Comparing Native and Non-native English Speakers' Behaviors in Collaborative Writing through Visual Analytics

Yuexi Chen
Department of Computer Science
University of Maryland
College Park, Maryland, USA
ychen151@umd.edu

Yimin Xiao
College of Information
University of Maryland
College Park, Maryland, USA
yxiao@umd.edu

Kazi Tasnim Zinat
Department of Computer Science
University of Maryland, College Park
College Park, Maryland, USA
kzintas@umd.edu

Naomi Yamashita
NTT
Keihanna, Japan
naomiy@acm.org

Ge Gao
College of Information
University of Maryland
College Park, Maryland, USA
gegao@umd.edu

Zhicheng Liu
University of Maryland
College Park, Maryland, USA
leozcliu@umd.edu

Abstract

Understanding collaborative writing dynamics between native speakers (NS) and non-native speakers (NNS) is critical for enhancing collaboration quality and team inclusivity. In this paper, we partnered with communication researchers to develop visual analytics solutions for comparing NS and NNS behaviors in 162 writing sessions across 27 teams. The primary challenges in analyzing writing behaviors are data complexity and the uncertainties introduced by automated methods. In response, we present COALA, a novel visual analytics tool that improves model interpretability by displaying uncertainties in author clusters, generating behavior summaries using large language models, and visualizing writing-related actions at multiple granularities. We validated the effectiveness of COALA through user studies with domain experts (N=2+2) and researchers with relevant experience (N=8). We present the insights discovered by participants using COALA, suggest features for future AI-assisted collaborative writing tools, and discuss the broader implications for analyzing collaborative processes beyond writing.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; **Collaborative and social computing**; **Visualization**.

Keywords

Collaborative writing, non-native speakers, event sequence analysis, human-AI interaction, collaboration, large language models

ACM Reference Format:
Yuexi Chen, Yimin Xiao, Kazi Tasnim Zinat, Naomi Yamashita, Ge Gao, and Zhicheng Liu. 2025. Comparing Native and Non-native English Speakers' Behaviors in Collaborative Writing through Visual Analytics. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713693

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713693

1 Introduction

Non-native speakers (NNS) actively engage in collaborative writing across various contexts: international students write reports with classmates [13] and advisors [107]; Wikipedia contributors edit articles in different languages [30]; employees in multi-national corporations collaborate on project pages [4]. Previous research shows that when writing in a non-native language, NNS tend to produce shorter and less complex content compare to writing in their native language [14, 93, 101]. However, limited research has examined the collaborative writing processes involving NNS and how they write differently from native speakers (NS) [104]. Such knowledge will be insightful in enhancing collaboration outcomes and fostering inclusive team dynamics involving NNS.

To bridge this gap, in collaboration with communication researchers, we collected document history and screen recordings of 162 collaborative writing sessions from 27 teams. We are interested in *comparing* authors' behaviors across different linguistic backgrounds and writing stages. For example, what are the common behavior patterns of NS and NNS, respectively? How do they differ in the early and late stages of collaborative writing?

However, existing document visualization and analytics tools fall short of supporting our analysis goal. Visualizations of document versions have long been used to investigate the dynamics of collaborative writing. For example, History Flow [96] uses a Sankey diagram to encode content contributions from different Wikipedia authors across different versions; Time Curves [5] project different document versions on a 2D curve based on their similarity and temporal order. However, such visualizations focus on the document content instead of the writers' behaviors during the writing process, such as browsing the internet or using translation tools.

Event sequence visualization tools are better suited for behavioral data but still do not effectively meet our needs. For example, TipoVis [31] allows comparisons between two sequences at a time, while our analysis requires comparing behavior sequences across multiple authors and sessions. CoCo [60] supports comparing event sequences belonging to different cohorts, but focuses primarily on aggregated metrics such as frequencies. In our case, we need to identify behavioral differences between NS and NNS at multiple levels of granularity with meaningful qualitative descriptions.

Furthermore, to effectively analyze such complex behavioral datasets, it is important to combine visualizations with automated methods such as data mining and clustering. Previous research highlights challenges related to the interpretability and trustworthiness of automated methods [10, 15, 55, 106]. As communication research experts, our collaborators possess contextual knowledge about their dataset and multilingual communication in general but are not familiar with the mechanisms of automated methods. Before incorporating the model-generated results into their analysis, they must interpret and trust them.

In comparing the behaviors of NS and NNS in collaborative writing, we thus face two challenges: 1) the limitations of existing text and event sequence visual analytics approaches, and 2) the lack of interpretability and trust in automatic data analytics. To address these challenges, we worked closely with the communication experts to formulate data models and task requirements, iterated on visualization designs to assess their applicability and scalability, and identified factors that might hinder interpretation and trust. Based on these design iterations, we make the following contributions:

- COALA, a visual analytics tool for comparing native and non-native speakers' behaviors in collaborative writing. COALA displays the uncertainties of multiple clustering results and supports interactive refinement of clusters while leveraging large language models (LLM) to generate cluster summaries.
- Empirical validation of COALA through a focus group session (N=2+2) and individual study sessions with researchers in related fields (N=8), where the participants used COALA to analyze behavioral differences between NS and NNS.
- Design lessons for developing interpretable visual analytics in the context of communication research and findings that inform future AI-assisted collaborative writing tools and collaborative processes beyond writing.

2 Related Work

2.1 Collaborative writing studies

Collaborative writing has been a topic of interest since the 1980s [6, 21, 22]. Early research focused on awareness and coordination during collaborative writing [20], common writing tasks, and the number of collaborators [42]. The rise of online collaborative writing tools like Google Docs, Microsoft Word, and Overleaf [71] has made collaborative writing a common practice. For instance, Olson et al. [69] analyzed collaborative writing patterns in 96 college assignments in Google Docs and found that balanced participation and leadership would result in higher writing quality [69]. Researchers also explored various aspects of collaboration, such as impression management [7], reluctance to write closely [98], preference over edits with explanations [72], differences of tasks across writing stages [81], and territorial behaviors [50].

Few studies focus on authors' off-document writing-related activities during collaborative writing, for example, navigating multiple applications (e.g., Google Docs and Adobe InDesign) [51] and coordinating writing tasks on Wikipedia discussion pages [82]. Collaborated with communication researchers, we focus on writing-related behaviors, including off-document behaviors like using a translator and browsing the internet, aiming to provide a new perspective on collaborative writing analysis.

2.2 Text visual analytics

Several visual analytics approaches have been designed to analyze the evolution of documents in collaborative writing. Itero [92] is a revision history analytics tool based on Google Docs that visualizes character insertion patterns and user contributions. History Flow [96] and DocuViz [97] encode each author's contribution as a colored vertical line, with the height of the line proportional to the content length. The flow-like visualization reveals the cooperation and conflict among co-authors by connecting the same line across different versions. Graphs are also widely used, where authors are represented as nodes, and edges could be disagreement [23] or revert actions [45]. Time Curves [5] is a timeline visualization based on points' similarity, which could visualize different document versions. Other visualizations include branch-based visualizations [75], revision maps [87] and color-coded words by authorship [24, 91]. Compared to text visual analytics approaches, COALA focuses on sequences of writing-related behaviors.

2.3 Non-native speakers vs. native speakers

Compared to native speakers (NS), non-native speakers (NNS) usually produce shorter and less complicated content and have difficulty transferring writing strategies from their mother tongue [14, 93, 101]. Though NNS need more help in the expression aspects [84], they may still contribute to the ideation aspect [26]. Cheng et al. [13] found in a case study that NS students had more power in collaborative writing at the beginning, but the NNS student developed academic literacy along the way, and overall the group writing experience has improved. NNS' writing could also be improved by receiving direct edit feedback at early versions [39, 107], or exposure to well-written model text by NS [38]. Compared to these previous case studies, we have a larger collaborative writing dataset of NS-NNS with video-recorded author behaviors, poised to reveal more patterns beyond anecdotal evidence.

2.4 Event sequence analysis

There are numerous methods to analyze event sequences. Besides *visualizing* sequences, we categorize analysis methods based on tasks: *comparing*, *clustering* and *summarizing*.

Visualizing event sequences. The most straightforward visualization design for event sequences is to arrange the events on a timeline [47, 76]. When the number of sequences is large, flow-based visualizations could show the trend of bundled sequences. For example, Sankey diagrams represent each event as a node, the length of the node and the thickness of the links between nodes encode event frequencies [74, 102]. Tree-based visualizations encode the frequency of events as the thickness of edges [34, 56]. Like tree-based visualizations, icicle plots encode events as stacked rectangles, ordered from top to bottom, usually colored by event categories [57, 103]. When subsequences are highly repetitive, matrix-based visualizations could show the transition trend clearly [73, 111].

Comparing event sequences. Multiple tools focus on comparison. CoCo compares two patient cohorts via statistical analysis with built-in metrics [60] distilled from domain expertise [64]. TipoVis compares event sequences of social and communicative behaviors by overlaying two sequences [31]. COQUITO [46] assists

users in defining cohorts with temporal constraints and comparing sequences by overlapped branches. Directly linking event subsequences for comparison is also common [61, 78, 112].

Clustering event sequences. Several interactive tools are designed for clustering sequences. For example, Wang et al. [99] built an unsupervised interactive clustering system to analyze large-scale clickstream data. EventThread [28] clusters event sequences by latent stage categories. Gotz et al. [27] group event sequences by dynamic hierarchical dimension aggregation. Sequen-C [59] adopts an align-score-simplify strategy to cluster sequences. VASABI [68] clusters user profiles by topic modeling and uses multi-dimensional distributions to characterize each cluster.

Summarizing event sequences. Numerous methods have been developed to find an overview of a cluster of sequences. Sequence Synopsis [12] constructs a high-level overview of sequences by balancing the minimum description length (MDL) principle and the information loss. CoreFlow [56] extracts branching patterns in temporal event sequences. Frequence [74] is a visual analytics tool built on a frequent pattern mining algorithm that handles multiple levels of details and concurrency. SentenTree [34] summarizes unstructured social media text in a tree structure.

COALA is also equipped with visualizing, comparing, clustering, and summarizing features. Compared to existing approaches, we focus on interpretation and trust by including multiple clustering methods and displaying uncertainties, supporting multi-level granularity of sequences, and leveraging large language models to generate more intuitive descriptions [9].

3 Study Background

3.1 Background

We collaborated with two communication researchers from a public university in the US. One is a professor who has studied multilingual communication for more than a decade, and the other is a Ph.D. advisee of the professor who also has rich experience in multilingual communication. They collected a dataset of collaborative writing between native and non-native speakers and contacted us for suggestions in visual analytics.

We conducted longitudinal co-design sessions with our collaborators to understand the communication research analysis better. We met weekly or bi-weekly for 30 weeks. After they introduced the study background and data, we initially adapted existing text visualizations like Time Curves [5] and History Flow [96] (see Appendix). Though such text visualizations provide a glimpse of how authors contribute to the document, and how co-authors revise or delete each others' contribution, they do not address the research questions to compare authors' behaviors, so we designed dedicated features for analyzing authors' behavioral sequences. During the process, we showed them visualizations and interface mockups, incorporated the feedback into the next iteration, and finalized the visualization and interaction designs with them.

3.2 Data collection

Our collaborators recruited 29 native English speakers (NS) and 29 non-native English speakers (NNS) from an American university and a Japanese university for an online collaborative writing study.

Figure 1: Turn-taking of NS (native speaker) and NNS (non-native speaker) of English in the multilingual collaborative writing study.

To ensure NNS are of similar English proficiency, all NNS are native Japanese speakers with *limited working proficiency* in English [1]. Participants are asked to act as if they were columnists for an English magazine who answer readers' questions about the role of technology in modern life. The topics include social media, remote learning, and digital privacy. To mitigate the influence of topic familiarity, one NS and one NNS are paired to form a writing group and are assigned a topic familiar to both authors. All participants are provided with preset Google accounts and links to blank Google Docs. NNS and NS of the same group do not know each other but are informed about co-authors' language proficiency. Participants are also asked to record their screen during writing. Participants are welcome to use any tools (e.g., search engines, Google Translate, Grammarly) that they normally use during writing. Since the study was conducted before ChatGPT's release, no participant used generative AI tools.

Then, NNS and NS take turns writing an English essay jointly. The turn-taking setup is shown in Figure 1: first, each author writes independently (turn_0), ensuring they have actively thought about the task instead of being a free-rider. Then, NS review the write-ups of both authors from turn_0 and merges them into a single document (turn_1). During turn_1, NS can add/delete/edit any content. Next, NNS revise the joint document turn_2, followed by NS turn_3, and finally concluded by NNS turn_4. Participants are allowed up to 1 hour for each turn, and they could finish early if they have nothing more to contribute. After removing two teams that did not follow instructions, we have 27 remaining.

Our collaborators carefully designed the setup of the above experiment. First, though current online writing tools support simultaneous writing, turn-taking is still a popular workflow adopted by co-writers in practice, as they minimize the burden of syncing content mentally [8], protect authors' territoriality [50] and thus promote editing other co-authors' writing [3]. Second, NNS may face difficulties identifying opportunities for contribution in flexibly structured collaboration with NS. Previous research has introduced several techniques to interrupt the natural conversation flow and impose contribution opportunities of NNS, including artificial silent gaps [105] and a conversational agent [54]. Therefore, having a designated opportunity to ensure NNS contributes to collaborative writing is necessary. Besides, since it's one of the first studies on multilingual collaborative writing, researchers chose a simplified

setup with only one NS and one NNS co-writer in a team to pinpoint the group dynamics easily, leaving more complex configurations for future research (e.g., NNS from different countries, unbalanced group settings where there are more NS than NNS or vice versa).

4 Data Abstraction

Based on previous research, we introduce a general data model for the collaborative writing processes in our study.

Authors. In collaborative writing, there must be at least two authors. Let $A = \{a_1, a_2, ...\}$ be the set of authors. For each author a_i, the meta-information M_i is a set of the author's quantitative or qualitative attributes, e.g., author's ID, linguistic background, seniority and so on. In our study, there are two authors in a team, so $A = \{a_1, a_2\}$. Since we only care about the linguistic background of authors in this study, $M_1 = \{native\text{-}English\text{-}speaker\}$ and $M_2 = \{non\text{-}native\text{-}English\text{-}speaker\}$.

Events. Let E be the set of all collaborative writing events. The events could be on-document (E_{on}) and related ($E_{related}$), and $E = E_{on} \cup E_{related}$. Event $e_{on,i}$ is an edit made by authors, e.g., $e_{on,i}$ could include the author, editing types (add/delete), locations, and so on. Event $e_{related,i}$ is an event related to the collaborative writing process, but not limited to editing, e.g., leave a comment [109], make a post [82] and browse the internet. $e_{related,i}$ could include the event name, start and end time. We can automatically obtain E_{on} by comparing different document versions. To obtain $E_{related}$, one communication researcher manually coded events by watching the recordings, including the author, event type, and start and end time. After that, another communication researcher helped categorize events into higher levels. There are six types of events in total, and we highlight each event in different colors:

Writing: activities include "On Google Docs", "Using online editing tools", "Checking for creating citation".

Note-taking: activities include "Writing a note". The difference between "Writing" and "Note-taking" is that the note does not go into the writing, but serves as an auxiliary role.

Wordsmith-crosslingual: though the goal is to write an English essay, many NNS chose to write in Japanese and translate to English, or translate the write-up and read in Japanese. Such activities include "Using translators to read", "Using translators to write", "Searching for language-related information", "Checking a dictionary or thesaurus" and "Checking for meanings".

Wordsmith-English: some NS also seek help with expression, such activities include "Checking dictionary or thesaurus", "Searching for language-related information" and "Checking for meanings".

Active-search: authors may seek external evidence to build their argument. Activities include "Searching for online information".

Passive-search: After an author cites an external source in the collaborative write-up, the other author may check the content. Such activities include "Opening a URL to read information".

Table 1 shows summarized frequencies and duration of six high-level writing-related actions: "Writing" is the most frequent and time-consuming action, followed by "Wordsmith-crosslingual", "Active-search", "Passive-search", "Wordsmith-English" and "Note-taking".

Turns and stages. In collaborative writing, authors take turns to write, which could be either sequential or parallel [69], depending on whether the timestamps of events overlap. In our study, as shown in Figure 1, $t_{0,ns}$ and $t_{0,nns}$ are turns where authors write individually, and t_1 to t_4 are collaborative turns. These turns are organized into writing stages; therefore, we have individual stages $t_{0,ns}$ for NS and $t_{0,nns}$ for NNS, and collaborative stages $\{t_{1,ns}, t_{3,ns}\}$ for NS and $\{t_{2,nns}, t_{4,nns}\}$ for NNS.

Document versions. Let V be the entire history versions of a single document, $V = \{v_1, v_2, ...\}$. Document v_{i+1} is the result of event E_i on v_i. Google Docs record word-level document history; however, our collaborators are not interested in such fine-grained analysis. Instead, they collected the document version at the end of each turn for each author $V_{end} = \{v_{0,ns}, v_{0,nns}, v_1, v_2, v_3, v_4\}$, then sampled three intermediate versions that reflected a person's writing progress during turn 2 to 4:

$$V_{inter} = \{v_{i,j} \mid i \in \{2,3,4\},\ j \in \{1,2,3\}\}.$$

In total, we have 15 versions: $V = V_{end} \cup V_{inter}$.

Sequences. The events happening in different turns from an author are grouped into a sequence based on the writing stages. As shown in Figure 1, we have four collections of sequences: $S_{NS,individual}$ are events in $turn_0$ by NS; $S_{NNS,individual}$ are events in $turn_0$ by NNS; $S_{NS,collaborative}$ are events in $turn_1$ and $turn_3$ concatenated; $S_{NNS,collaborative}$ are events in $turn_2$ and $turn_4$ concatenated. We chose to concatenate events in collaborative turns as suggested by communication researchers, as based on their experience, behaviors are similar across turns in the collaborative stage.

5 Methods

Communication researchers are interested in comparing authors' writing behaviors along two dimensions: writing stages (individual, collaborative), and author types (NS, NNS). For example, to answer the question "during the collaborative writing stage, how do NS and NNS writers' behaviors differ?", we need to compare two collections of sequences: $S_{NS,collaborative}$ and $S_{NNS,collaborative}$.

A primary challenge is that the sequences are highly heterogeneous, e.g., sequences in $S_{NNS,collaborative}$ range from 6 to 188 events, with an average of 75.6 and standard deviation of 47.0, making direct visual comparison impossible, as shown in Figure 2.

Therefore, we identified the following lower-level tasks to enable effective comparison:

- **T1: cluster similar sequences within each collection.** Besides automatic clustering methods, communication researchers should be able to incorporate their domain expertise into the clustering results.
- **T2: summarize each sequence cluster within each collection.** The summarization should provide rationales to help communication researchers interpret clusters.

In this section, we discuss computational methods to support task **T1** and **T2**, the interpretation challenges we faced in the design study, and our strategies to solve those challenges.

5.1 Computational methods

For T1, we decided to first automatically cluster sequences. Clustering algorithms are widely used to organize similar data points into groups [40]. Common methods include k-means [32], hierarchical clustering [65], DBSCAN [83], self-organizing maps [100] and so

Action	Duration		Frequency	
	NS	NNS	NS	NNS
Writing	52.5h (84.9%)	39.3h (61.9%)	706 (57.5%)	1548 (50.2%)
Passive-search	0.7h (1.1%)	0.1h (0.2%)	47 (3.8%)	22 (0.7%)
Active-search	8.1h (13.1%)	5.1h (8.1%)	425 (34.6%)	233 (7.6%)
Wordsmith-English	0.5h (0.8%)	0.0h (0.0%)	45 (3.7%)	1 (0.0%)
Wordsmith-crosslingual	0.1h (0.1%)	18.7h (29.6%)	4 (0.3%)	1277 (41.4%)
Note-taking	0.0h (0.0%)	0.1h (0.2%)	0 (0.0%)	4 (0.1%)

Table 1: Duration and frequencies of writing-related actions from video recordings

Figure 2: All sequences of non-native speakers' (NNS) and native speakers's (NS) behaviors during the individual and collaborative writing stage. The sequences in the collaborative writing stage are longer due to the concatenation of turns. The length of the rectangles indicates the duration of each event, and the color encodes the event types.

on. For T2, we reviewed the literature on visual summarization and data mining in event sequence analytics [29, 113, 114], and decided to use frequent patterns to summarize each cluster.

5.1.1 Method I: cluster and summarize sequences jointly. Our first approach is to cluster and summarize event sequences jointly. We chose Sequence Synopsis [12] because it is reported to produce higher quality visual summaries compared to other summarization methods [114]. Besides, sequential patterns are the most widely used summarization format [12, 57, 74, 77], compared to trees [56] and directed acyclic graphs (DAG) [34]. Sequence Synopsis clusters sequences and constructs a sequential pattern of each cluster by striking a balance between pattern conciseness and minimizing information loss from the original sequences. In our implementation, we can indirectly control the number of clusters K and the length of patterns by adjusting the weights of information loss and the number of patterns.

5.1.2 Method II: cluster first, summarize later. Our alternative approach is to cluster sequences first, and summarize each cluster. Here, we considered k-means and hierarchical clustering because these algorithms allow us to easily change the number of clusters. For each pair of sequences, we computed the Levenshtein distance (ignoring duration), which returns the minimal number of edits required to align the two sequences. Compared to other common distance metrics such as Euclidean distance, Levenshtein distance captures the order of the sequence and handles sequences of different lengths. Given K clusters, we also prefer nested results. For example, if two sequences are in the same cluster when $K = 4$, then we expected them to still be in the same cluster when $K = 3$. Therefore, we chose hierarchical clustering over k-means [90]. Since hierarchical clustering algorithms do not generate visual summaries for each cluster, we ran a commonly used maximal pattern mining algorithm VMSP [25] to extract patterns for each cluster. We set the minimum support to be 50%, i.e., the pattern has to be present in at least half of the sequences in the cluster. Different from Sequence Synopsis, which returns only one pattern for each cluster, VMSP returns multiple patterns, and we chose the longest one with the maximum support as the representative pattern.

Method I and method II differ in two key aspects: 1) the distance metric: method I does not rely on an explicit distance metric but learns how to cluster similar sequences and extract a representative pattern by balancing the pattern length and the information loss. In contrast, method II uses Levenshtein distance, which computes the edit distance between two sequences required for alignment;

Figure 3: Six patterns returned by Sequence Synopsis [12] for native speakers during the collaborative writing stage. Each pattern is a sequence of circles, representing a visual overview of sequences belonging to the cluster. The original sequences of the currently selected pattern are displayed below. They are sequences of rectangles, with event duration encoded by length. Events matched to the pattern are outlined in black.

2) the connection between patterns and clusters: for method I, the patterns and clusters are tightly coupled, each cluster is represented by a single pattern; for method II, clusters and patterns are loosely coupled, as pattern mining algorithms return multiple patterns for a cluster, it offers greater flexibility. By including both methods, COALA explores the computational space more thoroughly, provides multiple approaches to the clustering and summarization tasks, and thus brings different perspectives to the datasets.

5.2 Challenges in interpreting the computational results

We then designed an initial visualization to show the clustering and pattern mining results from these methods. Fig. 3 shows the result produced by Sequence Synopsis for NS' behaviors in the collaborative writing stage. There are six clusters, and for each cluster, Sequence Synopsis returns a representative pattern depicted as a sequence of circles. The number before each pattern is the cluster size, and the authors' IDs in the cluster are displayed after each pattern. To show the sequences in the cluster, users can click the radio button before each pattern (currently the 4th pattern is selected). In the raw event sequences displayed below the patterns, each event is a rectangle, where the color denotes the event type, and the length is the duration of the event. We highlight the event rectangles in black outlines if they are reflected in the pattern. We also implemented a similar interface for method II.

Visual summary. Though the patterns returned by Sequence Synopsis preserve the ordering of events in the sequences, the communication researchers expressed difficulty in interpreting such patterns and wonder whether it's possible to start with something more intuitive, for example, they suggest starting with one author's sequence and building clusters based on similar authors. Besides, it was unclear how each pattern differed from one another and how the algorithm performed the clustering. Take the sequences in Figure 3 for an example, the second and third cluster both feature a subsequence of Writing - Passive Search. However, it is unclear how these two clusters differ from each other. The same issue can be found in the first and the last clusters as well, where both clusters exhibit repetitive Writing - Active-Search subsequences.

Cluster membership. Since we have two sets of clustering results returned by Sequence Synopsis and hierarchical clustering, we let communication researchers switch to different results via a radio button. However, this caused great confusion. Even if we keep the number of clusters K the same for both methods, the clustering results returned by hierarchical clustering and Sequence Synopsis are not the same, and communication researchers are not sure which one to trust more. Besides, there are no explanations for why the sequences are grouped into each cluster or the differences between the clusters.

To summarize, the communication researchers encountered the following interpretation challenges:

- **C1:** two clustering methods' output differ.
- **C2:** lack of explanations of why the sequences are assigned to a particular cluster.
- **C3:** the extracted patterns are not easily interpretable.

5.3 Strategies to address the interpretation challenges

To address the above challenges, we devised several strategies, including ensemble and interactive clustering (**C1**), visualizing sequence-level information for clustering rationales (**C2**), and using large language models (LLM) to generate text summaries (**C3**).

5.3.1 Support ensemble and interactive clustering (C1). To address **C1**, we decided to show the consensus and discrepancies in the results produced by different methods, inspired by the idea behind ensemble clustering [95].

We obtained multiple clustering results with varying numbers of clusters, ranging from 2 to N, where N represents the maximum number of patterns identified by Sequence Synopsis. We use Sequence Synopsis as a constraint because hierarchical clustering can produce as many clusters as the data points. Then we evaluate the clustering results of both Sequence Synopsis and hierarchical clustering using the same number of clusters. Given two clustering results A and B, and a cluster number K ($2 <= K <= N$), we try to match each cluster in A to a cluster in B in a way that maximizes the total overlap between cluster members. We used the Hungarian algorithm [48] to obtain the assignments with the most overlap. Then, we only keep assignments when the intersection size is larger than 1 (it is trivial to have a cluster of only one sequence). Therefore, the size of consensus clusters is usually smaller than K. For unclustered sequences, we treat them as singletons.

Figure 4 shows a revised version of the visualization, where the sequences enclosed within a rectangle box have cluster assignments confirmed by both SequenceSynopsis and hierarchical clustering methods. In contrast, sequences without a consensus (i.e., NNS-2, NNS-3) are not enclosed, indicating they are singletons. Singletons are expected since the two methods leverage different computational techniques. Singletons remind communication researchers to give additional consideration to these authors, as computational methods differ in their cluster memberships. To help users assign

Figure 4: Consensus of clusters: sequences assigned to the same cluster by both Sequence Synopsis and hierarchical clustering are in the box; other sequences are outside of the box. Users can also change the number of clusters by dragging the slider, or manually add new clusters.

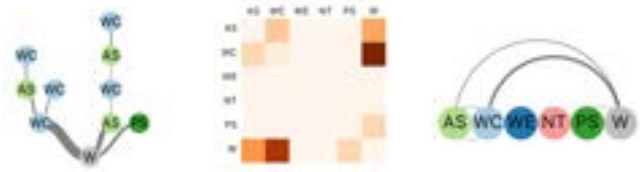

Figure 5: Design variants: tree, transition matrix, and arc diagram (final design) to visualize sequence information for author NNS-15.

singletons to clusters and revise existing cluster memberships, we provide a slider to adjust the number of clusters and support manually rearranging authors across clusters via drag-and-drop. Furthermore, to assist analysts in finding similar authors based on an author of interest, we also implemented a recommendation feature: we calculated the similarity between two sequences by summing over their sequence-level Levenshtein distance with the Levenshtein distance between their Sequence Synopsis [12] cluster patterns, and recommended top five to users.

5.3.2 Visualize sequence-level information for clustering rationales (C2). To help users understand why sequences are clustered (**C2**) and assess the similarities and differences between sequences within and across clusters, we devised two solutions: one focuses on local information of individual sequences, while the other focuses on the global context:

- **Local information:** for individual sequences, we improve the visualization design to support users in comparing pairs of sequences visually.
- **Global context:** for all sequences, we reveal their pairwise distances to support users in understanding the overall distribution of sequences in terms of pairwise similarity.

Visualizing local information of individual sequences. As shown in Figure 4, our earlier visualization design of an individual sequence presents the event sequences as it is, where colors encode the event types and the rectangle length encodes the duration. Such a design preserves all the information in the raw data, and communication researchers quickly conclude that NNS usually exhibit much more fragmented workflows than NS, frequently alternating between writing and other events. In contrast, NS usually allocates large chunks of uninterrupted time dedicated to writing. However, it is hard to generate additional insights. Therefore, we considered several design variants for visualizing individual sequences: trees, transition matrices, and arc diagrams.

- *Variant II: tree.* As communication researchers are interested in comparing NS and NNS' events before Writing, we extracted all unique subsequences ending with Writing. As depicted in Fig. 5, each tree node represents an event, and each edge denotes a transition, with the edge's thickness reflecting the frequency. While communication researchers appreciated the completeness of the tree visualization, they noted some redundancies, e.g., excessively extended tree branches like "WC - AS - WC - AS".
- *Variant III: transition matrix.* To mitigate the issue of redundant sequences, we adopted a transition matrix, where each row/column is an event, and each transition starts from the row and ends at the column. The cell's darkness signifies the normalized transition frequency. As illustrated in Fig. 5, it's easy to identify common event pairs, e.g., from Writing (W) to Active-Search (AS) and vice versa. However, the matrix is sparse as frequent transitions concentrate on a few cells, resulting in underutilized space.
- *Variant IV: arc diagram.* For the final design, we chose an arc diagram. We borrowed the event node design from the tree visualization, but instead of arranging them in a tree structure, we put all the nodes on the same row and connected them with arcs of different thicknesses, indicating the transition frequencies. The communication researchers like the simplicity and compactness of the design, as it is easy to spot which events precede Writing. Besides, unlike the transition matrix, where it is challenging to eliminate empty cells, we can easily conceal unwanted arcs. For example, we removed outgoing edges from Writing, as we are interested in events that precede each writing action instead of after.

Global pairwise distance of sequences. After improving the visualization for individual sequences, we helped users compare sequences globally via pairwise distances. For hierarchical clustering, we computed the Levenshtein distance between sequences and normalized it by the sequence length. Since Sequence Synopsis does not return distance scores, we used the Levenshtein distance between patterns as a proxy. Therefore, if sequences belong to the same cluster according to Sequence Synopsis, their distance is 0. We plot it on a 2D scatterplot, with each sequence represented as a dot, and the sequence of interest is placed at the origin.

Figure 6: The clustering panel of COALA. ① shows consensus clusters of authors returned by Sequence Synopsis [12] and hierarchical clustering; ②: unclustered sequences. When users select an author, its background turns orange, and recommended authors are highlighted in orange outlines. On the right, a 2D scatterplot (③) shows the current author's behavior distance between other authors. ④ shows detailed sequence information.

5.3.3 Use LLMs to generate text summaries (C3). To address **C3** (extracted patterns are not interpretable), we drew inspiration from the analysis process of communication researchers. We observed that they described clusters in natural language; for example, they described the visualization in Figure 4 as "*The first cluster mostly shows writing and wordsmith-crosslingual, so the NNS participants here spent most of their time figuring out how to produce writing in English through Japanese. The second cluster features more active search during the writing session even though participants still performed crosslingual language editing at the beginning and end of their sessions. These NNS participants followed a different writing path - they segmented the content (active information search in the middle) and the editing (wordsmith at the beginning and end) aspects of the writing and focused on one task at a time.*"

Recently, large language models (LLM) have shown great promise in data analysis [9, 58, 70] and LLM is found to have a high degree of agreement with human coders in thematic analysis [17]. Therefore, we leveraged LLM to generate text descriptions. We first tried capturing a screenshot of each cluster of arc diagrams (Figure 5) and prompted GPT-4V [2] with an explanation of the color encodings to describe each cluster. For example, we used the following prompt: "*The figure contains several event sequences, each showing an author's writing-related behaviors. There are six types of events: Active-Search, Wordsmith-Crosslingual, Wordsmith-English, Note-Taking, Passive-Search, and Writing. Each colored node is an event type, and you can find the event type in the colored legend. The arc thickness is the transition frequency of two events. Please name this cluster and provide a brief description.*"

However, GPT-4V sometimes ignored faint arcs between two nodes. To ensure GPT captures all transitions, including rare ones, we provided the transition data in JSON format, which was used to generate the arc diagram. Each entry contains the source event, the destination event, and the normalized frequencies. For example, source: Wordsmith-Crosslingual, destination: Active-Search, frequency: 0.25. Communication researchers found explanations returned by GPT fascinating and intuitive, and adopted some descriptions in their analysis. For example, GPT-4V calls one cluster "versatile writers" and explains that the cluster balances events between Wordsmith-Crosslingual and Active-Search, which suggests writers are comfortable with both actions without a dominant one.

Our prompting strategy to generate clusters could be generalized for other sequence datasets, and could be a plug-in for many existing visual analytics systems. We also found in later user study sessions that users borrow words from LLM-generated summaries during the analysis, especially for users new to the dataset.

6 COALA

Integrating all these strategies, we built COALA, a visual analytics tool to **co**mpa**r**e co**l**l**a**borative writing behaviors of native and non-native English authors. It has two tabs: one for inspecting and modifying the clustering results (Figure 6) and the other for comparing clustering results for authors of interest (Figure 7).

The first tab (Figure 6) supports refining sequence clusters and summaries. After selecting authors and writing stages in the dropdown menu, it displays consensus clusters of sequences by Sequence Synopsis and hierarchical clustering (①) and unclustered sequences (②). Users can drag the slider to change the number of clusters, add or delete clusters, revise cluster descriptions, and drag and drop arc diagrams directly across clusters. To facilitate the refinement process, when users select an author's arc diagram, its background turns orange, and recommended similar authors outside its cluster are highlighted in orange outlines. Currently, an author in the second cluster (orange background) is selected, and several authors outside the cluster are recommended. On the right,

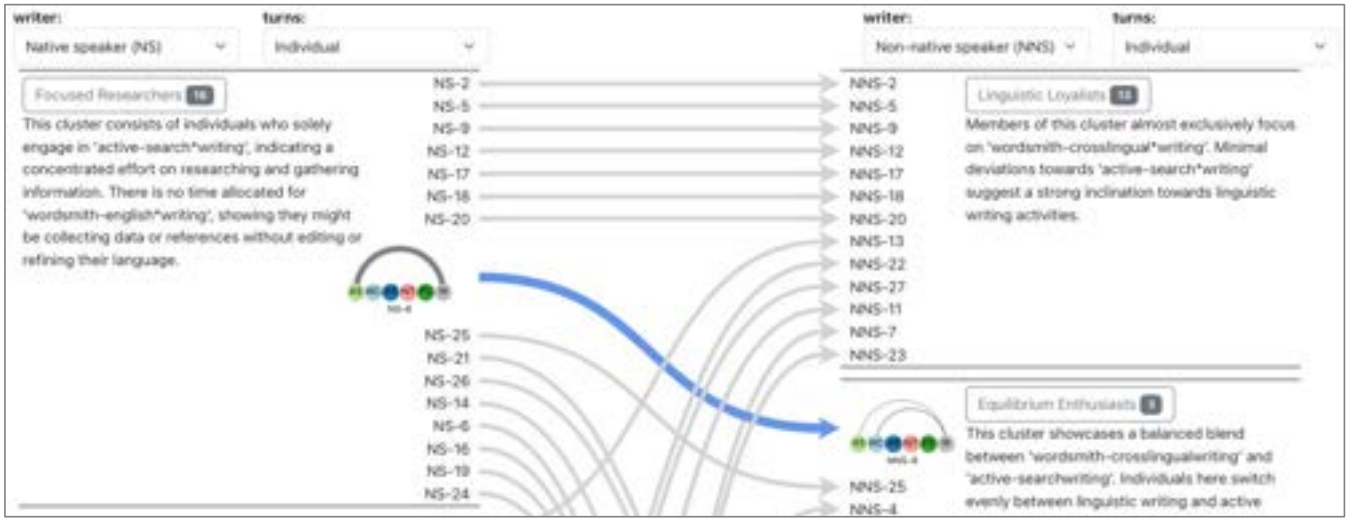

Figure 7: The comparison panel of COALA. It shows a two-panel layout: each has its dropdown menu for users to configure the author and writing stages. An arrow connects authors from the same team on both panels. Users can click an arrow to directly compare the two sides' arc diagrams.

a 2D scatterplot (③) shows the selected author's sequence distance between other authors according to Sequence Synopsis (method 1) and hierarchical clustering (method 2), and users can hover over each dot to see the sequence ID and distance. Users can also hover over ④ to inspect the event types and durations.

The second tab (Figure 7) supports directly comparing collections of sequences and individual sequences. It shows a two-panel layout; each panel has its dropdown menu for users to select collections of sequences. Currently, we select NS-individual on the left panel and NNS-individual on the right panel. An arrow connects authors of the same team. Authors are organized by clusters in Figure 6 and sorted by minimizing edge crossings. Once the user clicks an arrow, arc diagrams will appear on both panels. For example, now it shows that NS-8 transitions frequently between Active-Search and Writing, while NNS-8 has a balanced Active-Search and Wordsmith-Crosslingual activities before Writing, highlighting NNS' effort in information gathering and also translation. For cluster-level comparisons, since the arrows are sorted, users can observe the flow directions, similar to a Sankey diagram.

7 Validation: User Studies

The validity of our design is rooted in the user-centered design process reported in the previous sections. We also organized two user studies to evaluate COALA. The first, a focus group session with two existing and two new communication researchers, examined the usage of COALA in real-world setting; the second, individual sessions with 8 graduate students, evaluates the effectiveness of COALA for general users who work more independently.

7.1 User study setup

7.1.1 Focus group (N=2+2). We organized a focus group session with the two existing communication researchers along with two of their colleagues. All four participants belong to the same research group. The two new researchers are familiar with the multilingual communication research but are unfamiliar with the dataset details, nor have they seen COALA before. Though our collaborators are informed about the methods and individual visualizations, they also have not used COALA. Therefore, we conducted a 1-hour study session with all communication researchers (N=2+2). The goal of the group study is to use COALA in a realistic work setting to make sense of a dataset. In the beginning, we introduced the dataset and analytic tasks (**T1** and **T2**), and then we gave a tutorial on how to use the tool. We deployed COALA online, and each researcher accessed the tool on their laptop. They were encouraged to think aloud and discuss their findings during the session. We took notes during the session, and after the group session, we also invited them to write down their findings in a shared document.

7.1.2 Individual user studies (N=8). Besides the focus group, we also expanded the evaluation scope by recruiting eight graduate students with related research backgrounds (communication: 2, second-language education: 2, CSCW: 2, political science: 1, natural language processing: 1). All participants have first-hand collaborative writing experiences and are interested in understanding collaborative writing behaviors. Among them, four are NS, and four are NNS. To better evaluate the usability and design effectiveness of COALA, we set up individual user study sessions with them.

The procedure is similar: first, participants read a background introduction of the dataset (we prepared a shortened and simplified version of section 3 and section 4). We then demonstrated how to use COALA to cluster similar authors and compare their writing behaviors. Participants were asked to complete two tasks: analyzing authors' behaviors across two dimensions—writing stages (individual vs. collaborative) and author types (NS vs. NNS). Participants can either write their findings in COALA or describe them verbally. Participants are given one hour to complete the tasks. Participants were asked to think aloud during the study, and after they

completed the tasks, we conducted a brief interview to ask them questions about the effectiveness of COALA and the quality of model-generated results. We chose not to collect the ratings quantitatively due to the limited sample size and the lack of a baseline tool for comparison. Instead, we focus on gathering more qualitative feedback. All except one participant completed the tasks on time, and every participant was paid with a $20 Amazon gift card. The study is IRB-approved.

7.2 Effectiveness of COALA

7.2.1 Focus group (N=2+2). COALA facilitate discussions. All expert users had no difficulty in using COALA, and they were surprised to see that they focused on different aspects of the dataset during the analysis. For instance, one focused on the durations in the event sequence, and the other focused on the event transitions. This divergence triggered experts' discussion of why one aspect of collaborative writing was important. Ultimately, they concluded that COALA offered multiple angles to analyze the dataset they would otherwise miss due to their intuitions.

7.2.2 Individual user studies (N=8). COALA helps uncover insights. Among the seven participants who completed the study, three found the arc diagram design more effective, two preferred the sequence diagram, and two considered both effective. Participants prefer the arc diagram mostly because it summarizes the event sequences. For example, P8 appreciated how the arc diagram *"compresses information"*. In contrast, participants preferred the sequence diagram because it encoded the duration information (P5).

Regarding the model quality, five out of seven participants considered the initial clusters obtained by the consensus methods to be high quality and encouraged unbiased exploration. For example, P7 said, *"It's a really good starting point, because if you were to analyze this by yourself without any ideas...you can fall into typical stereotypes that...an AI model would not."* Furthermore, six participants found the recommendation feature helpful, using it as a starting point for the analysis (P8), a voting mechanism (P6), and a verification method (P7).

Participants also made suggestions to improve model-generated results, including more details about how the recommendation methods work (P4), more fine-grained clusters to start with (P5), and a 3rd recommendation method based on time duration (P6).

Notably, none of the participants had analyzed multilingual collaborative writing datasets before. Despite their unfamiliarity with the dataset and COALA, most of them were able to familiarize themselves with COALA and complete the analysis tasks. By leveraging the model-generated results and data visualization designs, they uncovered insights similar to those of expert users and connected the findings with their own experience. This outcome highlights the general usability and design effectiveness of COALA.

7.3 Findings: collaborative writing patterns

Here we aggregated participants' findings by using COALA, including the dataset itself and their reflections on their own collaborative writing experience.

7.3.1 Individual stage patterns. NS and NNS are asked to write independently before collaborating in this experiment to avoid free-riders. Communication researchers observed distinct behaviors during the individual stage.

NS research extensively. Using our tool, communication researchers easily identified several major clusters of NS. One cluster is characterized by dominant Active-Search - Writing behaviors, indicating they actively incorporated external information into their writing. Another cluster has more activities, besides AS - W, they also engaged in extensive paraphrasing activities: Wordsmith-English - Writing. Besides these two major clusters, there are also uncommon behaviors, e.g., one NS only displayed WE - W behavior without relying on external information; one NS went even further, solely engaged in writing, degenerating the arc diagram into a stick on the W node.

NNS encounter costly bilingual content production. Similarly, communication researchers identified several clusters of NNS. In one cluster, Wordsmith-Crosslingual - Writing dwarfed any other behaviors, implying NNS had significant usage of their native language to produce writing in English. Another cluster has more Active-Search - Writing behaviors, showing NNS also incorporated external information into their writing. Different from NS, several NNS also displayed AS - WC, meaning NNS also used their native language to transfer content from their content-related search to their writing output. Among the NNS who have engaged in both WC - W and AS - W, half are dominated by WC, and the rest are dominated by AS or are balanced. Notably, none of the NNS engaged with Wordsmith-English - Writing during the individual writing stage, highlighting the bilingual nature of NNS' writing process, as their limited English proficiency may have dissuaded them from writing directly in English. Compared to NS, many more NNS (n=10) only engaged in WC - W without relying on external information. This pattern hinted that NNS have dedicated much of their time to the costly process of writing in English - generating content in their native language first and then translating the content to English, either by themselves or by leveraging support from cross-lingual dictionaries and translation tools.

7.3.2 Collaborative stage patterns. Communication researchers found more diverse actions during collaborative stages, driven by co-authors' need to exchange information, refine text, and communicate with each other. For example, communication researchers observed multiple NS (n=8) and NNS (n=10) engaged in Passive-Search, indicating active information sharing. NS and NNS also display different distributions of behaviors in the collaborative stage.

NS shoulder more editing responsibilities. Transitions between AS - W and WE - W still have a strong presence, though the balance has shifted, with WE - W increases, and AS - W decreases. The shift towards editing indicates that NS is responsible for editing both parties' English expressions.

NNS gain more bandwidth for other tasks. Communication researchers witnessed an uptick in both WC - W and AS - W, suggesting NNS actively interpreting and incorporating NS-contributed content. In addition, more than a third of NNS engaged in PS - W.

These observations suggested that when co-writers completed initial drafts and transitioned to a collaborative editing stage, NNS were partially relieved from the demanding task of producing English content. Thus, they had more bandwidth to perform other tasks, such as AS to enrich the joint writing further.

7.3.3 Echo with first-hand experience. In the individual user study sessions, besides the findings similar to the ones in the group study session, our participants also compared their findings through the lens of their collaborative writing experience. For example, P6, a native English speaker, noted that too many NS *"just spitball on the fly without actually doing any research"* in this dataset, which aligned with his previous collaborative writing experience. In contrast, P2 (a non-native English speaker) observed that *"...native speakers become more cautious with their writing...when they are working with people, which is something that I wouldn't expect...if I am working with a non-native speaker of my language. I feel like I know more, so I would feel less need to actually check my writing."* P4 resonated with the Active-Search - Wordsmith-Crosslingual transition in the arc diagram: *"I was seeing myself do the same thing. I also like search for specific words in Korean, and then translated into English."*

Some participants also voiced their suggestions for best practices in collaborative writing. For example, P3, a native English speaker, pointed out that so much time spent on Wordsmith-Crosslingual is *"a waste of manpower, it makes more sense to focus on the ideation and don't care if it comes out ugly if the other person (NS) can fix it without having to do a lot of wordsmithing."*

7.4 Findings: participants' analysis strategies

We also studied participants' analysis strategies. For the group study, we analyzed the notes taken during the study and the document on which participants wrote findings. For the individual study, we analyzed the video recordings, including participants' screens and meeting transcripts.

7.4.1 Experts (N=2+2). COALA is pre-populated with clusters returned by methods described in section 5. Initially, all communication researchers used those clusters to gain an overview of the dataset. They read the descriptions generated by GPT-4 and gradually started to refine clusters based on their understanding.

For example, one researcher (E1) merged clusters based on the role of external resources in the writing processes: authors who use **Wordsmith-English** or **Wordsmith-Crosslingual** are classified as *"grammar-based"*, as they only need help with the expression, but not ideation; authors who leverage Active-Search and Passive-Search are classified as *"research-heavy"*, as such authors are still in the process of finding ideation.

On the opposite, another researcher (E2) broke down clusters into sub-clusters by examining the duration and transition frequencies of raw sequences carefully (Figure 6.④); for example, NS-20 and NS-22 were initially clustered into the same cluster by our algorithms as their behaviors are dominated by Active-Search, however, after examining the raw sequences, E2 found that NS-20 usually spent a long time on Active-Search before transitioning to Writing, different from NS-22, who rapidly transitioned between those two. Therefore, E2 created a new cluster called *"in the flow"*, and placed authors with less fragmented workflow into this cluster.

During the refinement, they also use the recommendations (orange border) to guide them in finding similar authors. After refinement, they moved to the comparison panel (Figure 7) and examined individual authors' behavior changes.

7.4.2 General users (N=8). Similar to focus group participants, participants in individual sessions also started by understanding the pre-populated clusters and text descriptions.

Clustering strategies. We found participants in individual sessions follow similar high-level clustering strategies of E1 and E2. Similar to E1, four participants also clustered authors by grouping multiple activities. For instance, P3 and P5 assigned authors performing more than two activities to a *"multi-tasking"* cluster; P3 assigned authors with AS or PS to the same cluster, as such behaviors implied that the authors incorporated external information.

Similar to E2, three participants refined the clusters with more fine-grained criteria. For example, P4 created two clusters for authors exhibiting dominant AS - W behaviors and labeled them as *"comfortable/uncomfortable with searching information in English"* based on the absence of AS - WC. P7 and P8 broke down clusters by considering the events' time durations, for example, P7 differentiated authors with WC - W behaviors by the event durations and labeled authors spent significant time on WC as *"long-term crosslingual enthusiasts"*. Similarly, P8 created several clusters with names like *"high/low active-search"* and *"strong/weak passive-search"* based on the durations.

Participants also interpreted the same behavior differently. For instance, for authors who wrote extensively without engaging in other activities, participants described them as *"have lots of knowledge and can pull citations from their memory"* (P4), *"stick to minimal approaches"* (P5), *"YOLO"* (P6) or *"confident"* (P7).

Reliance on the recommendation feature. Compared to E1 and E2, participants in the individual sessions rely more on the recommendation feature to discover similar authors. Some participants used it as a starting point to narrow the candidate pool; others used it to verify their assumptions. For example, P7 identified similar authors visually and then clicked the author of interest to see whether the recommendation confirmed her assumption. P6 employed the recommendation feature as a voting mechanism: when adding a new author to an existing group, P6 cross-checked the recommendation of multiple existing authors and only picked those endorsed by multiple existing cluster members.

Similarly, after refining clusters, participants moved to the comparison panel to get an overview of the authors' behavior changes and clicked authors of interest for more fine-grained comparison.

8 Discussion

In this section, we reflect upon the lessons learned from understanding collaborative writing processes through visual analytics and the implications for AI-assisted collaborative writing tools.

8.1 Reflections on developing visual analytics tools

Information-seeking and visual analytics mantras have limitations in guiding tool design. When we started working with the communication researchers, we aimed to create an overview of authors' behaviors, following the visual information-seeking mantra "overview first, zoom and filter, then details-on-demand" [85]. Therefore, we chose Sequence Synopsis [12], which clusters sequences and generates an overview pattern for each cluster. However, as discussed in section 5.2, communication researchers found such overviews confusing. Instead, they asked whether it was possible to start with one author's sequence and manually build clusters based on similar authors. This request directly contradicted the mantra, and we realized that the introduction of computational models resulted in interpretation challenges, and an overview would only be meaningful if users could reliably interpret it. Since it is tedious to inspect and cluster individual sequences manually, we kept the automatic clustering results but addressed the interpretation challenges with ensemble and interactive clustering and large language model (LLM)-generated text summaries. With the advances of AI, we expect more complex computational models to be incorporated into analysis. Although the visual analytics mantra ("Analyze first, show the important, zoom/filter, analyze further, details on demand") proposed by Keim et al. [41] tries to address the importance of analysis, what constitutes "the important" is highly dependent on the computational models and the tasks involved. Therefore, it is critical to emphasize the interpretability of model results and build a shared representation between analysts and models [33].

Visualizing ensemble clustering methods requires consolidating results. Clustering simplifies the analysis by grouping similar data points, but the variety of clustering methods and clustering results raised doubts among our collaborators about the output reliability. At first, we simply used radio buttons to toggle between methods, inadvertently presenting each method as interchangeable black boxes. However, having an agency over black boxes does not gain users' trust. Therefore, we visualized the uncertainty explicitly using a bounding box: sequences assigned to the same cluster by different methods were clustered, while discrepancies resulted in singletons. Besides, we also visualized the distance between different authors in a scatterplot to allow communication researchers to explore the similarity space intuitively. Our lessons show that mere juxtaposition is not enough when visualizing multiple models' results; instead, we should support analysts in interpreting multiple models' results without model-specific knowledge.

Caution needs to be exercised when updating model results based on user feedback. Our communication researchers think the automatically generated clusters are a useful starting point, yet still require revisions. Therefore, we implemented several user interactions, such as adjusting the number of clusters and rearranging cluster members. To minimize manual operations, we further suggested updating the recommendation interactively, e.g., if an author was moved from one cluster to another, then both clusters' descriptions would change, and the distance function would be revised. However, our collaborators found this distracting, preferring a more stable interface and manual revision of AI-generated results. Though incorporating users' feedback can improve the quality of model results [16], it also requires more thoughtful inputs from users and imposes extra cognitive load to examine the updated interface [86, 88]. Therefore, future tools should carefully balance the benefits of refined models and the additional burden on users.

8.2 Design implications for AI-assisted collaborative writing tools

With the recent advances of large language models (LLM), researchers have proposed multiple AI-assisted writing tools [18, 19, 43, 49, 52, 53, 79, 89, 110], some are tailored for non-native speakers [11, 37, 44]. However, we have not yet seen AI-assisted writing tools targeting collaborative writing, potentially due to the lack of understanding of the dynamics of collaborative writing. Here we propose several potential features derived from our study.

Detect diverse contribution types. Traditional collaborative writing tools [96, 97] for tracking authors' contributions in a collaborative document usually rely on word count. In our initial exploration (Appendix), we also found that NS contributed more text, and our collaborators mentioned that sometimes NS complained that NNS wrote too little. However, according to the communication researchers, there are diverse contribution types that are not necessarily manifested in word counts. Tools that quantify multi-dimensional contributions will help teammates understand each other's contributions and foster team dynamics. For example, NNS could contribute to the overall writing by providing an idea, which was expanded and refined by NS. Other contributions, such as scaffolding and improving the document flow by reorganizing, are also equally valuable. Automatically detecting such contributions requires tracking and deeply understanding the document, which may be potentially feasible by leveraging LLM.

Reduce context switching for NNS. NS and NNS spent about the same time on this lab study. However, the way they spent their time is notably different. NNS spent much more time transitioning between Wordsmith-Crosslingual and Writing, resulting in a more fragmented workflow. In contrast, NS usually dedicated extended time to writing, engaging in focused and uninterrupted work, as highlighted by Newport's concept of "deep work" [67]. To improve efficiency for NNS, we suggest that future collaborative writing tools introduce features to reduce multilingual context switching. For example, code editors like VSCode [62] support generating code by providing instructions in natural language, reducing the time for developers to search syntax of programming languages in external resources. Similarly, an AI-assisted editor could support NNS by providing high-level instructions in their native language and generating text in English.

Provide guardrails for machine translation. NNS is frequently involved in Wordsmith-Crosslingual during both individual and collaborative writing. However, machine translation is prone to information loss [63, 94], and NNS are usually unaware of it. For example, during an interview conducted by our collaborators, NS-1 mentioned that NNS-1 left a sentence, "*Why don't you stop comparing yourself to others and focus on yourself?*" Though NS-1 perceived an accusatory tone and wanted to change it, NS-1 eventually decided to respect NNS-1's writing and did not revise it. However, it

was later revealed that NNS-1 originally wrote it in Japanese and the soft tone was lost in translation. Therefore, future collaborative writing tools could have built-in translators and monitor the tone before and after translation.

Summarize collaborators' activities. During the collaborative stage, we observed that authors spent some time catching up on the latest changes in the document, for example, reading links provided by co-authors (Passive-Search). Though online collaborative writing tools like Google Docs have version history, they only reflect word-level change and do not summarize the semantic meanings of version changes. Therefore, future AI-assisted collaborative writing could support automatically summarizing co-authors' activities and facilitate the syncing up process.

8.3 Generalizability for understanding diverse collaborative processes

Though COALA is initially designed for a specific dataset that our collaborators have collected, COALA also has the potential to help analyze more generalized and diverse collaborative processes.

Diverse collaborative writing settings. Since our collaborators' study is one of the first studies on multilingual collaborative writing, they opted for a simplified setup: one NS and one NNS with clear turn-taking. In the future, communication researchers may conduct user studies with more complex scenarios, e.g., multiple NS and NNS, or NNS from different countries. Since COALA follows a generalized "cluster-summarize-compare" workflow, COALA can adapt to and process new sequences provided by researchers. Furthermore, the computational methods underpinning COALA are agnostic to event or author types, enabling researchers to introduce events and author types beyond those in Table 1. For example, if communication researchers want to study how people of varying Wikipedia editing experiences co-write an article, they could leverage a set of events recorded by Wikipedia, such as "create", "add", "delete", "revert". In this case, the factor of interest shifts from language proficiency to the authors' Wikipedia editing experience.

Collaborative processes beyond writing. While COALA focuses on analyzing multilingual collaborative writing, the design principles and computational techniques could extend to a wide range of collaborative processes that generate rich event sequences, such as visual information analysis, knowledge synthesis, and problem-solving. For instance, Isenberg et al. [36] examined how different teams behave in an information analysis task and derived a framework for such activities by analyzing temporal sequences. In another study, Isenberg et al. [35] recruited 15 pairs of participants to solve a problem by exploring 240 digital documents and identified eight collaboration styles. Similarly, Robinson et al. [80] recruited five pairs of geographers and disease biologists to complete a knowledge synthesis task, analyzing the participants' action frequency, time durations, and strategies. Recently, Yang et al. [108] studied how teams collaboratively engage and perform sensemaking tasks in an immersive environment. These studies highlight the importance of understanding the temporal sequences of participants' behaviors and identifying common patterns—an area where COALA's clustering and summarization features could provide significant benefits. Unlike previous methods which rely primarily on manual coding or frequency-based analysis, COALA offers a data-driven approach to interpret collaborative dynamics, and thus enhances both scalability and granularity in data analysis.

8.4 Limitations

One limitation of COALA is that it does not visualize text changes during collaborative writing. In the individual user studies, the only participant who struggled to complete the analysis tasks noted that having text alongside event sequences would make the analysis more concrete. While we explored adapting existing text visualizations to our dataset (see Appendix), our focus is on novel approaches to support behavioral analysis. Future tools could seamlessly integrate both aspects to enhance analytical depth. Another limitation is that COALA's comparison feature is designed for dichotomous analysis, such as collaborative vs. individual or NS vs. NNS. For teams without clear binary distinctions, COALA only supports clustering but lacks dedicated comparison capabilities.

9 Conclusion

To compare native and non-native English speakers' behaviors in collaborative writing, we partnered with two communication researchers to design visual analytics techniques. To address the limitations of existing text and event sequence visualizations, we implemented COALA to help communication researchers uncover insights during the collaborative writing process, mainly focusing on addressing interpretation and trust challenges. We validate COALA's effectiveness by conducting a focus group study (N=2+2) and an individual user study (N=8). Finally, we share design lessons learned during the development and potential features for AI-assisted collaborative writing tools. We believe this work will significantly contribute to collaborative writing communities, fostering a more diverse and inclusive work environment, especially for non-native speakers. Additionally, researchers aiming to understand other collaborative processes, such as collaborative knowledge synthesis and sensemaking, could also find inspiration in the designs of COALA.

Acknowledgments

We would like to thank user study participants and the anonymous reviewers for the constructive feedback. This work is supported by NSF awards IIS-2239130 and IIS-1947929.

References

[1] 2024. ILR scale - Wikipedia. https://en.wikipedia.org/wiki/ILR_scale. (Accessed on 02/15/2024).
[2] Open AI. 2024. Vision - OpenAI API. https://platform.openai.com/docs/guides/vision. (Accessed on 02/06/2024).
[3] Paul André, Robert E Kraut, and Aniket Kittur. 2014. Effects of simultaneous and sequential work structures on distributed collaborative interdependent tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 139–148.
[4] Atlassian. 2024. Confluence | Your Remote-Friendly Team Workspace | Atlassian. https://www.atlassian.com/software/confluence. (Accessed on 09/09/2024).
[5] Benjamin Bach, Conglei Shi, Nicolas Heulot, Tara Madhyastha, Tom Grabowski, and Pierre Dragicevic. 2015. Time curves: Folding time to visualize patterns of temporal evolution in data. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 559–568.
[6] Ronald M Baecker, Dimitrios Nastos, Ilona R Posner, and Kelly L Mawby. 1993. The user-centered iterative design of collaborative writing software. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 399–405.

[7] Jeremy Birnholtz, Stephanie Steinhardt, and Antonella Pavese. 2013. Write here, write now! An experimental study of group maintenance in collaborative writing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 961–970.

[8] Tom Boellstorff, Bonnie Nardi, Celia Pearce, and TL Taylor. 2013. Words with friends: Writing collaboratively online. *Interactions* 20, 5 (2013), 58–61.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[10] Angelos Chatzimparmpas, Rafael Messias Martins, Ilir Jusufi, Kostiantyn Kucher, Fabrice Rossi, and Andreas Kerren. 2020. The state of the art in enhancing trust in machine learning models with the use of visualizations. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 713–756.

[11] Yuexi Chen and Zhicheng Liu. 2024. WordDecipher: Enhancing Digital Workspace Communication with Explainable AI for Non-native English Speakers. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*. 7–10.

[12] Yuanzhe Chen, Panpan Xu, and Liu Ren. 2017. Sequence synopsis: Optimize visual summary of temporal event data. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 45–55.

[13] Rui Cheng. 2013. A non-native student's experience on collaborating with native peers in academic literacy development: A sociopolitical perspective. *Journal of English for Academic Purposes* 12, 1 (2013), 12–22.

[14] N Ann Chenoweth and John R Hayes. 2001. Fluency in writing: Generating text in L1 and L2. *Written communication* 18, 1 (2001), 80–98.

[15] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. 2012. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 443–452.

[16] David Cohn, Rich Caruana, and Andrew McCallum. 2003. Semi-supervised clustering with user feedback. *Constrained clustering: advances in algorithms, theory, and applications* 4, 1 (2003), 17–32.

[17] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100* (2023).

[18] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[19] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice over control: How users write with large language models using diegetic and non-diegetic prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[20] Paul Dourish and Victoria Bellotti. 1992. Awareness and coordination in shared workspaces. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*. 107–114.

[21] Douglas C Engelbart. 1984. Collaboration support provisions in AUGMENT. In *Proceedings of the 1984 AFIPS Office Automation Conference (Los Angeles CA, 1984, 51-58 p.)*.

[22] Robert S Fish, Robert E Kraut, and Mary DP Leland. 1988. Quilt: A collaborative tool for cooperative writing. In *Proceedings of the ACM SIGOIS and IEEECS TC-OA 1988 conference on Office information systems*. 30–37.

[23] Fabian Flöck and Maribel Acosta. 2015. whovis: Visualizing editor interactions and dynamics in collaborative writing over time. In *Proceedings of the 24th International Conference on World Wide Web*. 191–194.

[24] Fabian Flöck, David Laniado, Felix Stadthaus, and Maribel Acosta. 2015. Towards Better Visual Tools for Exploring Wikipedia Article Development—The Use Case of" Gamergate Controversy". In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 9. 48–55.

[25] Philippe Fournier-Viger, Cheng-Wei Wu, Antonio Gomariz, and Vincent S Tseng. 2014. VMSP: Efficient vertical mining of maximal sequential patterns. In *Advances in Artificial Intelligence: 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, 2014. Proceedings 27*. Springer, 83–94.

[26] Erving Goffman. 1981. *Forms of talk*. University of Pennsylvania Press.

[27] David Gotz, Jonathan Zhang, Wenyuan Wang, Joshua Shrestha, and David Borland. 2019. Visual analysis of high-dimensional event sequence data via dynamic hierarchical aggregation. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 440–450.

[28] Shunan Guo, Ke Xu, Rongwen Zhao, David Gotz, Hongyuan Zha, and Nan Cao. 2017. Eventthread: Visual summarization and stage analysis of event sequence data. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 56–65.

[29] Yi Guo, Shunan Guo, Zhuochen Jin, Smiti Kaul, David Gotz, and Nan Cao. 2021. Survey on visual analysis of event sequence data. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2021), 5091–5112.

[30] Scott A Hale. 2014. Multilinguals and Wikipedia editing. In *Proceedings of the 2014 ACM conference on Web science*. 99–108.

[31] Yi Han, Agata Rozga, Nevena Dimitrova, Gregory D Abowd, and John Stasko. 2015. Visual analysis of proximal temporal relationships of social and communicative behaviors. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 51–60.

[32] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.

[33] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.

[34] Mengdie Hu, Krist Wongsuphasawat, and John Stasko. 2016. Visualizing social media content with sententree. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 621–630.

[35] Petra Isenberg, Danyel Fisher, Sharoda A Paul, Meredith Ringel Morris, Kori Inkpen, and Mary Czerwinski. 2011. Co-located collaborative visual analytics around a tabletop display. *IEEE Transactions on visualization and Computer Graphics* 18, 5 (2011), 689–702.

[36] Petra Isenberg, Anthony Tang, and Sheelagh Carpendale. 2008. An exploratory study of visual information analysis. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1217–1226.

[37] Takumi Ito, Naomi Yamashita, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, Ge Gao, Jack Jamieson, and Kentaro Inui. 2023. Use of an AI-powered Rewriting Support Software in Context with Other Tools: A Study of Non-Native English Speakers. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[38] Eun Young Kang. 2020. Using model texts as a form of feedback in L2 writing. *System* 89 (2020), 102196.

[39] Khaled Karim and Hossein Nassaji. 2020. The revision and transfer effects of direct and indirect comprehensive corrective feedback on ESL students' writing. *Language Teaching Research* 24, 4 (2020), 519–539.

[40] Leonard Kaufman and Peter J Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

[41] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. *Visual analytics: Definition, process, and challenges*. Springer.

[42] Hee-Cheol Kim and Kerstin Severinson Eklundh. 2001. Reviewing practices in collaborative writing. *Computer Supported Cooperative Work (CSCW)* 10 (2001), 247–259.

[43] Taewan Kim, Donghoon Shin, Young-Ho Kim, and Hwajung Hong. 2024. DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[44] Yewon Kim, Mina Lee, Donghwi Kim, and Sung-Ju Lee. 2023. Towards explainable ai writing assistants for non-native english speakers. *arXiv preprint arXiv:2304.02625* (2023).

[45] Aniket Kittur, Bongwon Suh, Bryan A Pendleton, and Ed H Chi. 2007. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 453–462.

[46] Josua Krause, Adam Perer, and Harry Stavropoulos. 2015. Supporting iterative cohort construction with visual temporal queries. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 91–100.

[47] Milos Krstajic, Enrico Bertini, and Daniel Keim. 2011. Cloudlines: Compact display of event episodes in multiple time-series. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2432–2439.

[48] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2, 1-2 (1955), 83–97.

[49] Philippe Laban, Jesse Vig, Marti A Hearst, Caiming Xiong, and Chien-Sheng Wu. 2023. Beyond the Chat: Executable and Verifiable Text-Editing with LLMs. *arXiv preprint arXiv:2309.15337* (2023).

[50] Ida Larsen-Ledet and Henrik Korsgaard. 2019. Territorial functioning in collaborative writing: fragmented exchanges and common outcomes. *Computer Supported Cooperative Work (CSCW)* 28 (2019), 391–433.

[51] Ida Larsen-Ledet, Henrik Korsgaard, and Susanne Bødker. 2020. Collaborative writing across multiple artifact ecologies. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[52] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, et al. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–35.

[53] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[54] Xiaoyan Li, Naomi Yamashita, Wen Duan, Yoshinari Shirai, and Susan R Fussell. 2023. Improving Non-Native Speakers' Participation with an Automatic Agent

in Multilingual Groups. *Proceedings of the ACM on Human-Computer Interaction* 7, GROUP (2023), 1–28.

[55] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2023. MultiViz: Towards User-Centric Visualizations and Interpretations of Multimodal Models. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[56] Zhicheng Liu, Bernard Kerr, Mira Dontcheva, Justin Grover, Matthew Hoffman, and Alan Wilson. 2017. Coreflow: Extracting and visualizing branching patterns from event sequences. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 527–538.

[57] Zhicheng Liu, Yang Wang, Mira Dontcheva, Matthew Hoffman, Seth Walker, and Alan Wilson. 2016. Patterns and sequences: Interactive exploration of clickstreams to understand common visitor paths. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 321–330.

[58] Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. 2023. InsightPilot: An LLM-Empowered Automated Data Exploration System. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 346–352.

[59] Jessica Magallanes, Tony Stone, Paul D Morris, Suzanne Mason, Steven Wood, and Maria-Cruz Villa-Uriol. 2021. Sequen-c: A multilevel overview of temporal event sequences. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 901–911.

[60] Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. 2015. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. 38–49.

[61] Miriah Meyer, Tamara Munzner, and Hanspeter Pfister. 2009. MizBee: a multiscale synteny browser. *IEEE transactions on visualization and computer graphics* 15, 6 (2009), 897–904.

[62] Microsoft. 2024. Visual Studio Code - Code Editing. Redefined. https://code.visualstudio.com/. (Accessed on 09/09/2024).

[63] Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research* 55 (2016), 95–130.

[64] Megan Monroe, Krist Wongsuphasawat, Catherine Plaisant, Ben Shneiderman, Jeff Millstein, and Sigfried Gold. 2012. Exploring point and interval event patterns: Display methods and interactive visual query. *University of Maryland Technical Report* 80 (2012), 134.

[65] Fionn Murtagh and Pedro Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, 1 (2012), 86–97.

[66] Eugene W Myers. 1986. An O (ND) difference algorithm and its variations. *Algorithmica* 1, 1-4 (1986), 251–266.

[67] Cal Newport. 2016. *Deep work: Rules for focused success in a distracted world*. Hachette UK.

[68] Phong H Nguyen, Rafael Henkin, Siming Chen, Natalia Andrienko, Gennady Andrienko, Olivier Thonnard, and Cagatay Turkay. 2019. Vasabi: Hierarchical user profiles for interactive visual user behaviour analytics. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 77–86.

[69] Judith S Olson, Dakuo Wang, Gary M Olson, and Jingwen Zhang. 2017. How people write together now: Beginning the investigation with advanced undergraduates in a project course. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1 (2017), 1–40.

[70] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[71] Overleaf. 2024. Overleaf, Online LaTeX Editor. https://www.overleaf.com. (Accessed on 09/10/2024).

[72] So Yeon Park and Sang Won Lee. 2023. Why "why"? The Importance of Communicating Rationales for Edits in Collaborative Writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–25.

[73] Adam Perer and Jimeng Sun. 2012. Matrixflow: temporal network visual analytics to track symptom evolution during disease progression. In *AMIA annual symposium proceedings*, Vol. 2012. American Medical Informatics Association, 716.

[74] Adam Perer and Fei Wang. 2014. Frequence: Interactive mining and visualization of temporal frequent event sequences. In *Proceedings of the 19th international conference on Intelligent User Interfaces*. 153–162.

[75] Ignacio Perez-Messina, Claudio Gutierrez, and Eduardo Graells-Garrido. 2018. Organic visualization of document evolution. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*. 497–501.

[76] Catherine Plaisant, Brett Milash, Anne Rose, Seth Widoff, and Ben Shneiderman. 1996. LifeLines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 221–227.

[77] Peter J Polack Jr, Shang-Tse Chen, Minsuk Kahng, Kaya De Barbaro, Rahul Basole, Moushumi Sharmin, and Duen Horng Chau. 2018. Chronodes: Interactive multifocus exploration of event sequences. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 1 (2018), 1–21.

[78] Ji Qi, Vincent Bloemen, Shihan Wang, Jarke Van Wijk, and Huub Van De Wetering. 2019. STBins: Visual tracking and comparison of multiple data sequences using temporal binning. *IEEE Transactions on visualization and computer graphics* 26, 1 (2019), 1054–1063.

[79] Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan "Michael" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2024. ABScribe: Rapid Exploration & Organization of Multiple Writing Variations in Human-AI Co-Writing Tasks using Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[80] Anthony Christian Robinson. 2008. *Design for synthesis in geovisualization*. The Pennsylvania State University.

[81] Bahareh Sarrafzadeh, Sujay Kumar Jauhar, Michael Gamon, Edward Lank, and Ryen W White. 2021. Characterizing stage-aware writing assistance for collaborative document authoring. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–29.

[82] Jodi Schneider, Alexandre Passant, and John G Breslin. 2011. Understanding and improving Wikipedia article discussion spaces. In *Proceedings of the 2011 ACM Symposium on Applied Computing*. 808–813.

[83] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)* 42, 3 (2017), 1–21.

[84] Carol Severino, Jeffrey Swenson, and Jia Zhu. 2009. A comparison of online feedback requests by non-native English-speaking and native English-speaking writers. *The Writing Center Journal* 29, 1 (2009), 106–129.

[85] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 364–371.

[86] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *23rd International Conference on Intelligent User Interfaces*. 293–304.

[87] Vilaythong Southavilay, Kalina Yacef, Peter Reimann, and Rafael A Calvo. 2013. Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *Proceedings of the third international conference on learning analytics and knowledge*. 38–47.

[88] Hendrik Strobelt, Jambay Kinley, Robert Krueger, Johanna Beyer, Hanspeter Pfister, and Alexander M Rush. 2021. Genni: Human-ai collaboration for data-backed text generation. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 1106–1116.

[89] Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P Dow. 2024. ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 120–137.

[90] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2016. *Introduction to data mining*. Pearson Education India.

[91] Johnny Torres, Sixto García, and Enrique Peláez. 2019. Visualizing authorship and contribution of collaborative writing in e-learning environments. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 324–328.

[92] Selen Türkay, Daniel Seaton, and Andrew M Ang. 2018. Itero: A revision history analytics tool for exploring writing behavior and reflection. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[93] Kozue Uzawa and Alister Cumming. 1989. Writing strategies in Japanese as a foreign language: Lowering or keeping up the standards. *Canadian Modern Language Review* 46, 1 (1989), 178–194.

[94] Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. *arXiv preprint arXiv:1906.12068* (2019).

[95] Sandro Vega-Pons and José Ruiz-Shulcloper. 2011. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence* 25, 03 (2011), 337–372.

[96] Fernanda B Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 575–582.

[97] Dakuo Wang, Judith S Olson, Jingwen Zhang, Trung Nguyen, and Gary M Olson. 2015. DocuViz: visualizing collaborative writing. In *Proceedings of the 33rd Annual ACM conference on human factors in computing systems*. 1865–1874.

[98] Dakuo Wang, Haodan Tan, and Tun Lu. 2017. Why users do not want to write together when they are writing together: Users' rationales for today's collaborative writing practices. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–18.

[99] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y Zhao. 2016. Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 225–236.

[100] Jishang Wei, Zeqian Shen, Neel Sundaresan, and Kwan-Liu Ma. 2012. Visual cluster exploration of web clickstream data. In *2012 IEEE conference on visual analytics science and technology (VAST)*. IEEE, 3–12.

[101] Mark Wolfersberger. 2003. L1 to L2 writing process and strategy transfer: A look at lower proficiency writers. *TESL-EJ* 7, 2 (2003), 1–12.

[102] Krist Wongsuphasawat and David Gotz. 2011. Outflow: Visualizing patient flow by symptoms and outcome. In *IEEE VisWeek Workshop on Visual Analytics in Healthcare, Providence, Rhode Island, USA*. American Medical Informatics Association, 25–28.

[103] Krist Wongsuphasawat, John Alexis Guerra Gómez, Catherine Plaisant, Taowei David Wang, Meirav Taieb-Maimon, and Ben Shneiderman. 2011. LifeFlow: visualizing an overview of event sequences. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1747–1756.

[104] Yimin Xiao, Yuewen Chen, Naomi Yamashita, Yuexi Chen, Zhicheng Liu, and Ge Gao. 2024. (Dis) placed Contributions: Uncovering Hidden Hurdles to Collaborative Writing Involving Non-Native Speakers, Native Speakers, and AI-Powered Editing Tools. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–31.

[105] Naomi Yamashita, Andy Echenique, Toru Ishida, and Ari Hautasaari. 2013. Lost in transmittance: how transmission lag enhances and deteriorates multilingual collaboration. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 923–934.

[106] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th international conference on intelligent user interfaces*. 189–201.

[107] Weipeng Yang, Yongyan Li, and Hui Li. 2021. Supervisor as coauthor in writing for publication: Evidence from a cohort of non-native English-speaking Master of Education students. *SN Social Sciences* 1, 2 (2021), 44.

[108] Ying Yang, Tim Dwyer, Michael Wybrow, Benjamin Lee, Maxime Cordeil, Mark Billinghurst, and Bruce H Thomas. 2022. Towards immersive collaborative sensemaking. *Proceedings of the ACM on Human-Computer Interaction* 6, ISS (2022), 722–746.

[109] Xuchao Zhang, Dheeraj Rajagopal, Michael Gamon, Sujay Kumar Jauhar, and ChangTien Lu. 2019. Modeling the relationship between user comments and edits in document revision. 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)*. 5002–5011.

[110] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. Visar: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–30.

[111] Jian Zhao, Zhicheng Liu, Mira Dontcheva, Aaron Hertzmann, and Alan Wilson. 2015. Matrixwave: Visual comparison of event sequence data. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 259–268.

[112] Yuheng Zhao, Xinyu Wang, Chen Guo, Min Lu, and Siming Chen. 2023. ContextWing: Pair-wise Visual Comparison for Evolving Sequential Patterns of Contexts in Social Media Data Streams. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–31.

[113] Kazi Tasnim Zinat, Saimadhav Naga Sakhamuri, Aaron Sun Chen, and Zhicheng Liu. 2024. A Multi-Level Task Framework for Event Sequence Analysis. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[114] Kazi Tasnim Zinat, Jinhua Yang, Arjun Gandhi, Nistha Mitra, and Zhicheng Liu. 2023. A Comparative Evaluation of Visual Summarization Techniques for Event Sequences. *Computer Graphics Forum* 42, 3 (2023), 173–185. https://doi.org/10.1111/cgf.14821 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14821

A Understanding document modification behaviors

In this section, we focus on understanding document modification behaviors E_{on}. Such behaviors could be inferred by comparing consecutive document versions via the Myers difference algorithm [66].

A.1 Task Analysis

In our interviews with communication researchers, there are two tasks about analyzing document modification behaviors:

T1: identify editing dynamics. Communication researchers hypothesize that NS and NNS contribute differently to the joint document and would like to know the difference reflected by the evolution of documents.

T2: analyze specific content. Communication researchers are interested in finding frequently co-edited content and seeing how NS and NNS edit them during collaborative writing.

A.2 Understanding editing dynamics by branched time curves

To tackle **T1**, we draw inspirations from the folded time curve proposed by Bach et al. [5], where each dot represents a document version, the order on the curve denotes the temporal information, and the pairwise distance depends on the text similarity. We proposed a modified version as depicted in Figure 8: since authors write individually before collaborating on the same document, we adopted a two-head start, indicated by the gray arrows, which refer to the initial draft written by NS and NNS. Then, we use red and blue to refer to NNS and NS' contributions, respectively. The communication researchers suggest highlighting the differences between intermediate document versions V_{inter} and end-of-the-day versions V_{end}, so we apply big filled dots for V_{end} and small hollow dots for V_{inter}.

The branched time curves give an overview of the document's evolution, especially the "V" shape part, which shows how the merged version differs from NNS and NS' initial draft. Inspired by the proximity of the merged version to NS' version, in a recently published paper at a premier social computing conference (not cited here for anonymity), our collaborators conducted a statistical significance test and found that the merged document's lexical distance was indeed significantly closer to NS' initial writing. Meanwhile, in subsequent versions, NS' edits always result in a larger lexical distance than NNS.

A.3 Tracking specific content with Sentence Flow

Though the time curves provide an overview on the document-level contributions from NS and NNS, the coarse granularity limits us from uncovering more content-specific insights (**T2**). Inspired by history flow [96, 97], we propose a revised version called Sentence Flow, as depicted in Figure 9. The x-axis is the author; the y-axis are individual sentences in the document. The edits of a sentence is color-coded: white for no activity, blue and red for word-level add and remove, and black for deleting the entire sentence. The darkness of red and blue indicates the edit distance, the darker the larger. On top of the sentence flow, there are also bar charts quantify the total edit distance. If users click on a sentence, they can see the content and change logs.

Our collaborators identified two dominant editing patterns in sentence flow. One is within-author editing, where authors primarily revise their own sentences and rarely modify others', e.g., in Team 1, adjacent colored areas are uncommon, with only one instance shown in Figure 9, where the NS made minor modifications to NNS' sentences. The second pattern is between-author editing, as seen in Team 9 in Figure 9, where several sentences were consecutively revised by different authors, resulting in a more balanced distribution.

Though the modified versions of existing visualizations reveal document modification behaviors of NS and NNS, it does not answer questions about how content are generated in collaborative writing. To address these questions, we conduct and present an event sequence analysis of content generation behaviors in the next section.

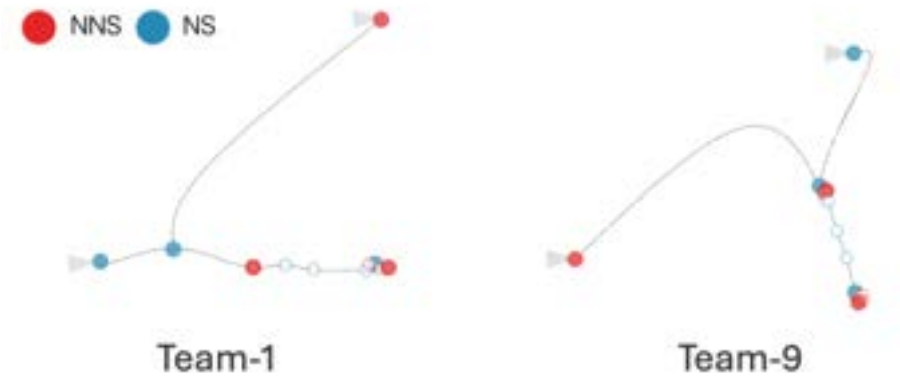

Figure 8: Branched time curves of team-1 and team-9. The gray arrows indicate the first versions written by NS and NNS. Big and filled dots are major versions, and small and hollow dots are intermediate versions.

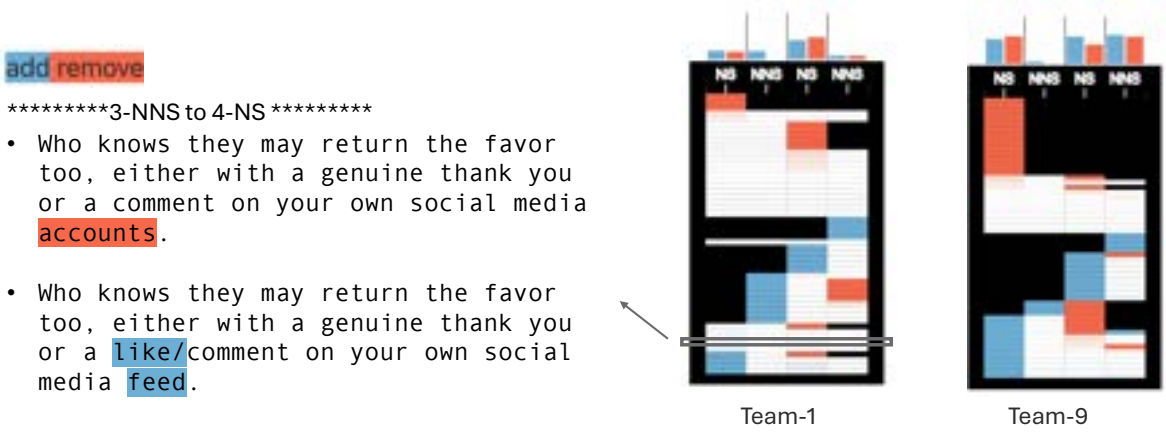

Figure 9: Sentence Flow of team-1 and team-9. Red: remove, blue: add, gray: edited by both authors.