MIMIC: Multi-party Dialogue Augmentation via Speaker Stylistic Transfer

Gaetano Cimino[δ], **Giuseppe Carenini**[δ], **Vincenzo Deufemia**[γ]
[δ]Department of Computer Science, University of British Columbia
V6T 1Z4, Vancouver, BC, Canada
[γ]Department of Computer Science, University of Salerno
84084, Fisciano, Salerno, Italy
gaetano.cimino@ubc.ca, carenini@cs.ubc.ca, deufemia@unisa.it

Abstract

Annotated data scarcity has long hindered progress in dialogue discourse parsing. To fill this gap, we introduce MIMIC, a framework for augmenting discourse-annotated corpora via speaker stylistic transfer using Large Language Models (LLMs). MIMIC rephrases utterances while preserving discourse coherence, using the MASK metric to identify speakers for replacement that enrich structural diversity and the MIRROR method to select substitute speakers who have experienced similar discourse interactions. Experimental results on STAC and Molweni corpora show that parsers trained with MIMIC-augmented data improve both link prediction and relation classification, with consistent gains for underrepresented discourse patterns and in low-resource scenarios.[1]

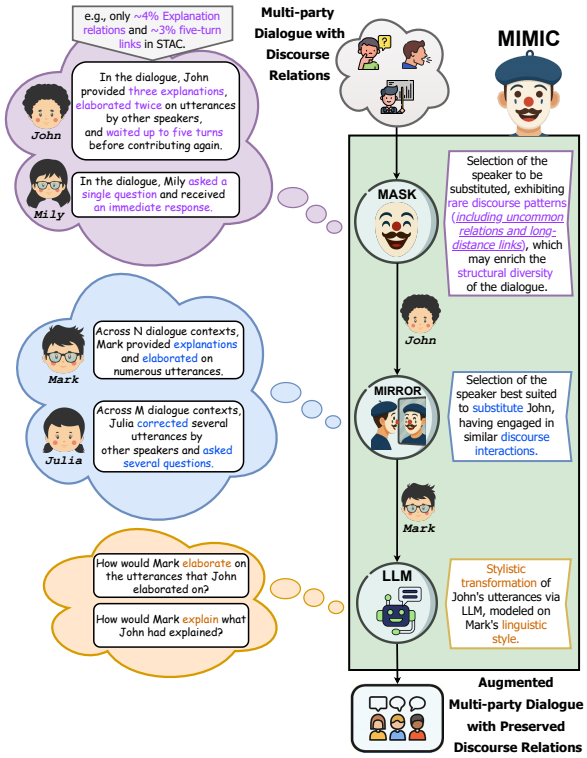


Figure 1: Illustration of MIMIC. Dialogues are generated by identifying structurally rare speakers using MASK, and rephrasing their utterances in the style of speakers engaged in similar discourse interactions selected via MIRROR.

1 Introduction

Discourse parsing in multi-party dialogues underpins tasks like action prediction (Chaturvedi et al., 2024), summarization (Feng et al., 2022), and dialogue comprehension (Ma et al., 2023), but progress is slowed by scarce annotated data. Few discourse-annotated corpora exist, most notably STAC (Asher et al., 2016) and Molweni (Li et al., 2020), but their small size and manual annotation hinder scalability (Zhang et al., 2018).

In this paper, we tackle data scarcity for dialogue discourse parsing through dialogue augmentation, a largely unexplored area, particularly for multi-party and discourse-level tasks (Mahajan and Shaikh, 2021). An augmented dialogue is valid for parser training only if it preserves the structural annotations of the dialogue it originates from. The key challenge, therefore, is to generate dialogues with consistent discourse structures. To this end, we propose MIMIC (**M**ulti-party d**I**alogue aug**M**entation via speaker styl**I**sti**C** transfer), a framework leveraging LLMs to perform controlled rephrasing of speaker utterances within annotated dialogues drawn from existing discourse parsing corpora. As shown in Figure 1, it comprises three steps: (i) selecting a speaker s_u to be substituted, (ii) selecting a speaker s_x to replace s_u, and (iii) rephrasing s_u's utterances in s_x's style. In the first step, MIMIC identifies the speaker whose utterances best support augmentation. Following data-balancing principles (Henning et al., 2023), which advocate augmenting rare patterns within a dataset, we introduce MASK (**M**easure

[1]Our code is available at MIMIC.

Findings of the Association for Computational Linguistics: EACL 2026, pages 2693–2719
March 24-29, 2026

of **A**typical **S**tructural **K**nowledge), a metric ranking speakers by the rarity of their discourse patterns. Prior work notes two core challenges in dialogue parsing (Chi and Rudnicky, 2022; Li et al., 2023; Cimino et al., 2024): (i) predicting long-distance links due to the predominance of direct links in corpora, and (ii) classifying infrequent relation types. Since relation and link distributions vary across datasets, MASK adapts its rankings accordingly. For example, in Figure 1, applied to the STAC corpus, MASK selects John rather than Mily because John participates in rare *Explanation* relations (≈4% in STAC) and links of distance five (≈3% in STAC), while Mily's utterances fall into common *Question-Answer* relations (≈24% in STAC) and direct links (≈56% in STAC).

Once a suitable speaker has been selected, in MIMIC second step such speaker's utterances are rephrased while preserving the original discourse relations. Existing dialogue augmentation methods rely on direct LLM rephrasing (Mehri et al., 2022; Chen et al., 2024), but LLMs struggle with rhetorical relations (Fan et al., 2024), potentially yielding paraphrases that distort original structures. To mitigate this, we propose MIRROR (**M**odeling **I**nterlocutor **R**oles and **R**elational **O**ve**R**lap), which identifies speakers with similar discourse patterns to serve as stylistic substitutes. Their dialogue histories guide the LLM to preserve relations, e.g., using Mark instead of Julia when rephrasing John's explanatory and elaborative utterances in Figure 1.

Finally, inspired by personalized dialogue systems that adapt to styles from dialogue histories rather than fixed personas (Ma et al., 2021; Qian et al., 2021; Lu et al., 2023; Cheng et al., 2024), we instruct the LLM to rephrase utterances based on the selected substitute speaker's history. This yields utterances that are both contextually grounded and stylistically aligned, enriching training data diversity while maintaining discourse structure[2].

To validate MIMIC, we conduct experiments on the STAC and Molweni corpora, evaluating two discourse parsers (Chi and Rudnicky, 2022; Fan et al., 2022) trained on the synthetically generated data. The results indicate that MIMIC enhances parser performance, yielding improvements of up to 3.7% in structure prediction and up to 8.2% in full parsing relative to parsers trained on the original corpora. Furthermore, MIMIC surpasses both speaker- and discourse-aware generation methods.

[2]Appendix J shows examples of augmented utterances.

2 Related Work

Dialogue and Parsing Augmentation via LLMs Latest dialogue augmentation advances use LLMs to generate synthetic data, boosting performance by synthesizing corpora from seed dialogues (Chen et al., 2022; Zheng et al., 2023). Systems like DIALOGIC (Li et al., 2022) and LAD (Mehri et al., 2022) target task-oriented exchanges, while SDA (Liu et al., 2024) and CONVAUG (Chen et al., 2024) aim to diversify open-domain and information-seeking dialogues. These methods generate fluent but dyadic conversations, struggling with multi-party dialogue, where overlapping turns, distributed discourse, and role diversity hinder discourse-level tasks such as parsing, which require structural coherence.

Recent work has applied LLMs to parsing. For example, Zhang et al. (2025) showed that LLMs can synthesize phrase structures for constituency parsing by manipulating existing subtrees, yielding diverse examples that improve cross-domain performance. In contrast, MIMIC targets parsing in multi-party dialogues, augmenting data for discourse-level rather than sentence-level analysis.

Persona-based Dialogue Generation Early approaches used explicit persona profiles to enforce consistent traits, e.g., Zhang et al. (2018) integrated structured persona data for identity-consistent responses. Later work embedded persona attributes directly into LLMs to enhance stylistic and behavioral fidelity (Shea and Yu, 2023; Jandaghi et al., 2024). However, explicit profiles are hard to maintain, as they require detailed user data (Zhao et al., 2024), and being static, they face cold-start issues and miss evolving behaviors (Gu et al., 2021). Recent studies instead leverage speaker dialogue histories to dynamically model user style (Ma et al., 2021; Zhong et al., 2022; Lu et al., 2023; Cheng et al., 2024). Building on this, MIMIC uses dialogue history to synthesize multi-party dialogues, rewriting utterances in alternative speakers' styles while preserving discourse coherence and structure.

Recent work generated dialogues via personas created by LLMs (Kirstein et al., 2025), asking the model to define aspects such as tone and style for each speaker. However, when augmenting dialogues for discourse parsing, generic personas hinder preserving speaker relations. Using dialogue histories instead lets the model draw on concrete examples of speaker behavior, ensuring rephrasings are contextually coherent and relation-preserving.

3 Dialogue Infilling for Discourse Parsing

Discourse Parsing Discourse parsing uncovers dialogue structure by identifying links between minimal textual segments aka *Elementary Discourse Units* (EDUs). Let $D = (\varepsilon_1, \ldots, \varepsilon_n)$ be a dialogue of n EDUs, each attributed to a speaker $s_u \in \mathcal{S}_D$, where $\mathcal{S}_D$ is the set of participants in D. The goal is to represent D as a graph $\mathcal{G} = (V, E, \varphi)$, where:

- $V = \{\varepsilon_1, \ldots, \varepsilon_n\}$ is the set of EDUs.
- $E \subseteq V \times V$ contains links $(\varepsilon_k, \varepsilon_i)$ with $1 \leq k < i$, in accordance with SDRT's (Segmented Discourse Representation Theory) *Turn Constraint* (Afantenos et al., 2015), which disallows *backward links*. The speaker of ε_k is referred to as the *initiator*, while the speaker of ε_i is called the *responder*. Each EDU must have at least one incoming link to ensure rhetorical connectedness.
- $\varphi : E \rightarrow \mathcal{R}$ maps each edge to a rhetorical relation in $\mathcal{R}$.

Discourse parsing involves two tasks. *Discourse structure prediction* identifies the set of unlabeled links E, capturing the "*naked*" graph structure of the dialogue (Cimino et al., 2024). *Full discourse parsing* extends this by assigning each link $(\varepsilon_k, \varepsilon_i)$ a rhetorical relation $r_{ki} \in \mathcal{R}$ (Fan et al., 2022), forming triples $\tau_p = (\varepsilon_k, \varepsilon_i, r_{ki})$. The final output is the full set of labeled triples $\mathcal{T} = \bigcup_{p=1}^{|E|} \tau_p$.

Dialogue Augmentation as Dialogue Infilling To enrich discourse parsing corpora, we frame augmentation as a *Dialogue Infilling* task (Lee and Berg-Kirkpatrick, 2022).

Let $\mathcal{C}$ be a training corpus involving speakers $\mathcal{S}_\mathcal{C}$, composed of samples $\{(\mathcal{G}_j, \mathcal{S}_{\mathcal{G}_j})\}_{j=1}^{N}$, where $\mathcal{G}_j = (V_{\mathcal{G}_j}, \mathcal{T}_{\mathcal{G}_j})$ is a discourse graph and $\mathcal{S}_{\mathcal{G}_j} \subseteq \mathcal{S}_\mathcal{C}$ the subset of active speakers. We simulate alternative discourse contributions by replacing the utterances of a speaker $s_u \in \mathcal{S}_{\mathcal{G}_j}$ with style-consistent paraphrases from another speaker $s_x \in \mathcal{S}_\mathcal{C} \setminus \mathcal{S}_{\mathcal{G}_j}$, who appears in other dialogues in $\mathcal{C}$.

We replace all EDUs of s_u, denoted $V_{\mathcal{G}_j}^{s_u} = \{\varepsilon_{u,1}, \ldots, \varepsilon_{u,m}\}$, with paraphrases $V_{\mathcal{G}_j}^{s_x} = \{\varepsilon'_{u,1}, \ldots, \varepsilon'_{u,m}\}$ in s_x's linguistic style. This extends the original corpus with a new dialogue, corresponding to a new graph $\mathcal{G}'_j = (V'_{\mathcal{G}_j}, \mathcal{T}'_{\mathcal{G}_j})$, with $V'_{\mathcal{G}_j} = (V_{\mathcal{G}_j} \setminus V_{\mathcal{G}_j}^{s_u}) \cup V_{\mathcal{G}_j}^{s_x}$ and updated triples $\mathcal{T}'_{\mathcal{G}_j}$.

Next, we present MIMIC, our dialogue infilling method for generating structurally consistent dialogue variants.

4 MIMIC

The MIMIC pipeline is shown in Figure 2. For an overview, the process includes: (a) **Target Speaker Selection** chooses speaker s_u based on the rarity of discourse patterns in their EDUs; (b) **Substitute Speaker Selection** identifies speaker s_x best positioned to rephrase s_u, drawing on prior dialogues; (c) **Interpersonal Relationship Modeling** captures interpersonal links of participants in $S_\mathcal{G}$ who interacted with s_u; (d) **EDU Generation** prompts an LLM to paraphrase s_u's EDUs in s_x's style. The model is guided by discourse-aware examples to ensure stylistic adaptation; and (e) **EDU Refinement** adjusts generated EDUs to preserve rhetorical relations[3]. Executing the pipeline doubles the size of the original corpus, while iterative application allows for arbitrary expansion, enabling controlled large-scale augmentation[4].

Target Speaker Selection To augment an annotated multi-party dialogue via infilling, we first select a speaker for replacement, prioritizing those engaged in rare relations and long-distance links to enhance corpus diversity. This selection is guided by MASK, a ranking metric that measures speakers' structural significance.

Given a discourse graph $\mathcal{G}$ and a speaker s, let $\mathcal{R}_s^\mathcal{G} \subseteq \mathcal{R}$ be the set of discourse relations that occur within the set of discourse triples $\mathcal{T}_s^\mathcal{G}$ associated with s in $\mathcal{G}$. For each $r \in \mathcal{R}_s^\mathcal{G}$, let $f_s^\mathcal{G}(r)$ denote its frequency in $\mathcal{T}_s^\mathcal{G}$, and define

$$p_\mathcal{C}(r) = \frac{\sum_{\mathcal{G} \in \mathcal{C}} \sum_{(\varepsilon_i, \varepsilon_j, r_{ij}) \in \mathcal{T}_\mathcal{G}} \mathbb{I}_{r_{ij}=r}}{\sum_{\mathcal{G} \in \mathcal{C}} |\mathcal{T}_\mathcal{G}|}$$

as the empirical probability of r across the training corpus $\mathcal{C}$, with $\mathbb{I}(\cdot)$ denoting the indicator function. The *relation rarity score* for s in $\mathcal{G}$ is:

$$\psi_{\text{rel}}^\mathcal{G}(s) = \frac{1}{|\mathcal{T}_s^\mathcal{G}|} \sum_{r \in \mathcal{R}_s^\mathcal{G}} \frac{f_s^\mathcal{G}(r)}{p_\mathcal{C}(r)}$$

The *link rarity score* reflects the average span of s's links:

$$\psi_{\text{lin}}^\mathcal{G}(s) = \frac{1}{|\mathcal{T}_s^\mathcal{G}|} \sum_{(\varepsilon_i, \varepsilon_j, r_{ij}) \in \mathcal{T}_s^\mathcal{G}} |i - j|$$

We apply min-max scaling over speakers in $\mathcal{G}$ to both scores, obtaining $\widetilde{\psi}_{\text{rel}}^\mathcal{G}$ and $\widetilde{\psi}_{\text{lin}}^\mathcal{G}$. The final selection criterion is defined as:

[3] The prompt templates for EDU generation and refinement are presented in Appendix B.

[4] Appendix H discusses sequential applications of MIMIC.

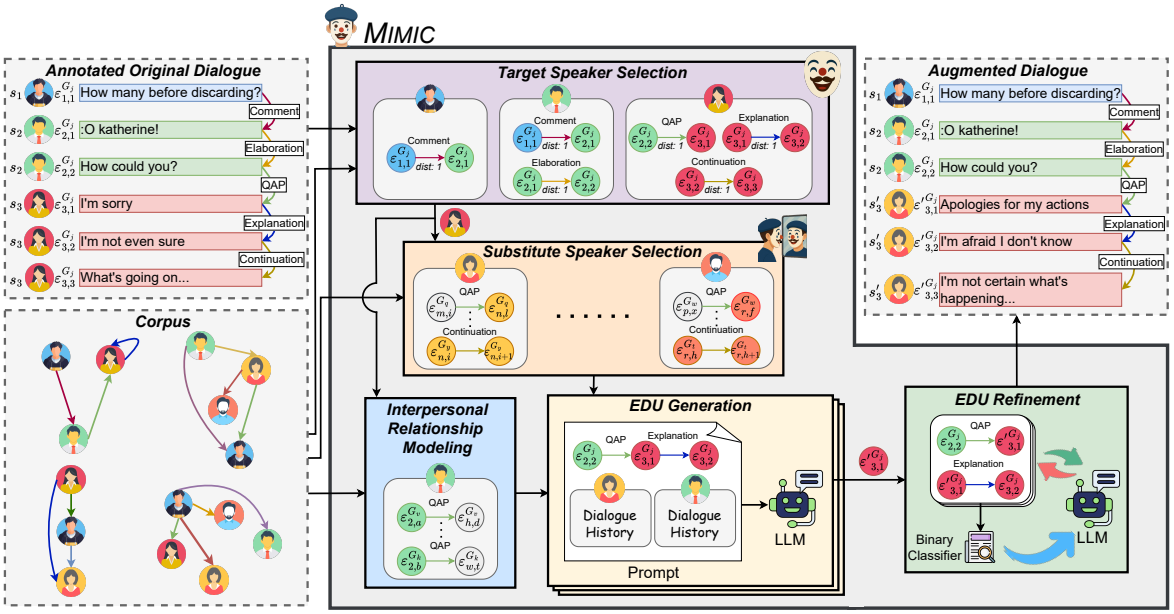


Figure 2: Overview of the MIMIC augmentation process on an annotated dialogue from STAC (id *pilot21_11*).

$$s_u = \arg\max_{s \in \mathcal{S}_\mathcal{G}} \left(\widetilde{\psi}^{\mathcal{G}}_{\text{rel}}(s) + \widetilde{\psi}^{\mathcal{G}}_{\text{lin}}(s) \right)$$

For example, in Figure 2, all speakers exhibit links of distance 1; consequently, the selection is based on the rarity of the relations. Since speaker s_3 participates in a particularly rare relation in STAC (see Table 5a), i.e., *Explanation*, it achieves the highest MASK score and is chosen for replacement. Notably, although framed for single-speaker substitution, MASK generalizes to multi-speaker replacement by selecting the top-K ranked candidates.

Substitute Speaker Selection Once a speaker $s_u \in \mathcal{S}_\mathcal{G}$ is selected for replacement in $\mathcal{G}$, MIRROR identifies a substitute s_x whose linguistic style is used to rephrase $V^{s_u}_\mathcal{G}$. We approximate s_x's style from past dialogue responses (Ma et al., 2021). This is feasible because, as Figure 2 shows, each speaker in corpus $\mathcal{C}$ participates in diverse conversations and relation types, providing sufficient stylistic context. Notably, substitutes are chosen not for linguistic style similarity with original speakers but for shared discourse functions, increasing the likelihood of preserving the intended rhetorical role. Formally, given the set of triples $\mathcal{T}^{\mathcal{G}}_{s_u}$, MIRROR finds a substitute speaker $s_x \in \mathcal{S}_\mathcal{C} \setminus \mathcal{S}_\mathcal{G}$ whose dialogue history contain relations most similar to those of s_u in $\mathcal{G}$, through the following two stages:

- *Exact Match Stage:* MIRROR first searches for a speaker who has, for each triple in $\mathcal{T}^{\mathcal{G}}_{s_u}$, a *representative*[5] dialogue history including triples with same relations and roles. If multiple speakers qualify, one is chosen at random to foster diversity. For example, to replace s_3 in Figure 2, MIRROR selects a speaker whose history includes triples exhibiting *Question-Answer* pairs (QAP) in the responder role, as well as triples corresponding to *Explanation* and *Continuation* pairs in both the initiator and responder roles.
- *Dominance-based Stage:* If no speaker meets the exact match condition, MIRROR applies a dominance-based strategy that prioritizes coverage of relations and roles. In this stage, candidates are ranked based on how well their dialogue histories reflect the discourse functions of the target speaker's utterances. For example, if no candidate exhibits the *QAP*, *Explanation*, and *Continuation* relations with their respective roles to replace s_3 in Figure 2, MIMIC selects a substitute that best approximates these patterns, ensuring structural consistency even without full discourse-role alignment. Specifically, each candidate s is encoded as a vector $\mathbf{v}_s \in \mathbb{N}^{2\times|R^{\mathcal{G}}_{s_u}|}$, where $v^{(r,\rho)}_s$ counts occurrences of relation r under role $\rho \in \{\text{INITIATOR}, \text{RESPONDER}\}$. A partial order over speakers is then defined via vector dominance: s_j *dominates* s_k if $v^{(r,\rho)}_{s_j} \geq v^{(r,\rho)}_{s_k}$ $\forall r \in R^{\mathcal{G}}_{s_u}, \forall \rho \in \{\text{INITIATOR}, \text{RESPONDER}\}$. The *dominance score* $\delta(s_j)$ is the number of candidate speakers dominated by s_j:

$$\delta(s_j) = \sum_{s_k \in \mathcal{S}_\mathcal{C} \setminus \mathcal{S}_\mathcal{G}} \mathbb{I}\left(\mathbf{v}_{s_j} \succeq \mathbf{v}_{s_k}\right)$$

[5] Representativeness is defined by a minimum number of triples, tuned on a validation set (see Appendix A).

where $\succeq$ denotes element-wise comparison. The substitute speaker is chosen as:

$$s_x = \arg \max_{s_j \in \mathcal{S_C} \setminus \mathcal{S_G}} \delta(s_j)$$

Interpersonal Relationship Modeling Interpersonal dynamics significantly shape multi-party dialogue (Lee et al., 2025), with initiator–responder relationships influencing response styles (Wei et al., 2023). These dynamics affect tone, language, and behavior (Floyd and Morman, 1997), making their modeling essential for generating socially and contextually appropriate responses.

To guide LLM rephrasing of EDUs for target speaker s_u, we enrich the prompt with dialogue history from each participant s_q in $\mathcal{G}$ who has interacted with s_u. Formally, for each triple $\tau = (\varepsilon_u, \varepsilon_q, r)$ or $(\varepsilon_q, \varepsilon_u, r) \in \mathcal{T}^{\mathcal{G}}_{s_u}$, we retrieve structurally similar triples $\tau = (\varepsilon_k, \varepsilon_q, r)$ or $(\varepsilon_q, \varepsilon_k, r)$ from other discourse graphs. For example, when replacing s_3 in Figure 2, prior triples with s_2 as a *QAP* initiator are included, grounding the prompt in social interaction patterns.

EDU Generation After establishing the knowledge for EDU rephrasing, generation proceeds sequentially. At each step, the LLM rephrases an EDU $\varepsilon^{\mathcal{G}}_{u,i}$ by speaker s_u using a constrained process, incorporating all triples from $\mathcal{T}_{\mathcal{G}}$ where $\varepsilon^{\mathcal{G}}_{u,i}$ acts as initiator or responder to preserve coherence and maintain alignment with the surrounding discourse.

Beyond coherence, rhetorical relations are critical for retaining the original communicative intent. Thus, each EDU is rephrased with explicit relation constraints. For example, rephrasing $\varepsilon^{\mathcal{G}_j}_{3,1}$ in Figure 2 requires maintaining its *QAP* link with $\varepsilon^{\mathcal{G}_j}_{2,2}$ and *Explanation* link with $\varepsilon^{\mathcal{G}_j}_{3,2}$.

An independent-based strategy is adopted for EDU rephrasing, wherein each EDU is generated exclusively with reference to the corresponding ground truths. This approach mitigates cascading errors, though potentially at the cost of inter-EDU coherence. For comparison, Appendix C reports the evaluation of an incremental strategy, in which the rephrasing of each EDU is autoregressively conditioned on previously generated outputs. Although this strategy should mirror human dialogue better, independent generation proves more robust for structure-sensitive tasks like discourse parsing.

EDU Refinement While LLMs can follow human instructions, hallucinations remain a key challenge in synthetic data generation (Ding et al., 2024). Our augmentation approach assumes that discourse structure, and thus rhetorical links, can be preserved. However, due to the fluidity of dialogue, an alternative response may remain coherent (Ma et al., 2021) while disrupting the original rhetorical relation, introducing noise[6]. Inspired by self-refinement techniques in LLMs (Madaan et al., 2023), we introduce a refinement step to address this issue. Specifically, for each relation type r, a binary classifier ϕ_r checks if candidate EDU pairs preserve intended relations. If any fail, an LLM is first prompted to produce a suggestion for the candidate EDU to preserve the relation. This suggestion is then added to the EDU generation prompt, and generation is repeated to enhance rhetorical fidelity.

Statistic	STAC	Molweni1k
#Dialogues	947	1000
#EDUs (min / max / avg)	2 / 105 / 11.04	7 / 14 / 8.78
Avg EDU length (in tokens)	4.73	13.74
#Speakers (min / max / avg)	1 / 6 / 2.99	2 / 8 / 3.54
% of dialogues with $\geq n$ speakers		
$\geq$ 2 speakers	96.3%	100.0%
$\geq$ 3 speakers	72.3%	77.2%
$\geq$ 4 speakers	29.5%	46.6%
$\geq$ 5 speakers	0.8%	21.5%
$\geq$ 6 speakers	0.1%	6.6%
$\geq$ 7 speakers	–	2.0%
$\geq$ 8 speakers	–	0.3%

Table 1: Statistics for STAC and Molweni1k datasets.

5 Experimental Setup

Datasets We apply MIMIC to two popular discourse parsing datasets[7]: STAC (Asher et al., 2016), comprising dialogues from the game Settlers of Catan, and Molweni (Li et al., 2020), based on the Ubuntu Chat Corpus (Lowe et al., 2015). Given their limited lexical overlap (Liu and Chen, 2021), these corpora are ideal for cross-domain testing.

For STAC, we use the split from Shi and Huang (2019): 947 train, 103 validation, and 109 test dialogues. We augmented 943 of the training dialogues, excluding 4 that contained missing links, yielding a total of 1,890 dialogues. To emulate a comparable low-resource setting, we randomly selected and augmented 1,000 training dialogues from Molweni, hereafter referred to as Molweni1k, resulting in an augmented dataset consisting of 2,000 dialogues. We use the original validation and test splits of the Molweni dataset, each comprising 500 dialogues. Both datasets follow SDRT and have the same relation types ($|\mathcal{R}| = 16$).

Considered Models For dialogue infilling, we

[6] Appendix I provides an illustrative example.

[7] Detailed corpus statistics are presented in Table 1.

Parser	Approach	STAC		Molweni	
		UAS	*LAS*	*UAS*	*LAS*
SDDP	Naïve	73.41 ± 0.5	57.12 ± 0.6	79.11 ± 0.2	54.04 ± 0.3
	Context-free Rephrasing	72.95 ± 0.8	57.28 ± 0.5	79.21 ± 0.3	54.88 ± 0.2
	Persona-based Rephrasing	73.02 ± 0.3	57.33 ± 0.4	79.38 ± 0.2	55.01 ± 0.2
	MIMIC (Ours)	**74.41** ± 0.3	**58.10** ± 0.3	**80.60** ± 0.4	**55.75** ± 0.3
DAMT	Naïve	70.57 ± 0.5	49.66 ± 0.4	76.75 ± 0.4	54.15 ± 0.2
	Context-free Rephrasing	71.06 ± 0.2	51.28 ± 0.5	77.08 ± 0.5	54.21 ± 0.5
	Persona-based Rephrasing	71.12 ± 0.3	51.30 ± 0.4	77.47 ± 0.5	54.36 ± 0.5
	MIMIC (Ours)	**71.72** ± 0.4	**52.34** ± 0.3	**79.12** ± 0.4	**55.29** ± 0.4

Table 2: Performance of SDDP and DAMT on STAC and Molweni test sets across four training setups. Results are average UAS/LAS over five runs. Statistical testing using t-tests with Holm-Bonferroni correction (Holm, 1979) confirmed MIMIC significantly outperforms the others ($p < 0.05$).

use the 8-bit quantized Llama 3.3 70B model (Grattafiori et al., 2024) with temperature set to 0 for reproducibility. EDU refinement uses binary BERT classifiers to predict relations between EDU pairs split by [SEP], all achieving $\geq 71\%$ accuracy and $\geq 70\%$ precision on test sets[8].

Discourse Parsers and Baselines We chose SOTA models based on: (a) publicly available code, (b) strong task performance, and (c) no reliance on proprietary LLMs for reproducibility. We evaluate two models[9]: (i) SDDP (Chi and Rudnicky, 2022), using matrix-tree learning and a modified CLE algorithm (Edmonds et al., 1967) for non-projective parsing, and (ii) DAMT (Fan et al., 2022), a distance-aware, multi-task model combining transition- and graph-based methods.

As existing methods generate dialogues without considering discourse relations, we adapt prior research to discourse parsing to implement three baselines[10] for benchmarking MIMIC. The first, *speaker-level context-free rephrasing*, follows earlier work (Mehri et al., 2022; Chen et al., 2024) by paraphrasing utterances with Llama while omitting speaker cues, isolating fine-grained style transfer from generic effects of diverse paraphrases. The second, *speaker-level persona-based rephrasing*, extends the work of Kirstein et al. (2025) by rephrasing utterances in alignment with LLM-generated personas, using Llama for both persona construction and rephrasing. The third, *dialogue-level discourse-aware generation*, creates new dialogues from contextual information (Kim et al., 2023). Specifically, it leverages discourse relations derived from annotated graphs in existing corpora as structural backbones and instructs Llama to generate dialogues that maintain these relations.

Evaluation Metrics We evaluate parsers with two metrics: (i) *Unlabeled Attachment Score* (UAS) for structure prediction, and (ii) *Labeled Attachment Score* (LAS) for full parsing. Following Ma et al. (2021), we complement automatic metrics with a human evaluation detailed in Appendix F, showing that MIMIC rephrasings preserve rhetorical relations, maintain coherence, and reflect personalization through the use of dialogue histories.

6 Experimental Results

We evaluate SDDP and DAMT parsers trained on original and augmented versions of STAC and Molweni1k. Evaluation uses the original STAC and Molweni test sets. Parsers' score variations from those reported in prior work, as also noted in (Wang et al., 2021; Bennis et al., 2023), may stem from hardware differences. To ensure fairness, all experiments, with and without augmentation, were executed on the same machine[11].

Performance on Discourse Parsing Table 2 presents the performance of SDDP and DAMT across four training configurations: (i) Naïve, using only the original corpus; (ii) augmentation via context-free rephrasing without incorporating speaker contexts; (iii) augmentation via persona-based rephrasing; and (iv) augmentation with MIMIC-generated dialogues[12]. In each configu-

[8] Classifier performance is reported in Appendix D.
[9] Hyperparameters of the models are listed in Appendix A.
[10] Baseline approaches are detailed in Appendix E.
[11] A workstation with an NVIDIA RTX 4090 GPU (24GB).
[12] Appendices G and I provide, respectively, an analysis and a discussion of the augmented training data.

Parser	Strategy	STAC		Molweni	
		UAS	*LAS*	*UAS*	*LAS*
SDDP	CG	72.54 ± 0.5	56.99 ± 0.2	79.01 ± 0.4	54.26 ± 0.4
	+MA	72.95 ± 0.8	57.28 ± 0.5	79.21 ± 0.3	54.88 ± 0.2
	+MA+MI	73.53 ± 0.3	57.63 ± 0.4	79.89 ± 0.2	55.36 ± 0.4
	+MA+MI+IR	73.72 ± 0.4	57.76 ± 0.2	79.99 ± 0.4	55.43 ± 0.4
	MIMIC	**74.41** ± 0.3	**58.10** ± 0.3	**80.60** ± 0.4	**55.75** ± 0.3
DAMT	CG	70.79 ± 0.4	50.22 ± 0.6	76.85 ± 0.5	54.02 ± 0.5
	+MA	71.06 ± 0.2	51.28 ± 0.5	77.08 ± 0.5	54.21 ± 0.5
	+MA+MI	71.39 ± 0.3	51.70 ± 0.4	78.01 ± 0.3	54.68 ± 0.3
	+MA+MI+IR	71.48 ± 0.2	51.85 ± 0.4	78.24 ± 0.3	54.83 ± 0.3
	MIMIC	**71.72** ± 0.4	**52.34** ± 0.3	**79.12** ± 0.4	**55.29** ± 0.4

Table 3: Ablation results for SDDP and DAMT trained on augmented datasets.

ration, rephrasing is applied to the utterances of the top MASK-ranked speaker.

Results show that rephrasing-based augmentation strategies generally outperforms the Naïve baseline, though UAS drops for context-free and persona-based rephrasing with SDDP on STAC. Among baselines, persona-based performs better than context-free, highlighting the value of speaker personalization in the rephrasing process. MIMIC achieves even stronger results by using dialogue histories instead of predefined personas. Indeed, although persona-based rephrasing can produce utterances that resemble human-like style, the findings suggest that predefined personas might be less effective than concrete examples of how a speaker reacts in specific relational contexts. Overall, generic personas offer limited support for context-sensitive rephrasing, as the LLM has to depend on broad traits when reformulating utterances in relationally grounded situations. In contrast, MIMIC leverages dialogue histories that are intrinsically tied to the relational context of the utterances being rephrased, potentially producing more faithful rewritings.

Ablation Study Table 3 presents the ablation study evaluating the contribution of each MIMIC component. The results suggest that integrating constrained generation (CG) with MASK (CG+MA) outperform CG alone, indicating that selecting speakers by discourse rarity is better than random choice. Further gains come from MIRROR (CG+MA+MI), which appears to make rephrasing more effective via stylistic transfer. Adding interpersonal modeling (CG+MA+MI+IR) continues this trend, hinting at a potential role of relational context between participants. The full pipeline (MIMIC), with EDU-level refinement, further increases both UAS and LAS, showing its impact on rhetorical accuracy and semantic coherence.

Performance with Multi-speaker Replacement To further evaluate MIMIC, we test whether progressively replacing speakers ranked by MASK improves parsing accuracy, as detailed in Table 4. In STAC (Molweni1k, resp.) only 0.9% (8.9%, resp.) of dialogues involve ≥ 5 (≥ 6, resp.) speakers, so these cases were grouped into 4 (5, resp.) speaker replacements. Since MASK prioritizes structurally rare speakers, top-ranked ones add richer discourse patterns, while lower-ranked yield diminishing returns and may add noise. This plausibly explains why gains plateau after a point: peak accuracy occurs with three replacements on STAC and four on Molweni1k, reflecting their differing speaker distributions (see Table 1). Overall, multi-speaker replacement outperforms the Naïve baseline, with SDDP improving $+1.9\%/+2.1\%$ on STAC and $+3.1\%/+4.6\%$ on Molweni, and DAMT yielding $+3.5\%/+8.2\%$ on STAC and $+3.7\%/+3.2\%$ on Molweni (UAS/LAS).

Performance Gains in Relation Prediction We assess MIMIC's impact on relation accuracy by testing whether synthetic dialogues help parsers predict underrepresented relations. Table 5a shows relation distributions in STAC and Molweni1k, while Figure 3 reports per-relation accuracy for SDDP and DAMT. Results show MIMIC improves accuracy on underrepresented relations.

While synthetic dialogues may add noise to frequent relations, as seen in the slight decrease for *Q-Elab* in STAC with SDDP, this is not consistent. MIMIC even improves accuracy on common relations like *Comment*, *QAP*, and *Clarification-Question* with DAMT on Molweni1k. Overall, both

Parser	#RS	STAC		Molweni	
		UAS	*LAS*	*UAS*	*LAS*
SDDP	1	74.41 ± 0.3	58.10 ± 0.3	80.60 ± 0.4	55.75 ± 0.3
	2	74.55 ± 0.4	58.28 ± 0.2	80.89 ± 0.3	55.91 ± 0.4
	3	**74.80** ± 0.4	**58.31** ± 0.2	81.53 ± 0.3	56.11 ± 0.2
	4	74.75 ± 0.3	58.27 ± 0.3	**81.56** ± 0.2	**56.54** ± 0.2
	5	–	–	**81.56** ± 0.3	56.47 ± 0.3
DAMT	1	71.72 ± 0.4	52.34 ± 0.3	79.12 ± 0.4	55.29 ± 0.4
	2	72.39 ± 0.3	52.40 ± 0.3	79.30 ± 0.3	55.45 ± 0.2
	3	**73.01** ± 0.2	**53.74** ± 0.3	79.57 ± 0.2	55.61 ± 0.3
	4	72.99 ± 0.3	53.68 ± 0.4	**79.70** ± 0.3	**55.73** ± 0.3
	5	–	–	79.64 ± 0.3	55.70 ± 0.2

Table 4: Impact of multi-speaker replacement in MIMIC. #RS = # of replaced speakers per dialogue.

ID	Relation	STAC	Molweni1k
0	Comment	16.94	31.38
1	Elaboration	8.57	2.61
2	QAP	23.93	20.01
3	Q-Elab	5.09	3.16
4	Explanation	4.36	1.40
5	Result	3.54	2.61
6	Continuation	9.58	6.43
7	Ack	12.32	3.31
8	Contrast	4.42	1.34
9	Conditional	1.15	1.05
10	Correction	1.91	1.16
11	Background	0.86	0.41
12	Parallel	1.73	0.30
13	Alternation	0.98	0.32
14	Clarification-Question	3.79	24.11
15	Narration	0.83	0.40

(a)

Distance	STAC	Molweni1k
1	55.81	64.36
2	21.18	21.83
3	10.55	7.49
4	5.15	3.18
5	3.10	1.51
≥ 6	< 1.50	< 1.50

(b)

Table 5: Distribution (%) of relation types (a) and link distances (b) in original training corpora. No link distance ≥ 6 reaches 1.50% of the overall distribution.

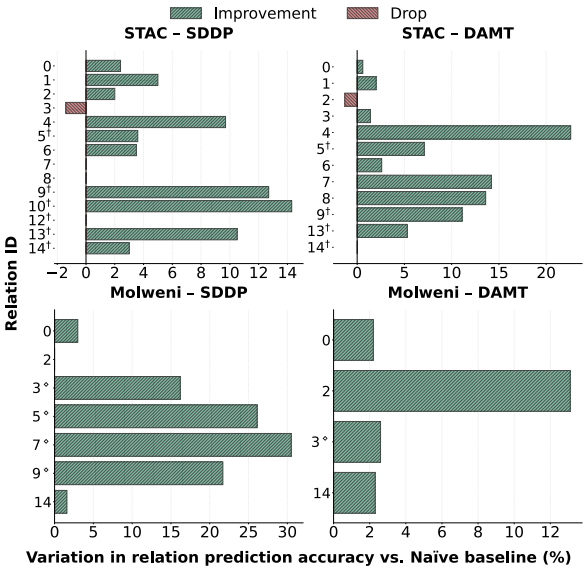


Figure 3: Relation prediction accuracy (%) for SDDP and DAMT trained with MIMIC, and variation (%) vs. the Naïve baseline. IDs align with Table 5a. † and ⋄ denote underrepresented relation types (<4%) in STAC and Molweni1k. Unpredicted relations are omitted.

parsers benefit, with notable gains on underrepresented relations, highlighting MIMIC's impact on relation-level parsing.

Performance Gains in Link Prediction Figure 4 indicates that incorporating MIMIC-generated dialogues in parser training also improves link prediction, especially for rare links (see Table 5b for link-distance distributions in STAC and Molweni1k). For example, SDDP improves on links of lengths 4 and 5 in both datasets and even predicts a long-distance link of length 13 in STAC. Similarly, DAMT outperforms the Naïve baseline on links of length 6.

As with relation prediction, slight drops sometimes occur on frequent links, e.g., length 2 for SDDP on STAC or length 1 for DAMT on Molweni, though the trend is inconsistent. MIMIC also boosts frequent links, such as length 1 for SDDP on STAC and length 2 for DAMT on Molweni.

Comparison with Unconstrained Generation Table 6 shows that LLM-generated synthetic dialogues improves UAS over the Naïve baseline (see Table 2). However, MIMIC attains even greater improvements. This is likely due to its constrained rephrasing grounded in authentic speaker styles, in contrast to the open-ended generation of the

Parser	Approach	STAC		Molweni	
		UAS	*LAS*	*UAS*	*LAS*
SDDP	Dialogue-level Generation	73.71 ± 0.2	57.69 ± 0.3	79.73 ± 0.1	54.41 ± 0.4
	MIMIC (Ours)	**74.80** ± 0.4	**58.31** ± 0.2	**81.56** ± 0.2	**56.54** ± 0.2
DAMT	Dialogue-level Generation	71.04 ± 0.3	51.18 ± 0.4	77.81 ± 0.3	54.64 ± 0.2
	MIMIC (Ours)	**73.01** ± 0.2	**53.74** ± 0.3	**79.70** ± 0.3	**55.73** ± 0.3

Table 6: Performance of SDDP and DAMT on STAC and Molweni test sets across two training setups. MIMIC improvements are statistically significant ($p < 0.05$).

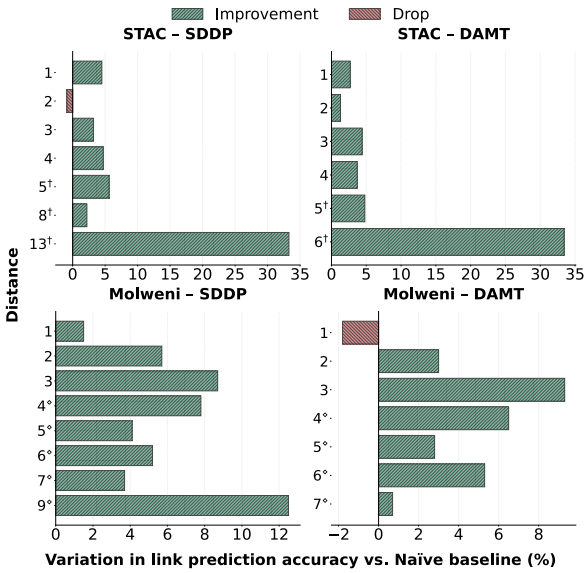


Figure 4: Link prediction accuracy (%) for SDDP and DAMT trained with MIMIC, and variation (%) vs. the Naïve baseline. † and ⋄ denote underrepresented distances (<4%) in STAC and Molweni1k. Unpredicted link distances are omitted.

dialogue-level approach, which is more prone to producing hallucinations (Yang et al., 2023).

LAS gains over the Naïve baseline stem from better prediction of relations like *QAP* and *Comment*, highlighting the challenge of dialogue-level generation, which demands deep understanding of discourse relations, a known difficulty for LLMs (Chan et al., 2023; Fan et al., 2024). MIMIC mitigates this by rephrasing existing utterances instead of generating full dialogues, with an EDU refinement step enforcing rhetorical consistency. This constrained approach preserves discourse structure more reliably than unconstrained generation.

Incremental Evaluation on Molweni We evaluate MIMIC on Molweni using an incremental setup that mirrors real-world augmentation practices. Starting with $1k$ real dialogues, we add $1k$ randomly selected instances up to $9k$ (the full dataset), generating an equal number of synthetic dialogues with MIMIC to double the training set at each step.

Figure 5 reports SDDP and DAMT performance under Naïve and MIMIC-based training setups. MIMIC consistently improves UAS/LAS, with largest gains in low-resource conditions and continued benefits as data grows. Notably, MIMIC's improvements parallel those from adding real data, showing that (a) discourse parsing remains difficult even with more gold data, and (b) MIMIC helps close the gap. In many cases, MIMIC-augmented setups even outperform larger real-data baselines (e.g., $2k$ real + $2k$ MIMIC $> 3k$ real), confirming its value as a complement to costly annotations.

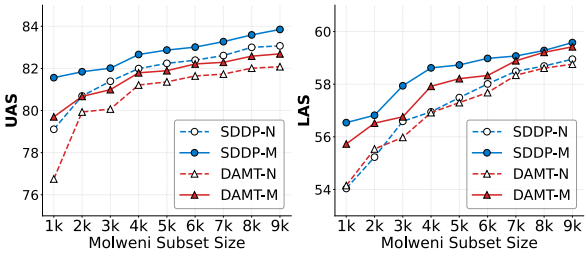


Figure 5: Parser performance on Molweni with incremental data augmentation. N = Naïve, M = MIMIC.

7 Conclusion

This paper presents MIMIC, a framework for dialogue augmentation in discourse parsing. Across both STAC and Molweni, parsers trained with MIMIC consistently outperformed those trained on original or alternatively augmented data, achieving higher accuracy in both structure and relation prediction. Results demonstrate that MIMIC is especially effective for underrepresented discourse patterns and in low-resource conditions, showing that style-informed rephrasings can enrich training corpora and strengthen parser robustness.

8 Limitations

The computational complexity of MIMIC stems from its three components: MASK, MIRROR, and LLM-based rephrasing. The MASK metric, which ranks speakers by the rarity of their discourse relations and link distances, scales linearly with the number of active speakers in a dialogue, i.e., $O(|S_D|)$. MIRROR, which identifies stylistically appropriate substitutes from the corpus, scales with the number of speakers in the entire dataset, excluding those in the current dialogue, i.e., $O(|S_C \setminus S_D|)$. Although both are theoretically linear, practical optimizations reduce runtime. For MASK, efficiency improves through precomputing and indexing speaker-level discourse statistics, enabling constant-time lookups instead of repeated scans. For MIRROR, clustering speakers by discourse style or using approximate nearest-neighbor search (Malkov and Yashunin, 2018) yields sub-linear performance while maintaining substitute quality. The main computational cost lies in EDU generation and refinement, which depend on LLM inference speed. On a Mac Studio (M2 Ultra, 192 GB RAM) running Llama 3.3 70B (8-bit), rephrasing a single EDU averages 6 seconds without refinement and 17 seconds with it. This can be mitigated using optimized inference frameworks such as *vLLM*[13].

The refinement phase in MIMIC utilizes binary classifiers trained on labeled data to evaluate rhetorical coherence. While their effectiveness may be limited by the scarcity of annotated examples, experimental results show that parsers trained on augmented datasets achieve improved performance even for rare discourse relations. This suggests that, despite potential noise in classifier predictions, the augmentation mechanism effectively facilitates their learning. Furthermore, relation classification remains challenging even for advanced models such as ChatGPT (Chan et al., 2023; Fan et al., 2024), indicating that supervised methods still offer more reliable performance.

Concerns arise regarding the scale of augmentation. Although MIMIC enables large-scale corpus expansion, prior work shows that beyond a certain point additional data may not improve performance (Longpre et al., 2020) and may even lead to degradation (Okimura et al., 2022). Future work should therefore focus on assessing and regulating the diversity of generated dialogues to avoid redundancy.

The space of speaker substitution combinations in MIMIC is extremely large. Empirical results show that replacing multiple speakers yields measurable gains, yet the full combinatorial space remains unexplored due to high computational and resource demands. This space also includes cross-domain stylistic substitution, where speakers from different datasets rephrase utterances in the target corpus. Exploring such cross-domain transfer could further enhance MIMIC's generalizability.

Finally, although MIMIC has significantly improved discourse parser performance across benchmark datasets, it is important to note that all synthetic data in this study was generated using a quantized version of the Llama 3.3 70B model. While quantization provides substantial computational benefits, it can also limit the model's expressive capacity and stylistic fidelity. Consequently, employing more powerful or full-precision LLMs could further enhance the quality and effectiveness of MIMIC's augmentations, highlighting clear opportunities for future improvement.

9 Ethical Considerations

MIMIC is a framework for augmenting multi-party dialogue corpora using speaker stylistic transfer via LLMs. As with any system generating synthetic data, it raises ethical concerns related to data provenance, representation, and downstream use.

First, MIMIC operates solely on publicly available datasets, namely STAC and Molweni, and does not rely on personally identifiable or sensitive user data. The framework further avoids predefined speaker profiles, instead inferring stylistic traits from the dialogue context, thereby reducing risks related to profiling and privacy.

Second, while the generation pipeline includes mechanisms to ensure coherence, such as discourse-constrained prompting and classifier-based refinement, synthetic utterances may still deviate from the intended meaning or introduce inappropriate content. This limitation reflects broader safety concerns in LLM-based generation and motivates the need for human oversight and post-generation filtering when using synthetic data.

Finally, although synthetic augmentation can improve model robustness, it may also amplify biases present in the original corpora. Careful evaluation of fairness and generalizability, together with transparent reporting of generation settings and failure cases, is essential for responsible deployment.

[13] https://github.com/vllm-project/vllm

References

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, page 928–937.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC*. European Language Resources Association (ELRA).

Zineb Bennis, Julie Hunter, and Nicholas Asher. 2023. A simple but effective model for attachment in discourse parsing with multi-task learning for relation labeling. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)*, pages 3404–3409. Association for Computational Linguistics.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

Akshay Chaturvedi, Kate Thompson, and Nicholas Asher. 2024. Nebula: A discourse aware minecraft builder. In *Findings of the Association for Computational Linguistics (EMNLP 2024)*, pages 6431–6443. Association for Computational Linguistics.

Haonan Chen, Zhicheng Dou, Kelong Mao, Jiongnan Liu, and Ziliang Zhao. 2024. Generalizing conversational dense retrieval via LLM-cognition data augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages 2700–2718. Association for Computational Linguistics.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022. Weakly supervised data augmentation through prompting for dialogue understanding. *CoRR*, abs/2210.14169.

Chuanqi Cheng, Quan Tu, Wei Wu, Shuo Shang, Cunli Mao, Zhengtao Yu, and Rui Yan. 2024. "in-dialogues we learn": Towards personalized dialogue without pre-defined profiles through in-dialogue learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pages 10408–10422. Association for Computational Linguistics.

Ta-Chung Chi and Alexander I. Rudnicky. 2022. Structured dialogue discourse parsing. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, (SIGDIAL 2022)*, pages 325–335. Association for Computational Linguistics.

Gaetano Cimino, Chuyuan Li, Giuseppe Carenini, and Vincenzo Deufemia. 2024. Coherence-based dialogue discourse structure extraction using open-source large language models. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, (SIGDIAL 2024)*, pages 297–316. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics (ACL 2024)*, pages 1679–1705. Association for Computational Linguistics.

Jack Edmonds and 1 others. 1967. Optimum branchings. *Journal of Research of the national Bureau of Standards B*, 71(4):233–240.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2024. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, (LREC/COLING 2024)*, pages 16998–17010. ELRA and ICCL.

Yaxin Fan, Peifeng Li, Fang Kong, and Qiaoming Zhu. 2022. A distance-aware multi-task framework for conversational discourse parsing. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*, pages 912–921. International Committee on Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, (IJCAI 2022)*, pages 5453–5460. ijcai.org.

Kory Floyd and Mark T Morman. 1997. Affectionate communication in nonromantic relationships: Influences of communicator, relational, and contextual factors. *Western Journal of Communication (includes Communication Reports)*, 61(3):279–298.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jia-Chen Gu, Zhen-Hua Ling, Yu Wu, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Detecting speaker personas from conversational texts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pages 1126–1136. Association for Computational Linguistics.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based

natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)*, pages 523–540. Association for Computational Linguistics.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful persona-based conversational dataset generation with large language models. In *Findings of the Association for Computational Linguistics (ACL 2024)*, pages 15245–15270. Association for Computational Linguistics.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 12930–12949. Association for Computational Linguistics.

Frederic Kirstein, Muneeb Khan, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. You need to MIMIC to get FAME: Solving meeting transcript scarcity with multi-agent conversations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11482–11525, Vienna, Austria. Association for Computational Linguistics.

Ivan Lee and Taylor Berg-Kirkpatrick. 2022. Helo: Learning-free lookahead decoding for conversation infilling. In *Findings of the Association for Computational Linguistics (EMNLP 2022)*, pages 4996–5008. Association for Computational Linguistics.

Seungmi Lee, Midan Shim, and Kyong-Ho Lee. 2025. Mirror: Multi-party dialogue generation based on interpersonal relationship-aware persona retrieval. *Expert Systems with Applications*, 276:127141.

Chuyuan Li, Patrick Huber, Wen Xiao, Maxime Amblard, Chloe Braud, and Giuseppe Carenini. 2023. Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2562–2579, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 2642–2652. International Committee on Computational Linguistics.

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022. Controllable dialogue simulation with in-context learning. In *Findings of the Association for Computational Linguistics (EMNLP 2022)*, pages 4330–4347. Association for Computational Linguistics.

Zhengyuan Liu and Nancy F. Chen. 2021. Improving multi-party dialogue discourse parsing via domain integration. *CoRR*, abs/2110.04526.

Zhenhua Liu, Tong Zhu, Jianxiang Xiang, and Wenliang Chen. 2024. Controllable and diverse data augmentation with large language model for low-resource open-domain dialogue generation. *CoRR*, abs/2404.00361.

Shayne Longpre, Yu Wang, and Chris DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4401–4411, Online. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2015)*, pages 285–294. The Association for Computer Linguistics.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning LLMs to use memos for consistent long-range open-domain conversation. *CoRR*, abs/2308.08239.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2023. Enhanced speaker-aware multi-party multi-turn dialogue comprehension. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, 31:2410–2423.

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 555–564. ACM.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Khyati Mahajan and Samira Shaikh. 2021. On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2021)*, pages 338–352. Association for Computational Linguistics.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Shikib Mehri, Yasemin Altun, and Maxine Eskénazi. 2022. LAD: language models as data for zero-shot dialog. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, (SIGDIAL 2022)*, pages 595–604. Association for Computational Linguistics.

Itsuki Okimura, Machel Reid, Makoto Kawano, and Yutaka Matsuo. 2022. On the impact of data augmentation on downstream performance in natural language processing. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 88–93, Dublin, Ireland. Association for Computational Linguistics.

Hongjin Qian, Zhicheng Dou, Yutao Zhu, Yueyuan Ma, and Ji-Rong Wen. 2021. Learning implicit user profile for personalized retrieval-based chatbot. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM 2021)*, pages 1467–1477.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Ryan Shea and Zhou Yu. 2023. Building persona consistent dialogue agents with offline reinforcement learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 1778–1795. Association for Computational Linguistics.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 7007–7014. AAAI Press.

Ante Wang, Linfeng Song, Hui Jiang, Shaopeng Lai, Junfeng Yao, Min Zhang, and Jinsong Su. 2021. A structure self-aware model for discourse parsing on multi-party dialogues. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021)*, pages 3943–3949.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-party chat: Conversational agents in group settings with humans and models. *CoRR*, abs/2304.13835.

Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang, and Hai Zhao. 2023. RefGPT: Dialogue generation of GPT, by GPT, and for GPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2511–2535. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 2204–2213. Association for Computational Linguistics.

Ziyan Zhang, Yang Hou, Chen Gong, and Zhenghua Li. 2025. Data augmentation for cross-domain parsing via lightweight LLM generation and tree hybridization. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 11235–11247. Association for Computational Linguistics.

Meng Zhao, Lifang Wang, Zejun Jiang, Yushuang Liu, Ronghan Li, Zhongtian Hu, and Xinyu Lu. 2024. From easy to hard: Improving personalized response generation of task-oriented dialogue systems by leveraging capacity in open-domain dialogues. *Knowl. Based Syst.*, 295:111843.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics (ACL 2023)*, pages 1552–1568. Association for Computational Linguistics.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL 2022)*, pages 5808–5820. Association for Computational Linguistics.

A Implementation Details

In the context of the EDU refinement process, binary classifiers were individually trained on EDU pairs corresponding to specific relations from the training set and evaluated on the respective pairs from the test set. Training was conducted using a maximum sequence length of 128, a batch size of 15, 10 epochs, and a learning rate of 1×10^{-5}. Each classifier was applied in accordance with the relevant discourse relations, with the refinement process iterated up to two times.

For both EDU generation and refinement, prompt tuning was conducted using a validation set. This process involved evaluating the quality of

rephrased utterances with respect to the linguistic style of the substitute speaker and the nature of the underlying interpersonal relationships. Consequently, for each relevant relation considered during the augmentation process, up to five discourse triples were selected.

For the training of SDDP and DAMT, we largely relied on the hyperparameters specified in their respective original publications (Chi and Rudnicky, 2022; Fan et al., 2022), applying only minor adjustments to accommodate the newly augmented corpora. In particular, for DAMT, a batch size of 180 was used when training on the augmented STAC dataset. In contrast, SDDP was trained on the same dataset for 4 epochs with a batch size of 2. When training SDDP on the augmented Molweni dataset, we employed a batch size of 8 and allocated the initial 5% of training steps for warm-up.

B Prompt Templates

The structured prompt templates supporting the MIMIC augmentation framework's generation and refinement steps are embedded throughout the following discussion. The generation phase begins with a system prompt that instructs the model to rephrase Speaker A's utterance in the style of Speaker B, ensuring that both discourse coherence and interactional intent are preserved:

EDU Generation - System Prompt

You are an advanced assistant specializing in transforming dialogues while preserving the coherence of discourse relations. You will be provided with a structured dialogue where speaker A's utterances interact with one or more other speakers, creating multiple discourse relations.

Your task is to rephrase a specific utterance from speaker A according to the linguistic style of speaker B while also considering the speaking tendencies of the other participants in the dialogue.

Your primary objectives are:
- Rephrase speaker A's utterance using speaker B's typical language patterns, ensuring that speaker B's linguistic style is preserved.
- The utterance should be rewritten based on examples of speaker B's past interactions.
- Maintain all discourse relations involving the rephrased utterance. If speaker A's utterance is part of multiple relations, ensure the rewritten version satisfies all of them.
- If linguistic tendencies of the other speakers are provided, take them into account to ensure the rephrased utterance fits naturally within the broader conversational dynamics. Otherwise, focus solely on speaker B's past interactions.
- Ensure that the logical coherence and meaning of the conversation are preserved.

You will be provided with:
- The original dialogue, including all relevant interactions, where one specific utterance by speaker A needs to be rephrased.
- Examples of speaker B's past interactions categorized by discourse relation type.
- (Optional) Examples of how the other speakers in the dialogue typically communicate, to account for their linguistic tendencies.

For each relation type, a brief explanation is provided to help you preserve its meaning:
{Relation Explanations}

This is followed by a user prompt that supplies the dialogue context and stylistic examples to guide the model in producing an accurate rephrasing:

EDU Generation - User Prompt

Here is a structured dialogue where speaker A's utterance needs to be rephrased according to the linguistic style of speaker B. Ensure that speaker B's rewritten utterance matches the provided examples of their past interactions and preserves all discourse relations in the original dialogue.

If examples of the linguistic tendencies of the other speakers are provided, ensure the rephrased utterance also aligns with these interactional nuances. If not, focus solely on reproducing speaker B's style.

Speaker A's Original Dialogue:
{Speaker A's Discourse Relations}

Speaker B's Dialogue History:
{Speaker B's Dialogue History}

Interpersonal Relationships:
{Other Speakers' Dialogue History}

Please rephrase the **{Speaker A's Utterance}** while ensuring that:
- Speaker B's linguistic style is maintained.
- All discourse relations in the original dialogue remain intact.
- The relational and interpersonal dynamics in the dialogue are preserved when possible.

Generate only the rephrased utterance, without any explanation or reasoning.

The process then transitions to refinement. A

dedicated system prompt prompts the model to identify potential disruptions in rhetorical relations within the rephrased EDU and to propose high-level suggestions for improvement:

EDU Refinement - System Prompt

You are an expert in understanding why a rephrased utterance no longer satisfies the **{Relation Type}** relation when compared to another utterance.

The **{Relation Type}** relation is defined as follows:
{Relation Explanation}

Given the original pair of utterances that previously satisfied this relation, one of the utterances has now been rephrased. Your task is to provide high-level guidance on how the rephrased utterance should be adjusted to restore the relation, without providing a specific rephrasing. The suggestion should focus on the key aspects that need to be preserved or modified in the rephrased utterance to ensure it aligns with the original utterance in terms of the defined relation.

This is paired with a user prompt that presents both the original and rephrased utterances, asking the model to offer revisions that enhance alignment while retaining meaning:

EDU Refinement - User Prompt

You are given the following inputs:
- **{Original Utterance 1}**
- **{Original Utterance 2}**
- **{Relation Type}**
- **{Rephrased Utterance}**

The rephrased utterance is a modified version of **{Original Utterance 1}**, and it no longer satisfies the given relation with **{Original Utterance 2}**. Your task is to generate a high-level suggestion on how **{Rephrased Utterance}** can be improved to restore the relation **{Relation Type}** with **{Original Utterance 2}**.

Here's what you need to do:
- Analyze the two original utterances and understand the relation **{Relation Type}** between them.
- Examine **{Rephrased Utterance}** and determine why it no longer satisfies the relation with **{Original Utterance 2}**.
- Provide a high-level suggestion on how **{Rephrased Utterance}** can be modified to restore the relation. Focus on the key aspects that need to be preserved or modified in **{Rephrased Utterance}** to ensure it aligns with **{Original Utterance 2}** according to the given relation.

Do not generate a new rephrased utterance, nor reference the current one in your output. Instead, provide concise, one-sentence guidance on how the rephrasing should be approached.

Collectively, these templates embody the framework's emphasis on preserving rhetorical coherence and speaker-specific interaction patterns across the augmentation pipeline. The discourse relations referenced in these prompts are detailed in Table 7.

C Comparison of EDU Generation Strategies

A key design consideration in the MIMIC framework concerns the strategy for generating EDUs. While the main experiments adopt an independent (IND) approach—where each EDU is rephrased in isolation using gold-standard context—it is also natural to consider an incremental (INC) alternative that conditions each generation step on previously produced outputs.

Table 8 reports the results of this comparison. The IND strategy yields consistently superior performance compared to INC, irrespective of the parser or dataset. The gap stems from context handling: INC conditions each EDU on previously generated ones, improving local fluency but risking error accumulation that harms coherence. IND avoids this by generating each EDU in isolation, relying on gold context, thus reducing cascading errors and producing more structurally sound output. These observations carry important practical implications. While INC should better mirrors human conversational patterns, results suggest that for structure-sensitive tasks like discourse parsing, robustness outweighs realism. IND generation seems therefore better suited for augmentation, especially in structurally complex settings.

D Performance Evaluation of Relation Classification Models

Tables 9 and 10 show the results of the BERT-based classifiers used for relation classification in EDU refinement on the STAC and Molweni corpora, respectively. The classifiers were trained on the complete STAC corpus and the Molweni1k subset of the Molweni corpus. As shown, all classifiers achieve a minimum accuracy of 71%, with precision not falling below 70%, demonstrating their potential for validating relations between EDU pairs. The variations in performance observed can be attributed to the

Relation	Description
Comment	A Comment relation typically indicates that one utterance provides a comment or opinion on the content of another utterance. It shows a speaker's perspective or evaluation of the preceding statement.
Clarification-Question	In a Clarification-Question relation, one utterance poses a question seeking clarification or additional information about the content of another utterance. It implies a request for further explanation.
Elaboration	An Elaboration relation signifies that one utterance expands upon or provides more details about the content of another utterance. It is used to enhance understanding by offering additional information or context.
Acknowledgment	An Acknowledgment relation indicates that one utterance acknowledges or recognizes the content of another utterance. It signifies that the speaker has taken note of what was said.
Continuation	A Continuation relation suggests that one utterance continues the topic or discussion from a previous utterance. It signifies a logical progression in the conversation.
Explanation	An Explanation relation pertains to one utterance offering an explanation or clarification in response to a question or confusion expressed in another utterance. It aids in providing clarity.
Conditional	A Conditional relation implies that one utterance presents a condition or hypothetical scenario related to the content of another utterance. It often involves "if-then" statements.
Question-Answer	A Question-Answer relation indicates that one utterance contains a question, and another utterance immediately follows with an answer to that question. It demonstrates a direct question-and-answer interaction.
Alternation	An Alternation relation shows that two utterances present alternative options or choices. It is used when discussing multiple possibilities or courses of action.
Q-Elab	A Q-Elab relation signifies that one utterance asks a question, and another utterance follows with an elaboration or further explanation of the question or its context.
Result	A Result relation indicates that one utterance discusses the outcome or consequence of the content presented in another utterance. It shows a cause-and-effect relation.
Background	In a Background relation, one utterance provides background information or context that is relevant to the content of another utterance. It helps set the stage for the discussion.
Narration	A Narration relation holds when the main eventualities of two utterances occur in sequence.
Correction	A Correction relation shows that one utterance corrects or revises the content of another utterance. It is used to rectify errors or inaccuracies.
Parallel	A Parallel relation occurs when two or more utterances share similar or related content, often in a parallel or analogous manner. It emphasizes similarities or comparisons.
Contrast	A Contrast relation signifies that one utterance presents content that is in contrast or opposition to the content of another utterance. It highlights differences or contradictions in the conversation.

Table 7: Descriptions of discourse relations in STAC and Molweni corpora.

Parser	Approach	STAC		Molweni	
		UAS	*LAS*	*UAS*	*LAS*
SDDP	MIMIC$_{\text{INC}}$	72.22 ± 0.3	57.28 ± 0.2	79.74 ± 0.4	55.11 ± 0.3
	MIMIC$_{\text{IND}}$	**74.41** ± 0.3	**58.10** ± 0.3	**80.60** ± 0.4	**55.75** ± 0.3
DAMT	MIMIC$_{\text{INC}}$	71.09 ± 0.3	51.49 ± 0.3	78.02 ± 0.3	54.68 ± 0.4
	MIMIC$_{\text{IND}}$	**71.72** ± 0.4	**52.34** ± 0.3	**79.12** ± 0.4	**55.29** ± 0.4

Table 8: Performance of SDDP and DAMT on STAC and Molweni test sets across Incremental (INC) and Independent (IND) strategies.

differing amounts of training pairs available for each relation type.

E Baseline Approaches

To assess the effectiveness of our proposed augmentation strategy, we compare it against three baseline approaches that differ in their use of speaker and discourse information: (i) *Speaker-Level Context-Free Rephrasing*, (ii) *Speaker-Level Persona-Based Rephrasing*, and (iii) *Dialogue-Level Discourse-Aware Generation*.

Speaker-Level Context-Free Rephrasing This baseline is designed to evaluate whether using the dialogue history of a target speaker yields better rephrasings than context-free paraphrasing. Specifically, it generates a reformulation of each utterance without accounting for speaker interaction patterns. The model receives the utterance along with its associated discourse relations and produces a rephrased version intended to preserve the original semantic content and relational coherence.

Context-Free Rephrasing - System Prompt

```
You are an advanced assistant specializing
in transforming dialogues while preserving
the coherence of discourse relations. You
```

Metric (%)	Relation															
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Accuracy	83	83	93	87	81	73	77	89	73	89	71	92	87	92	89	77
Precision	88	86	93	86	79	72	79	85	75	82	74	86	82	86	96	89
Recall	77	78	94	88	84	75	75	95	72	100	71	100	93	100	82	74
F_1*-score*	82	82	93	87	82	73	77	90	73	90	72	92	87	92	88	81

Table 9: Performance metrics of BERT-based classifiers for each relation type on STAC. The relation IDs correspond to those listed in Table 5a.

Metric (%)	Relation															
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Accuracy	71	75	89	79	73	81	71	87	80	86	73	81	83	78	89	75
Precision	72	73	89	79	76	83	71	88	78	84	73	86	79	86	92	70
Recall	70	78	88	79	71	78	84	87	84	89	72	75	92	74	86	88
F_1*-score*	71	75	88	79	74	80	77	87	81	86	72	80	85	80	89	78

Table 10: Performance metrics of BERT-based classifiers for each relation type on Molweni. The relation IDs correspond to those listed in Table 5a.

will be provided with a structured dialogue where speaker A's utterances interact with one or more other speakers, creating multiple discourse relations.

Your task is to rephrase a designated utterance from Speaker A in such a way that all discourse relations connected to this utterance remain intact, ensuring the original meaning is preserved while the phrasing is altered.

For each relation type, a brief explanation is provided to help you preserve its meaning:
{Relation Explanations}

Context-Free Rephrasing - User Prompt

Here is a structured dialogue where speaker A's utterance **{Speaker A's Utterance}** needs to be rephrased. Ensure that the rephrased utterance preserves all discourse relations.

Speaker A's Original Dialogue:
{Speaker A's Discourse Relations}

Generate only the rephrased utterance, without any explanation or reasoning.

Speaker-Level Persona-Based Rephrasing This baseline extends context-free rephrasing by incorporating LLM-generated speaker personas to guide the reformulation of utterances. The objective is to evaluate whether rephrasings guided by generic personas, rather than treated as neutral paraphrases, offer improvements in coherence and relation-preserving, and to compare these gains against MIMIC, which instead grounds rephrasings in dialogue histories. The process unfolds in two stages. First, a persona generation step creates a structured personality profile for each target speaker. Separate prompts were designed for the two domains considered: for STAC, the persona describes communication traits of a Catan player (e.g., competitive tone, negotiation style, argument structure), while for Molweni, it captures interactional patterns of a participant in Ubuntu-related technical discussions (e.g., cooperative troubleshooting style, choice of technical language). Each persona is represented across five dimensions (Kirstein et al., 2025): Tone, Language, Communication, Structure, and Other, ensuring a systematic description of speaker behavior. Second, a persona-based rephrasing step uses these generated traits to guide the rewriting of target utterances. The system prompt specifies that the LLM must produce a rephrasing that simultaneously (i) preserves all annotated discourse relations, (ii) maintains the original semantic intent, and (iii) reflects the stylistic attributes encoded in the provided persona. The user prompt then delivers the structured dialogue context, the utterance to be rephrased, and the corresponding persona traits, instructing the model to output the rephrased utterance.

STAC Persona Generation - System Prompt

You are a persona generator for role-playing tasks.
Your goal is to create a coherent personality profile for a player in the board game Catan.
Catan is a strategic board game where players compete to build settlements, roads, and cities using five types of resources: wheat, wood, ore, brick, and sheep. Players collect resources based on dice rolls, trade with one another, and may use the robber to block opponents.
Focus on how the player communicates and interacts during the game, considering actions like trading, negotiating, and strategic planning. Output the persona structured into the following five categories:

- **Tone:** emotional tone in speech (e.g., enthusiastic, sarcastic, calm, competitive).
- **Language:** choice of words and expressions (e.g., formal, slangy, diplomatic, blunt).
- **Communication:** typical interaction style (e.g., persuasive, confrontational, cooperative, secretive).
- **Structure:** how sentences and arguments are organized (e.g., short and direct, long and elaborate, logical step-by-step).
- **Other:** any additional quirks or habits in communication (e.g., often jokes, tends to repeat themselves, uses metaphors, impatient interruptions).

STAC Persona Generation - User Prompt

Generate a persona for a Catan player.
Describe how this player typically speaks and interacts during the game.
Output only a list containing the 5 personality traits categories: Tone, Language, Communication, Structure, Other.
For each category, provide a description of the player's style.
Do not include any additional text or explanations outside of this list.

Molweni Persona Generation - System Prompt

You are a persona generator for role-playing tasks.
Your goal is to create a coherent personality profile for a participant in Ubuntu-related conversations.
Ubuntu is a Linux-based operating system. The dialogues take place between users discussing technical issues, sharing solutions, and offering advice related to system usage and configuration.
Focus on how the participant communicates and interacts during discussions, considering behaviors like asking for help, providing guidance, debating solutions, and troubleshooting collaboratively.
Output the persona structured into the following five categories:

- **Tone:** emotional tone in speech (e.g., enthusiastic, sarcastic, calm, competitive).
- **Language:** choice of words and expressions (e.g., formal, slangy, diplomatic, blunt).
- **Communication:** typical interaction style (e.g., persuasive, confrontational, cooperative, secretive).
- **Structure:** how sentences and arguments are organized (e.g., short and direct, long and elaborate, logical step-by-step).
- **Other:** any additional quirks or habits in communication (e.g., often jokes, tends to repeat themselves, uses metaphors, impatient interruptions).

Molweni Persona Generation - User Prompt

Generate a persona for a participant in Ubuntu-related conversations.
Describe how this participant typically speaks and interacts during discussions about system usage, troubleshooting, and configuration.
Output only a list containing the 5 personality traits categories: Tone, Language, Communication, Structure, Other.
For each category, provide a description of the player's style.
Do not include any additional text or explanations outside of this list.

Persona-Based Rephrasing - System Prompt

You are an advanced assistant specializing in transforming dialogues while preserving the coherence of discourse relations.
You will be provided with a structured dialogue where speaker A's utterances interact with one or more other speakers, creating multiple discourse relations, accompanied by a list of personality traits to inform the rephrasing of a single utterance from Speaker A.
Your task is to rephrase a designated utterance from Speaker A in such a way that:

- All discourse relations connected to this utterance remain intact.
- The original meaning is preserved.
- The utterance reflects the provided personality traits.

For each relation type, a brief explanation is provided to help you preserve its meaning:
{Relation Explanations}

Persona-Based Rephrasing - User Prompt

Here is a structured dialogue where speaker A's utterance **{Speaker A's Utterance}** needs to be rephrased. Ensure that the rephrased utterance preserves all discourse relations and reflects the provided personality traits.

Speaker A's Original Dialogue:
{Speaker A's Discourse Relations}

Personality Traits:
{Personality Traits}

Generate only the rephrased utterance, without any explanation or reasoning.

Dialogue-Level Discourse-Aware Generation This baseline constructs full dialogues from scratch, conditioned on a set of predefined rhetorical relations between EDUs. This approach aims to ensure global coherence by aligning the generated content with a discourse graph and maintaining a consistent interaction flow among participants. The objective is to evaluate whether generating entire dialogues conditioned on existing discourse structures—rather than rephrasing utterances as in MIMIC—can effectively preserve coherence and relational accuracy, thereby assessing the impact of unconstrained generation in dialogue discourse parsing. Prompts include background information about the dialogue domain (e.g., board gameplay for STAC or technical discussion for Molweni), the list of relevant discourse relation definitions, and the number of distinct speakers.

STAC Dialogue Generation - System Prompt

You are an advanced assistant specialized in generating realistic and coherent conversations between players during a game of Catan.
Catan is a strategic board game where players compete to build settlements, roads, and cities using five types of resources: wheat, wood, ore, brick, and sheep. Players collect resources based on dice rolls, trade with one another, and may use the robber to block opponents.
You will be given a list of discourse relations between fictional Elementary Discourse Units (EDUs). Each EDU is a short utterance spoken by a player during gameplay.
Each dialogue you generate must:
- Take place within the context of a Catan game.
- Involve exactly {**Number of Speakers**} players.
- Respect the discourse relations provided, which define how the EDUs are logically or rhetorically connected.

Guidelines:
- Invent the content of each EDU using natural and realistic language.
- The EDUs should reflect typical interactions in Catan, such as proposing trades, responding to dice outcomes, blocking with the robber, building, or reacting to others' actions.
- Assign each EDU to one of the fictional players (e.g., Alice, Bob, Clara, etc.).

Here are the types of discourse relations that will appear in the current dialogue:
{Relation Explanations}

STAC Dialogue Generation - User Prompt

Below is a list of discourse relations between EDUs in a fictional dialogue between players during a game of Catan.
Each relation connects two EDUs, indicating how one relates to the other (e.g., comment, explanation, question-answer, etc.).
Your task is to generate the content of each EDU (EDU0, EDU1, EDU2, ...) so that:
- The conversation reflects typical player interaction in Catan.
- The rhetorical/discourse structure aligns with the relations.
- The flow of the conversation is coherent and natural.

Return the output in the following format:
"edus": ["speaker": "...", "text": "...", "speechturn": 0, "speaker": "...", "text": "...", "speechturn": 1,
...]
Here are the discourse relations to follow:
{**Relations**}

Molweni Dialogue Generation - System Prompt

You are an advanced assistant specialized in generating realistic and coherent conversations between users discussing topics related to Ubuntu.
Ubuntu is a Linux-based operating system. The dialogues take place between users discussing technical issues, sharing solutions, and offering advice related to system usage and configuration.
You will be given a list of discourse relations between fictional Elementary Discourse Units (EDUs). Each EDU is a short utterance spoken by a user during the discussion.
Each dialogue you generate must:
- Take place within the context of a discussion about Ubuntu.
- Involve exactly {**Number of Speakers**} users.
- Respect the discourse relations provided, which define how the EDUs are logically or rhetorically connected.

Guidelines:
- Invent the content of each EDU using natural and realistic language.
- The EDUs should reflect typical interactions in Ubuntu-related discussions, such as asking for technical help, offering solutions, reporting issues, suggesting commands, or reacting to others' suggestions.
- Assign each EDU to one of the fictional users (e.g., Alice, Bob, Clara, etc.).

Here are the types of discourse relations that will appear in the current dialogue:
{Relation Explanations}

Molweni Dialogue Generation - User Prompt

Below is a list of discourse relations between EDUs in a fictional dialogue between users discussing Ubuntu.
Each relation connects two EDUs, indicating how one relates to the other (e.g., comment, explanation, question-answer, etc.).
Your task is to generate the content of each EDU (EDU0, EDU1, EDU2, ...) so that:
- The conversation reflects typical user interaction in Ubuntu-related discussions.
- The rhetorical/discourse structure aligns with the relations.
- The flow of the conversation is coherent and natural.

Return the output in the following format:
"edus": ["speaker": "...", "text": "...",
"speechturn": 0, "speaker": "...", "text": "...", "speechturn": 1,
...]
Here are the discourse relations to follow:
{**Relations**}

F Human Evaluation

We conducted a human evaluation of rephrased utterances, equally sampled from STAC and Molweni. A total of 100 dialogues (50 from each dataset) were assessed.

Annotators Three researchers, all fluent in English and with backgrounds in computer science and discourse parsing, served as annotators. Each had prior familiarity with SDRT. To ensure reliable judgments across datasets, the annotators received additional preparation: they were introduced to the rules and objectives of the Settlers of Catan game for STAC, while their technical expertise facilitated accurate assessment of Molweni. The tasks were carried out during regular working hours without specific monetary compensation. Annotators required on average 3–4 minutes per STAC dialogue and 5–7 minutes per Molweni dialogue, with the longer EDUs in Molweni accounting for the additional time.

Annotation procedure For each dialogue, annotators were presented with (i) the original dialogue with a missing utterance, (ii) the generated filler, (iii) the rhetorical relations to be preserved, and (iv) the substitute speaker's dialogue history.

Evaluation criteria Dialogues were rated along three dimensions:

- *Coherence* — how well the generated utterance fits the dialogue context (1 = not coherent, 2 = partially coherent, 3 = fully coherent).
- *Intentionality* — the extent to which rhetorical relations are preserved (1 = none, 2 = partial, 3 = full preservation).
- *Personalization* (Ma et al., 2021) — whether the utterance reflects the substitute speaker's stylistic traits (e.g., lexical choices, modality), scored 0 for not personalized and 1 for personalized.

Results Table 11 presents the results of the human evaluation conducted on STAC and Molweni. The three augmentation approaches plausibly reflect distinct strengths depending on the dataset characteristics. In STAC, where utterances are typically short and fragmentary, context-free rephrasing often produces less coherent outputs. Persona-based rephrasing offers moderate improvements by introducing stylistic markers that contribute to textual coherence. MIMIC attains the highest scores in this setting: by leveraging dialogue history and refining EDUs, it more effectively maintains rhetorical flow. Annotators even rated some MIMIC outputs more effective than gold utterances in conveying rhetorical relations, likely due to EDU refinement enforcing rhetorical fidelity. In Molweni, where utterances are generally longer and more self-contained, rhetorical cues remain explicit even without dialogue history. This likely accounts for the ability of context-free rephrasing to reach intentionality levels slightly higher than the other approaches. Nevertheless, MIMIC still improves coherence by grounding generation in prior speaker responses, which plausibly explains its superior performance.

Across both datasets, annotators frequently noted authentic personalization, showing MIMIC's gains stem not just from paraphrase diversity but from genuine stylistic transfer.

Approach	STAC			Molweni		
	C	*I*	*P*	*C*	*I*	*P*
MIMIC (Ours)	**2.79**	**2.91**	0.85	**2.77**	2.71	0.80
Context-free Rephrasing	2.51	2.61	N/A	2.74	**2.76**	N/A
Persona-based Rephrasing	2.60	2.68	N/A	2.74	2.70	N/A
Ground-Truth	2.83	2.87	N/A	2.83	2.75	N/A

Table 11: Human evaluation results. **C**: *Coherence*; **I**: *Intentionality*; **P**: *Personalization*. Fleiss Kappa scores: **C**=0.57, **I**=0.64, **P**=0.48 for STAC; **C**=0.43, **I**=0.50, **P**=0.73 for Molweni. All Fleiss Kappa scores (≥ 0.43) indicate substantial inter-annotator agreement.

G Effect of Data Quantity and Quality

To evaluate the impact of both the *quality* and *quantity* of augmented data on parser performance, we conduct an analysis varying these two factors. For quantity, the augmented dialogues are randomly partitioned into three equally sized subsets: the 943 dialogues in STAC are split into three subsets, as are the 9,000 dialogues in Molweni. For quality, we modify a small, fixed number of turns per dialogue, following prior shuffling and perturbation approaches applied to the same datasets (Li et al., 2023). Specifically, one sentence is replaced in dialogues with fewer than five EDUs, two sentences for dialogues containing six to nine EDUs, and three sentences for dialogues exceeding ten EDUs. Three levels of quality are considered: (i) *low*, in which selected utterances are randomly replaced with utterances from other dialogues; (ii) *medium*, in which selected utterances are randomly replaced with utterances drawn from those having a similarity score above 0.6, derived from cosine similarity applied to sentence embeddings obtained via the SentenceBERT model (Reimers and Gurevych, 2019); and (iii) *high*, where the original dialogues generated by MIMIC are retained. The first two configurations simulate potential hallucination scenarios that may occur with LLMs: in the low-quality setting, the replacement utterance may be entirely inconsistent with the dialogue context, while in the medium-quality setting, the replacement utterance may partially align with the context but may not be fully coherent.

Figures 6 and 7 present the results of the analysis for DAMT and SDDP, respectively. In these figures, the size of each circle corresponds to the UAS or LAS score obtained for a specific combination of augmented data quantity and quality, combined with the original corpora for parser training. Larger circles indicate higher UAS/LAS scores. As we can see, data quality exerts a decisive influence on both UAS and LAS. For each level of data quantity, optimal performance is consistently observed under the high-quality setting, in which the original MIMIC-generated dialogues are preserved without perturbations. Furthermore, a monotonic trend emerges with respect to data quantity: increasing the volume of MIMIC-generated data invariably results in higher UAS and LAS values.

In the low- and medium-quality settings, distinct trends emerge for UAS and LAS. Even when parsers are trained on augmented data containing partially incoherent EDUs, increasing the number of dialogues generally leads to improved UAS scores, indicating that larger training samples positively affect structure prediction. Conversely, as the number of dialogues with incoherent EDUs (and thus erroneous rhetorical relations) increases, parser performance on LAS either declines or remains stable. This divergence indicates that, while structure prediction is relatively robust to certain levels of noise, relation classification is more sensitive to data fidelity.

Overall, these results suggest that expanding the number of dialogues can enhance parser performance on structure prediction, even under hallucination-like conditions introduced by low- and medium-quality samples. In contrast, substantial improvements in the full parsing task require higher-quality data that more effectively preserve rhetorical consistency between EDUs.

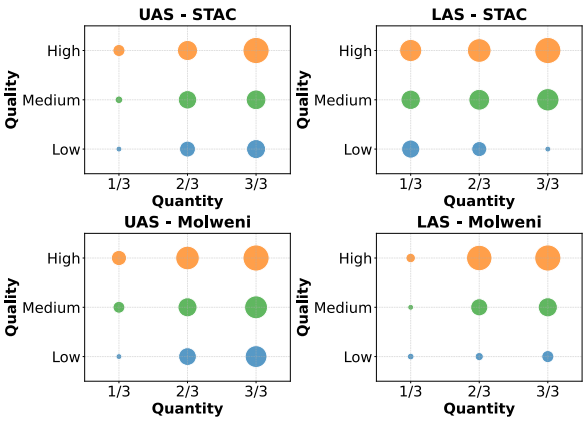


Figure 6: DAMT parser performance with varying quantities and qualities of augmented data. Circle size corresponds to UAS/LAS scores. Larger circles correspond to higher scores.

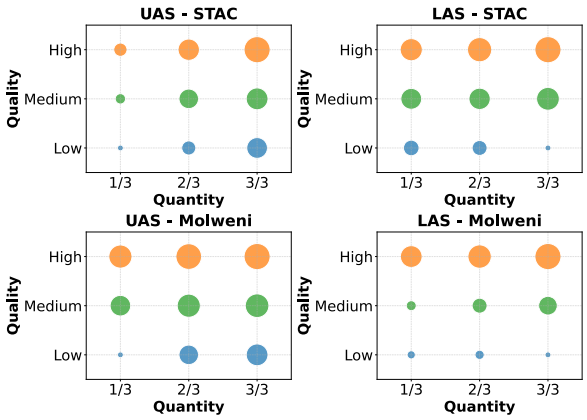


Figure 7: SDDP parser performance with varying quantities and qualities of augmented data. Circle size corresponds to UAS/LAS scores. Larger circles correspond to higher scores.

Parser	Approach	STAC		Molweni	
		UAS	*LAS*	*UAS*	*LAS*
SDDP	MIMIC^1	**74.80** $\pm$ 0.4	**58.31** $\pm$ 0.2	81.56 $\pm$ 0.2	56.54 $\pm$ 0.2
	MIMIC^2	74.45 $\pm$ 0.4	58.17 $\pm$ 0.5	**81.84** $\pm$ 0.3	**56.71** $\pm$ 0.4
	MIMIC^3	74.12 $\pm$ 0.3	57.89 $\pm$ 0.4	81.27 $\pm$ 0.3	56.40 $\pm$ 0.5
DAMT	MIMIC^1	**73.01** $\pm$ 0.2	**53.74** $\pm$ 0.3	79.70 $\pm$ 0.3	55.73 $\pm$ 0.3
	MIMIC^2	72.84 $\pm$ 0.3	53.38 $\pm$ 0.4	**79.85** $\pm$ 0.2	**56.06** $\pm$ 0.4
	MIMIC^3	72.76 $\pm$ 0.4	52.98 $\pm$ 0.3	78.99 $\pm$ 0.5	55.51 $\pm$ 0.4

Table 12: Impact of progressively expanded training data on discourse parsing performance (UAS/LAS) for SDDP and DAMT across STAC and Molweni.

H Sequential Dialogue Augmentation

In the previous experiments, we applied MIMIC to augment discourse corpora by doubling their size. However, the framework naturally extends to large-scale, iterative augmentation.

Let $\mathcal{C}_M^1$ denote the set of n synthetic dialogues generated by applying MIMIC to a corpus $\mathcal{C}$ of size n. This process replaces speakers in the original dialogues with speakers drawn from other dialogues in $\mathcal{C}$, implicitly creating new dialogue histories that simulate how these speakers might respond in different situations while retaining their stylistic tendencies. Using these generated histories, MIMIC can be reapplied to produce a new set of n synthetic dialogues $\mathcal{C}_M^2$, synthesizing utterances in $\mathcal{C}$ according to the stylistic distributions in $\mathcal{C}_M^1$. More generally, this supports an iterative augmentation chain, where each new set $\mathcal{C}_M^i$ is produced by applying MIMIC to $\mathcal{C}$ with substitute speakers from $\mathcal{C}_M^{i-1}$. This iterative design enables scalable generation of arbitrarily large synthetic corpora.

Table 12 presents the results achieved by the SDDP and DAMT parsers when trained on progressively augmented corpora. In this context, MIMIC^i indicates that the parser is trained on the original corpus augmented with the automatically generated corpora $\mathcal{C}_M^1, \ldots, \mathcal{C}_M^i$. Compared to training with MIMIC^1, the results show improved performance on Molweni when MIMIC^2 is considered, whereas performance on STAC declines. Instead, training with MIMIC^3 leads to a decrease in performance for both corpora. These results suggest that successive augmentations may introduce redundancy, leading to diminishing returns. In corpora with relatively homogeneous dialogue contexts, such as STAC, where interactions frequently revolve around repetitive selling and purchasing scenarios, additional synthetic examples plausibly tend to replicate existing conversational structures, producing diminishing returns. Conversely, in corpora with greater contextual diversity, such as Molweni, where dialogues cover varied Ubuntu-related issues, successive augmentations can potentially generate genuinely novel variations, improving parser performance. However, even in this case, an excessive enlargement of the training set may ultimately lead to redundancy, constraining the effectiveness of additional augmentation. These findings align with prior research suggesting that beyond a certain point, additional augmentation yields diminishing returns, as the new information may already be captured by earlier learning (Longpre et al., 2020), and substantially enlarging a corpus can sometimes even degrade performance (Okimura et al., 2022). Accordingly, while Appendix G shows that increasing dialogue count can improve parser performance, such gains likely depend on introducing meaningful variation; without it, the benefits are expected to be limited. Evaluating the diversity of newly generated samples relative to existing data could address this limitation, which we leave for future work.

I Impact of Training Data on Parsing

Training Data Volume The proposed MIMIC framework enables a substantial expansion of the original training corpora, as described in Appendix H. Moreover, applying MIMIC across domains (e.g., introducing Molweni-style speakers into the STAC corpus, or vice versa) facilitates the integration of additional training data, which may further enhance performance. It should be noted, however, that training time scales linearly with the amount of

augmented data, highlighting a practical trade-off between computational cost and parsing accuracy.

Increasing the number of substituted speakers in the training data has been shown to enhance discourse parsing performance. However, this improvement is most pronounced when the substituted speakers exhibit structurally rare discourse patterns. Extending substitution to more commonly represented speaker profiles may yield diminishing returns, as the added data contributes less to structural diversity.

Training Data Fidelity In the absence of the EDU-level refinement step, parser performance deteriorates due to the inclusion of noisy or rhetorically inconsistent utterances. This underscores the critical importance of preserving fine-grained rhetorical relations during the augmentation process. To exemplify this issue, consider the following excerpt from the STAC corpus (id *pilot01_2*):

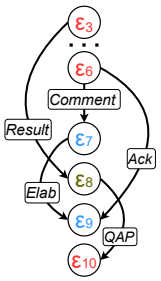
(ε_3) **A:** Don't give Tomm wheat
...
(ε_6) **A:** If he gets wheat

(ε_7) **B:** Aww :(

(ε_8) **C:** Can you give wheat to me, Dave?

(ε_9) **B:** Sadly, the man does have a point

(ε_{10}) **A:** Uh, I might get some on my turn, so maybe



In this example, where Elab stands for *Elaboration* and Ack for *Acknowledgement*, if MASK selects A for substitution, when rephrasing ε_{10}, the LLM receives $(\varepsilon_8, \varepsilon_{10}, QAP)$ as ε_{10} has only one incoming link and no outgoing ones. Under these conditions, Llama outputs "*Wait, do you mean you want to trade right now?*", which replaces "*Uh, I might get some on my turn, so maybe*". While structurally valid, it changes the rhetorical intent. The new utterance implies a *Clarification-Question*, not a *QAP* relation, breaking rhetorical consistency. By contrast, the EDU refinement yields "*Maybe wheat on my turn*", which better preserves the original rhetorical relation. As a result, the refinement module enhances the alignment of generated EDUs with their corresponding discourse relations, thereby improving communicative clarity.

J Examples of Augmented Utterances

Tables 13–15 present examples of utterances augmented using the MIMIC framework. In these examples, brown text represents portions of the dialogue history that the LLM (Llama) referenced when rephrasing the original utterance. Blue text indicates discourse relations present across the original dialogue and the dialogue histories.

In Table 13, speaker B's original utterance—"*You restart x and it works automatically*"—is rephrased as "*You may want to try restarting x, it should work automatically then*", using the stylistic profile of speaker D. The revised utterance maintains the original's explanatory function while adapting both tone and modality to reflect speaker D's communicative style. In particular, the expression "*You may want to try restarting x*" echoes the more tentative, advisory phrasing used by speaker D in "*You may want to try the main repositories*". Here, the direct imperative ("*You restart x*") is softened into a suggestion, aligning with speaker D's tendency to avoid assertive directives. Moreover, the phrase "*it should work automatically then*" introduces a degree of epistemic hedging through the modal *should*, contrasting with the original's more definitive "*it works automatically*". This nuance appears influenced by another utterance in speaker D's dialogue history—"*I would accept the install. Probably some minor error with the mirror you are using*"—which also exhibits a cautious tone via the inclusion of *probably*. Overall, this example illustrates MIMIC's effectiveness in replicating speaker D's stylistic tendencies, not only lexically but also in terms of pragmatic and epistemic framing.

Table 14 offers a further illustrative example, wherein the English expression "*You too*" is rephrased as its French counterpart "*À toi aussi*". This transformation likely stems from the fact that the substitute speaker, speaker C, had previously used the French phrase "*Pas de problèmes*" in their dialogue history. The occurrence of this non-English expression—presumably employed in a light-hearted or humorous tone—served as a stylistic signal prompting the model to adopt a similarly playful and multilingual register in the generated utterance. Importantly, the rephrased version preserves the original communicative function, while infusing it with the stylistic character of the substitute speaker. This example highlights MIMIC's sensitivity not only to syntactic and pragmatic cues but also to sparse and contextually subtle stylistic markers, such as multilingual expressions embedded in speaker histories.

Finally, Table 15 presents a case that underscores the relevance of modeling interpersonal dynamics in the rephrasing process. The original utter-

ance—"*No danger of me winning here*"—was associated with a speaker whose substitute, speaker C, lacked directly corresponding stylistic examples in their dialogue history. Nevertheless, the model produced a rephrased version—"*I'm not looking like a winner this time around*". This alternative draws upon the broader relational context of speaker C's dialogue history. Specifically, it echoes an earlier utterance by another speaker—"*Looks like it caught up*"—which was directed at speaker A within a Contrast relation. The rephrased version mirrors the syntactic structure and indirect tone of self-deprecation found in that earlier turn. This case exemplifies MIMIC's ability to transcend speaker-specific modeling by incorporating relational and interactional information, ensuring that the generated utterance remains socially and contextually grounded even in the absence of direct stylistic precedents.

Molweni	**ID:** *4412*
Original Dialogue	... **A**: *How do I turn on 3d in ubuntu when the driver is installed?* ... **B**: *You restart x and it works automatically* (`Original utterance`) ... **C**: *Press altf2 and run : compiz – replace*
Discourse Relations (Speaker B)	The utterance "*How do I turn on 3d in ubuntu when the driver is installed?*" by speaker **A** is in a QAP relation with the utterance "*You restart x and it works automatically*" by speaker **B** The utterance "*You restart x and it works automatically*" by speaker **B** is in an Explanation relation with the utterance "*Press altf2 and run : compiz – replace*" by speaker **C**
Discourse Relations (Speaker D)	The utterance "*You may want to try the main repositories*" by speaker **D** is in an Explanation relation with the utterance "*I changed it to main server, I hope it works.*" The utterance "*I would accept the install. Probably some minor error with the mirror you are using*" by speaker **D** is in an Explanation relation with the utterance "*I went to software sources and I have there multiverse enabled, if this helps.*" ...
Interpersonal Relationships	The utterance "*I have xp shares, typical shares, no server*" is in an Explanation relation with the utterance "*The system sharing the file is essentially a fileserver*" by speaker **C** ...
Augmented Dialogue	... **A**: *How do I turn on 3d in ubuntu when the driver is installed?* ... (`Augmented utterance`) **D**: *You may want to try restarting x, it should work automatically then* ... **C**: *Press altf2 and run : compiz – replace*

Table 13: Example of MIMIC-augmented utterance demonstrating style-aware rephrasing grounded in dialogue history. The original response "*You restart x and it works automatically*" is rewritten as "*You may want to try restarting x, it should work automatically then*", reflecting the indirect and advisory tone characteristic of the substitute speaker, whose dialogue history includes utterances such as "*You may want to try the main repositories*". Brown text highlights the retrieved utterances from the substitute speaker's dialogue history used to guide stylistic transfer. Blue text indicates discourse relations present across the original dialogue and the dialogue histories.

STAC	ID: *s1-league3-game6_12*
Original Dialogue	. . . **A**: *Good luck* **B**: *You too* (`Original utterance`) **B**: *:)*
Discourse Relations (Speaker B)	The utterance "*Good luck*" by speaker **A** is in a Continuation relation with the utterance "*You too*" by speaker **B** The utterance "*You too*" by speaker **B** is in a Comment relation with the utterance ":)" by speaker **B**
Discourse Relations (Speaker C)	The utterance "*Good luck all*" is in a Continuation relation with the utterance "*Good luck!*" by speaker **C** The utterance "*Pas de problèmes ;)*" by speaker **C** is in a Comment relation with the utterance "*My french is not so good*" by speaker **C**
Interpersonal Relationships	The utterance "*I'm innis,*" by speaker **A** is in a Continuation relation with the utterance "*He's niko*" The utterance "*Thanks for everything*" by speaker **A** is in a Continuation relation with the utterance "*See you!*" . . .
Augmented Dialogue	. . . **A**: *Good luck* **C**: *À toi aussi* (`Augmented utterance`) **C**: *:)*

Table 14: Illustration of stylistic transfer across languages within MIMIC augmentation. The original English utterance "*You too*" is rephrased as the French equivalent "*À toi aussi*", influenced by the substitute speaker's prior use of the French expression "*Pas de problèmes*". This example underscores MIMIC's sensitivity to multilingual cues and informal registers. Brown text denotes the relevant non-English utterance in the dialogue history, while blue text indicates discourse relations present across the original dialogue and the dialogue histories.

STAC	ID: *s2-league4-game3_15*
Original Dialogue	**A**: *Last time you won...* . . . **B**: *No danger of me winning here* (`Original utterance`) **A**: *Expect the unexpected..*
Discourse Relations (Speaker B)	The utterance "*Last time you won...*" by speaker **A** is in a Contrast relation with the utterance "*No danger of me winning here*" by speaker **B** The utterance "*No danger of me winning here*" by speaker **B** is in a Contrast relation with the utterance "*Expect the unexpected..*" by speaker **A**
Discourse Relations (Speaker C)	The utterance "*I was hoping to build another road last time to protect it*" is in a Contrast relation with the utterance "*You'll pop back up to 8 next turn*" by speaker **C** The utterance "*I need clay*" by speaker **C** is in a Contrast relation with the utterance "*I have sheep*" . . .
Interpersonal Relationships	The utterance "*Looks like it caught up*" is in a Contrast relation with the utterance "*Ok.. seems like it's working*" by speaker **A** . . .
Augmented Dialogue	**A**: *Last time you won...* . . . **C**: *I'm not looking like a winner this time around* (`Augmented utterance`) **A**: *Expect the unexpected..*

Table 15: Example showcasing the importance of interpersonal relationship modeling in MIMIC augmentation. The original utterance "*No danger of me winning here*" is rephrased as "*I'm not looking like a winner this time around*". Although the substitute speaker's own history lacked stylistically relevant content, the model successfully mimicked the interactional tone of prior Contrast relations involving speaker A, drawing on the utterance "*Looks like it caught up*". Brown text refers to the retrieved utterance from related interpersonal interactions, and blue text indicates discourse relations present across the original dialogue and the dialogue histories.