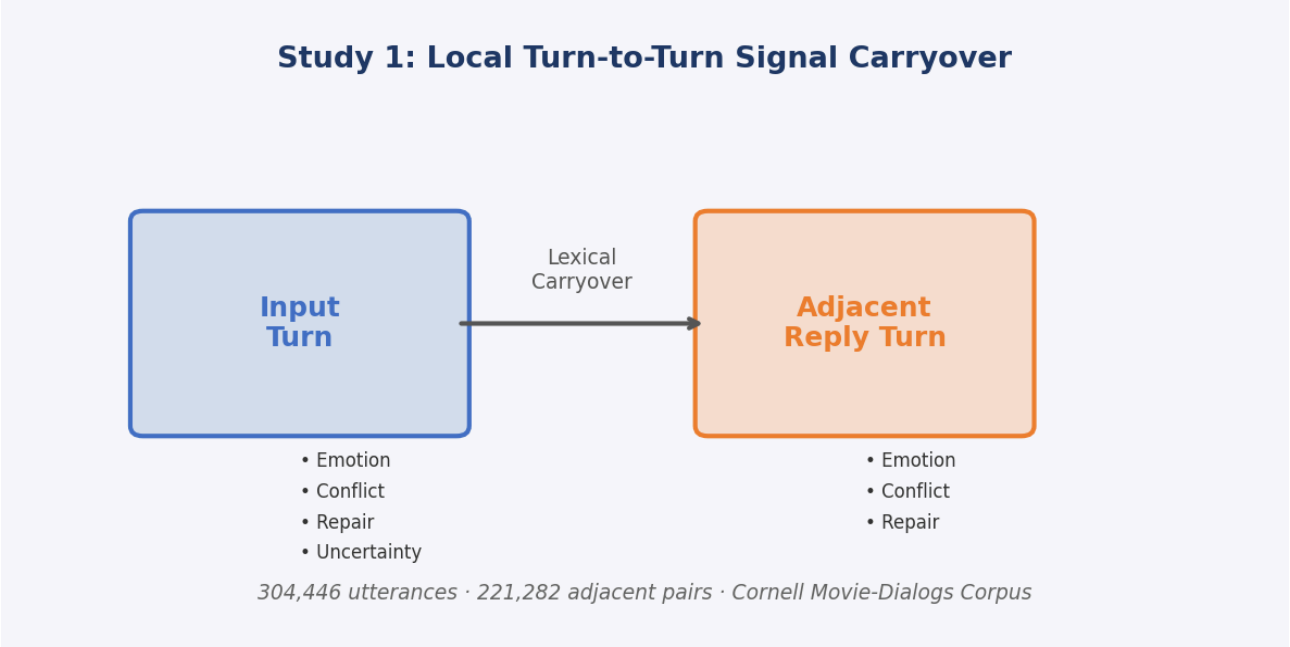

Figure 1: Our Study 1 design: lexical features extracted from input turns predicting carryover into adjacent reply turns (Cornell Movie-Dialogs Corpus, 221,282 adjacent pairs).

3.1 Research Question and Rationale

In Study 1, we ask whether simple lexical indicators recover local structure across adjacent turns in dialogue. We begin here for pragmatic reasons. If even coarse features such as emotion-related, conflict-related, repair-related, and uncertainty-related lexical counts fail to detect any turn-level carryover, then our broader empirical program on relationship decision support would lack a basic operational foundation.

3.2 Data and Features

The Cornell Movie-Dialogs Corpus provides a large collection of fictional conversational exchanges suitable for turn-level dialogue analysis (Danescu-Niculescu-Mizil & Lee, 2011). Our parsed corpus contained 304,446 individual utterances, 83,097 conversations, and 221,282 adjacent dialogue pairs. Each pair consisted of one input utterance and the immediately following reply.

We applied a lexicon-based pipeline to both input and reply turns (Tausczik & Pennebaker, 2010). The extracted variables included emotion-related lexical score, conflict-related lexical score, repair/conciliation-related lexical score, uncertainty-related lexical score, question counts, and word counts. We treated these variables as coarse surface markers rather than deep psychological measures.

3.3 Analytic Strategy

We proceeded in three stages. First, we used descriptive correlations to summarize relationships among input- and reply-side features. Second, we estimated ordinary least squares models predicting reply emotion-related, conflict-related, and repair-related lexical scores from input-side features. Third, we captured binary reply indicators for whether the next utterance contained emotion-related, conflict-related, or repair-related language at all, and estimated logistic regressions using the same predictor set.

3.4 Results

Our main finding was weak but interpretable local carryover. Reply emotion-related lexical score was moderately associated with reply repair-related lexical score (r = .324) and more weakly associated with reply conflict-related lexical score (r = .158). Input emotion-related lexical score showed a small positive association with reply emotion-related lexical score (r = .086), while input conflict-related lexical score showed a small positive association with reply conflict-related lexical score (r = .060).

In our OLS models, input emotion-related lexical score was the strongest predictor of reply emotion-related lexical score (b = 0.089, p < .001; R^2 = .008). Input conflict-related lexical score was the strongest predictor of reply conflict-related lexical score (b = 0.062, p < .001; R^2 = .004). For reply repair-related lexical score, input emotion-related lexical score was the strongest predictor (b = 0.018, p < .001) with a smaller positive contribution from input conflict-related lexical score (b = 0.005, p = .016), but overall fit remained minimal (R^2 = .001).

Our logistic models sharpened the same conclusion. Input emotion-related lexical score strongly predicted whether the reply contained emotion-related language (coef = 0.867, p < .001; pseudo-R^2 = .012). Input conflict-related lexical score strongly predicted whether the reply contained conflict-related language (coef = 0.717, p < .001; pseudo-R^2 = .0068). Input emotion-related lexical score also predicted the presence of repair-related language in the reply (coef = 0.338, p < .001; pseudo-R^2 = .0034).

3.5 Interpretation

Study 1 provides us with a minimal proof of operational viability. We find that adjacent dialogue turns contain weak but recoverable local carryover, and that simple lexical features can detect a small portion of that structure. At the same time, the very low R^2 and pseudo-R^2 values show that we are capturing only a surface layer. It is worth noting that repair, as an interactional phenomenon, involves sequential contingency that word-list features cannot fully capture (Robinson, 2006). Our study does not approximate deep dialogue understanding, and it does not test recurrence-aware support directly. Its role in our paper is foundational: it establishes the lowest-level unit in the empirical bridge we are building.

4. STUDY 2 - EMOTIONAL AND INTERACTIONAL STRUCTURE IN RELATIONSHIP-HELP NARRATIVES

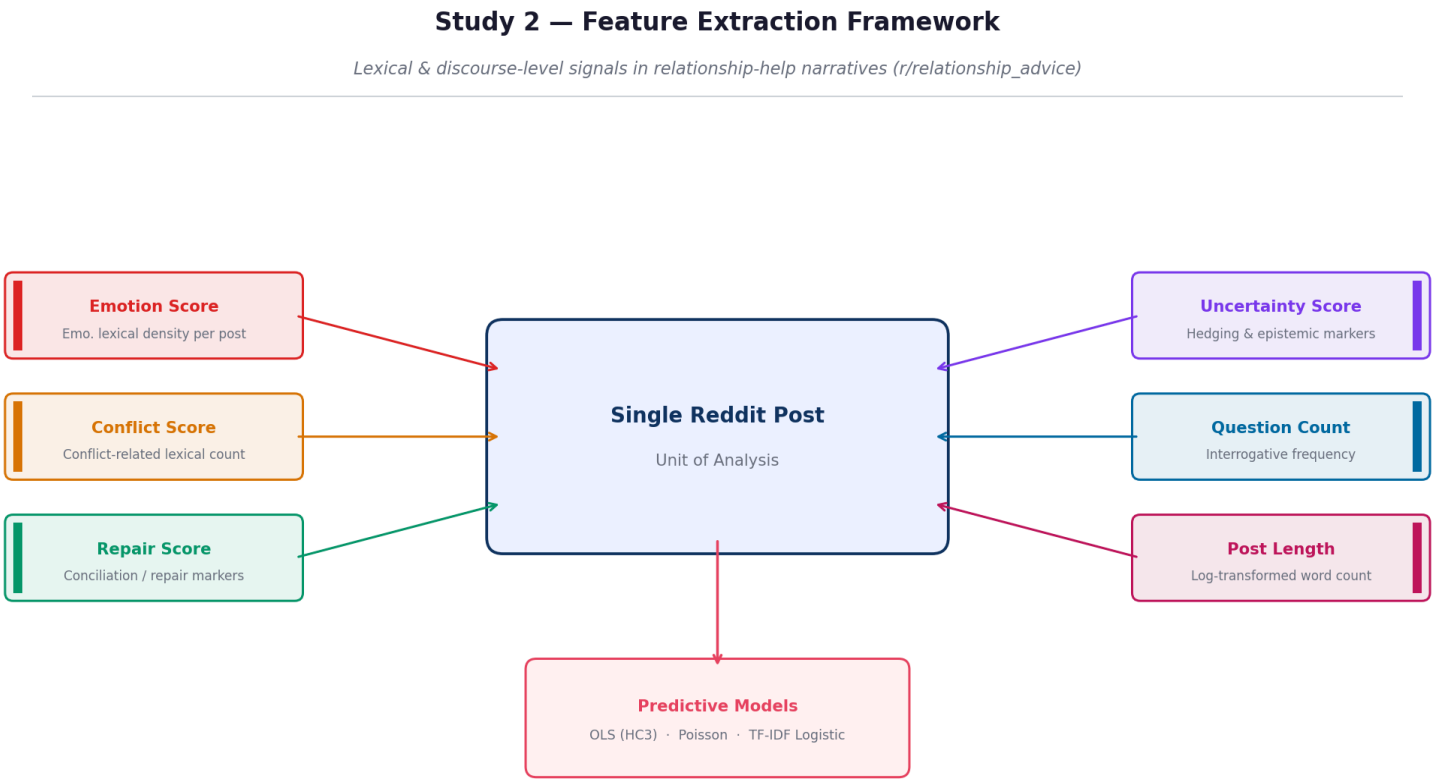

Figure 2: Our Study 2 framework: six lexical and discourse-level features extracted from r/relationship_advice posts, fed into OLS, Poisson, and TF-IDF logistic models to recover emotional and interactional structure.

4.1 Research Question and Rationale

In Study 2, we ask: to what extent do simple lexical and discourse-level features capture local emotional and interactional signals in online relationship-help narratives? Our move from Study 1 to Study 2 is deliberate. Anonymous online platforms now provide large-scale naturalistic data on relationship help-seeking behavior (Entwistle et al., 2021), and the broader literature on relationship help-seeking (Stewart et al., 2016) indicates that informal digital channels play an increasingly significant role alongside formal support. We shifted into relationship-relevant text in Study 2, which is much closer to the target domain implied by our conceptual paper.

4.2 Data, Hypotheses, and Measures

Our dataset consisted of posts from Reddit's r/relationship_advice, with the individual post used as the unit of analysis. For each post, we extracted simple lexical features related to emotion, conflict, repair, uncertainty, questioning behavior, and post length.

Our hypotheses were intentionally modest. We expected emotional intensity to be positively associated with conflict-related language, expected some positive relationship between uncertainty and questioning behavior, anticipated that conflict-related and repair-related language might co-occur in unstable ways across model families, and predicted that explanatory power would remain limited even when statistically reliable effects were present.

4.3 Analytic Strategy

Computerized lexical text analysis can provide psychologically informative, though necessarily coarse, indicators of language use (Tausczik & Pennebaker, 2010). Consistent with this logic, we used descriptive statistics and correlations, log-transformed OLS models with HC3 robust standard errors, Poisson count models, and a TF-IDF logistic classifier as a predictive robustness check for conflict-related language. Our goal was not to identify one final model, but to evaluate whether basic patterns remained visible across different lightweight specifications.

4.4 Results

Across model families, we found that emotional lexical content emerged as the most robust and consistent signal. It showed the clearest positive association with conflict-related language and also positively predicted repair-related language in several specifications. By contrast, effects involving uncertainty-related language and question-related language were less stable in our analyses. The relationship between conflict-related and repair-related language was also model-sensitive rather than uniformly strong.

As a predictive robustness check, our TF-IDF classifier identifying whether a post contained any conflict-related lexical content achieved high precision but relatively low recall for the positive class. That pattern suggests that richer text representations can capture some conflict signal, but still do so imperfectly under a lightweight pipeline.

4.5 Interpretation

Study 2 advances our empirical bridge by showing that relationship-help narratives contain detectable surface-level emotional and interactional structure. Prior work documents that conflict communication in romantic relationships is consequential but context-sensitive (Overall & McNulty, 2017; Domingue & Mollen, 2009); our findings suggest that even simple text features can capture partial traces of this structure in naturalistic self-report data. The main takeaway is not that simple text features adequately model relational reasoning. They do not. The stronger point is narrower: once we move the analysis into relationship-domain text, coarse computational features still recover modest regularities, especially around emotional intensity, conflict, and repair-related language. This positions Study 2 as a domain-specific extension of the turn-level operational logic we introduced in Study 1.

5. STUDY 3 - UPDATE-BASED RELATIONSHIP NARRATIVES AS A COARSE EPISODE-LEVEL TRAJECTORY PROXY

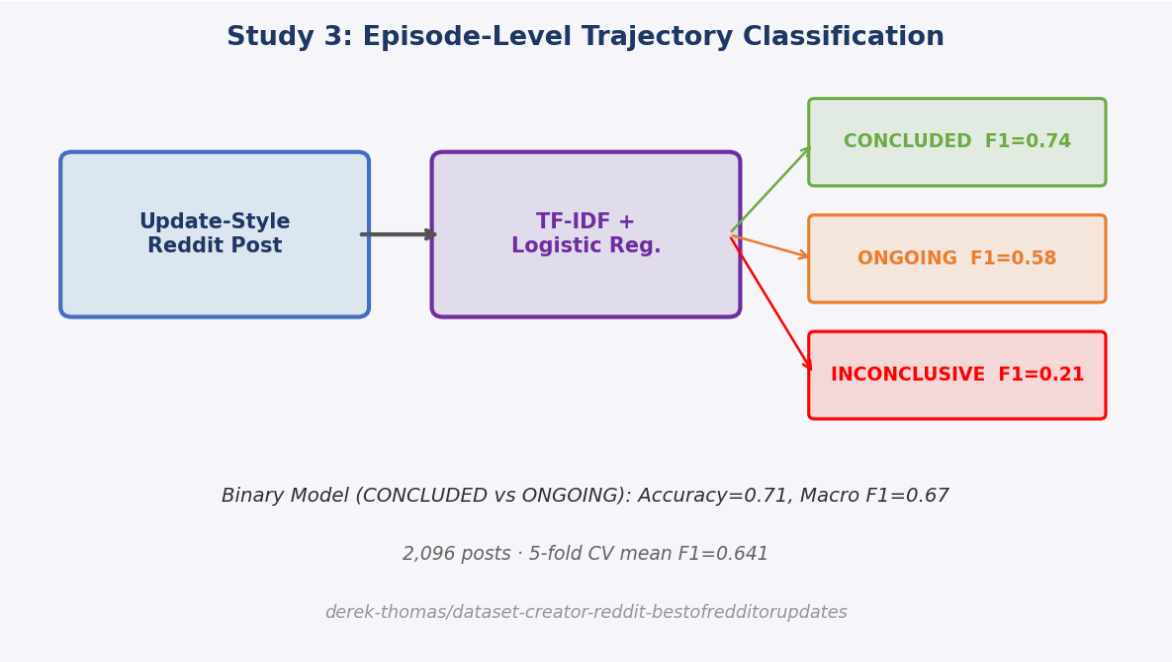

Figure 3: Our Study 3 classification pipeline: TF-IDF + logistic regression applied to update-style Reddit posts to distinguish CONCLUDED, ONGOING, and INCONCLUSIVE episode labels (2,096 posts; binary model accuracy = 0.71, macro F1 = 0.67).

5.1 Research Question and Rationale

In Study 3, we ask whether simple textual features can distinguish different reported relationship-update outcomes in romantic relationship narratives. Relative to our first two studies, this is our clearest move toward episode-level structure. Instead of analyzing adjacent turns or single relationship-help posts, we analyze update-style narratives that more directly reflect rough relational trajectory.

5.2 Dataset and Sample Construction

We used the dataset-creator-reddit-bestofredditorupdates corpus (Thomas, n.d.), a curated collection of Reddit update-style posts from Best of Redditor Updates hosted on Hugging Face. The full dataset contained 11,595 rows and 10 fields. We applied a multi-stage filtering pipeline to identify romantic relationship narratives. We removed unusable rows, used relationship-related title keywords to build a stricter relationship subset, retained update-oriented rows through title markers, flair information, and the updated field, and then applied a conversational-content filter. Our final candidate subset contained 2,446 posts, and our final modeling subset contained 2,096 posts.

5.3 Outcome Labels and Model

Our primary multiclass outcome was derived from flair labels: CONCLUDED, ONGOING, and INCONCLUSIVE. Our final three-class subset contained 1,290 concluded cases, 618 ongoing cases, and 188 inconclusive cases. Because INCONCLUSIVE was both small and semantically ambiguous, we conducted a secondary binary analysis retaining only CONCLUDED and ONGOING.

We used a simple baseline text-classification pipeline. We concatenated title and content into a single text field. TF-IDF remains a standard baseline representation for text retrieval and classification tasks (Salton & Buckley, 1988). A TF-IDF vectorizer with up to 20,000 unigram and bigram features, English stopword removal, and a minimum document frequency of 3 fed into a logistic regression classifier with max_iter = 2000 and class_weight = balanced. The full pipeline was implemented in scikit-learn (Pedregosa et al., 2011). We compared performance against a dummy most-frequent baseline and evaluated with precision, recall, F1-score, accuracy, confusion matrices, macro F1, and 5-fold cross-validation for the binary classifier.

5.4 Results

Our dummy classifier, which always predicted the majority class, achieved accuracy = 0.61, weighted F1 = 0.47, and macro F1 = 0.25. Our three-class TF-IDF plus logistic regression model improved performance to accuracy = 0.64, weighted F1 = 0.64, and macro F1 = 0.51. Class-specific performance was strongest for CONCLUDED (precision = 0.77, recall = 0.70, F1

= 0.74), moderate for ONGOING (precision = 0.52, recall = 0.65, F1 = 0.58), and weak for INCONCLUSIVE (precision = 0.23, recall = 0.18, F1 = 0.21).

Our binary model produced the clearest result. When we simplified the task to distinguishing concluded from ongoing narratives, we achieved accuracy = 0.71, weighted F1 = 0.71, and macro F1 = 0.67. Five-fold cross-validation yielded macro-F1 scores of 0.639, 0.631, 0.659, 0.682, and 0.596, for a mean of 0.641 and a standard deviation of 0.029.

Feature inspection suggested that our model learned a mixture of meaningful narrative cues and dataset-specific artifacts. Terms such as relationship and final were strongly weighted for CONCLUDED, while divorce was weighted toward ONGOING. At the same time, names, years, repost markers, and formatting artifacts also contributed to prediction, which limits interpretability.

5.5 Interpretation

Study 3 is our strongest empirical step toward the conceptual paper because we begin to treat relationship text as episode-level narrative evidence rather than isolated lexical content. Even so, our model predicts coarse dataset labels, not validated interpersonal outcomes. Our contribution is therefore bounded. We show that update-based relationship narratives contain recoverable trajectory-related textual structure, but only at a rough and noisy level. It functions as a partial bridge toward recurrence-aware modeling, not as a test of repeated relational adaptation itself.

6. GENERAL DISCUSSION

Taken together, our three studies support a cautious but coherent empirical claim. Relationship-relevant text contains recoverable structure at multiple levels of analysis. At the smallest scale, we find that adjacent turns show weak carryover in emotion-related, conflict-related, and repair-related language. At the narrative scale, we show that relationship-help posts contain detectable emotional and interactional organization, especially around emotional intensity and conflict-related language. At the broadest scale we examined, our update-based narratives contain enough regularity for simple models to distinguish coarse trajectory labels better than a majority-class baseline.

The more important point is how our results should and should not be interpreted. They do not validate recurrence-aware conversation decision support. They do not demonstrate branch preparedness, strategic alignment, post-conversation learning, or repeated-episode adaptation. What we do show is that the text environments relevant to such a framework are not structureless. Even under shallow lexical and n-gram methods, there is enough signal to justify more theory-aligned empirical work.

This reframing also clarifies the role of our paper within the larger research program. The conceptual paper defined the target constructs. Our empirical paper builds a measurement ladder beneath them. Study 1 establishes local response structure. Study 2 establishes relationship-domain emotional and interactional structure. Study 3 introduces a rough trajectory proxy at the episode level. That sequence gives the conceptual framework a more credible empirical spine, even though the core constructs of recurrence-aware support remain for future work.

7. CROSS-STUDY LIMITATIONS

Our paper has several limitations that apply across all three studies. First, none of our datasets constitutes a clean, validated corpus of real romantic conversations tracked longitudinally over repeated episodes. The dialogue corpus in Study 1 is fictional, the Reddit posts in Study 2 are one-sided self-reports, and the update corpus in Study 3 is a curated and artifact-prone Reddit-derived dataset.

Second, our operationalizations are intentionally simple. Lexicon-based features and TF-IDF representations are transparent and easy to deploy, but they are shallow. They do not model goals, intentions, speaker identities, pragmatic force, conversational memory, or objective-strategy mismatch. We therefore treat them as useful baseline operational tools, not as deep relational models.

Third, some of our findings are statistically reliable but substantively weak. This is especially clear in Study 1, where model fit is extremely low. Statistical significance in large samples should not be confused with strong explanatory power.

Fourth, label quality is limited. Study 3 in particular depends on rough flair-based narrative status labels, including an INCONCLUSIVE category that appears semantically unstable. That constraint makes our classification results informative but necessarily provisional.

Finally, our paper remains indirectly rather than directly tied to the central conceptual constructs. Recurrence, branching, strategic objective clarification, adaptive learning, and post-conversation outcome integration are still not operationalized in a way that permits direct empirical testing.

8. FUTURE EMPIRICAL AGENDA

Our next phase should move from surface structure to theory-aligned structure. A stronger empirical continuation would require datasets that track repeated relationship episodes over time, labels that directly encode goals and conversational outcomes, and models that represent branching, adaptation, and mismatch between intended objective and chosen conversational strategy.

One practical implication is that future datasets should be built around episodes rather than isolated texts. An episode-level design would ideally include the precipitating situation, the user's goal, the message or conversation attempt, the partner's response type, the immediate outcome, and follow-up information about whether the same issue recurs. That kind of structure would allow recurrence-aware decision support to be tested directly instead of inferred indirectly from lexical regularities.

A second implication is methodological. We recommend that simple lexicon and TF-IDF baselines remain in the pipeline for transparency and sanity checking, but that they be supplemented by richer semantic representations and theory-linked annotations. The purpose of our paper is not to freeze the research program at a baseline stage; it is to justify moving beyond that stage with a clearer empirical foundation.

9. CONCLUSION

Our paper reorganizes three small quantitative studies into a single empirical bridge for a prior conceptual framework on recurrence-aware conversation decision support. We connect the studies through a clear escalation in unit of analysis: adjacent dialogue turns, single relationship-help narratives, and update-based romantic relationship episodes. Across all three levels, our simple computational methods recover some interpretable signal, but that signal remains limited and often shallow.

Our paper's contribution is therefore best understood as foundational. We do not confirm the full conceptual theory. Instead, we demonstrate that multiple layers of relationship-relevant text contain enough structure to justify more direct empirical work on recurrence, branching, strategic alignment, and post-conversation adaptation. In that sense, our three pilot studies do not replace the conceptual paper. They provide its first empirical spine.

REFERENCES

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 Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011.

Domingue, R., & Mollen, D. (2009). Attachment and conflict communication in adult romantic relationships. Journal of Social and Personal Relationships, 26(5), 678–697.

Entwistle, C., Horn, A. B., Meier, T., & Boyd, R. L. (2021). Dirty laundry: The nature and substance of seeking relationship help from strangers online. Journal of Social and Personal Relationships.

Overall, N. C., & McNulty, J. K. (2017). What type of communication during conflict is beneficial for intimate relationships? Current Opinion in Psychology, 13, 1–5.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Robinson, J. D. (2006). Managing trouble responsibility and relationships during conversational repair. Communication Monographs, 73(2), 137–161.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

Stewart, J. W., Bradford, K., Higginbotham, B. J., & Skogrand, L. (2016). Relationship help-seeking: A review of the efficacy and reach. Marriage & Family Review.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29(1), 24–54.

Thomas,D.(n.d.).dataset-creator-reddit-bestofredditorupdates[Dataset].HuggingFace. https://huggingface.co/datasets/derek-thomas/dataset-creator-reddit-bestofredditorupdates