

Characterizing Stage-Aware Writing Assistance in Collaborative Document Authoring

BAHAREH SARRAFZADEH*, Microsoft, USA
SUJAY KUMAR JAUHAR, Microsoft Research AI, USA
MICHAEL GAMON, Microsoft Research AI, USA
EDWARD LANK, University of Waterloo, Canada
RYEN W. WHITE, Microsoft Research AI, USA

Writing is a complex non-linear process that begins with a mental model of intent, and progresses through an outline of ideas, to words on paper (and their subsequent refinement). Despite past research in understanding writing, Web-scale consumer and enterprise collaborative digital writing environments are yet to greatly benefit from intelligent systems that understand the stages of document evolution, providing opportune assistance based on authors' situated actions and context. In this paper, we present three studies that explore temporal stages of document authoring. We first survey information workers at a large technology company about their writing habits and preferences, concluding that writers do in fact conceptually progress through several distinct phases while authoring documents. We also explore, qualitatively, how writing stages are linked to document lifespan. We supplement these qualitative findings with an analysis of the longitudinal user interaction logs of a popular digital writing platform over several million documents. Finally, as a first step towards facilitating an intelligent digital writing assistant, we conduct a preliminary investigation into the utility of user interaction log data for predicting the temporal stage of a document. Our results support the benefit of tools tailored to writing stages, identify primary tasks associated with these stages, and show that it is possible to predict stages from anonymous interaction logs. Together, these results argue for the benefit and feasibility of more tailored digital writing assistance.

Additional Key Words and Phrases: Writing; Collaborative Writing; Intelligent writing assistance; Workplace; Productivity

ACM Reference Format:

Bahareh Sarrafzadeh, Sujay Kumar Jauhar, Michael Gamon, Edward Lank, and Ryen W. White. 2020. Characterizing Stage-Aware Writing Assistance in Collaborative Document Authoring. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW, Article 1 (October 2020), 29 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

When authoring a document, a writer does not go from a blank page to a fully completed text in a linear fashion. Rather they may go through several overlapping, sometimes cyclical, *stages* [21, 48] of planning, outlining, drafting, revising, soliciting feedback, and copy-editing to crystallize their

*Work done while a PhD candidate at University of Waterloo and interning at Microsoft Research

Authors' addresses: Bahareh Sarrafzadeh, Microsoft, Redmond, WA, USA, bsarrafz@uwaterloo.ca; Sujay Kumar Jauhar, Microsoft Research AI, Redmond, WA, USA, sjauhar@microsoft.com; Michael Gamon, Microsoft Research AI, Redmond, WA, USA, mgamon@microsoft.com; Edward Lank, University of Waterloo, Waterloo, ON, Canada, lank@uwaterloo.ca; Ryen W. White, Microsoft Research AI, Redmond, WA, USA, ryenw@microsoft.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Association for Computing Machinery.

2573-0142/2020/10-ART1 \$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

thoughts into words. Progressing through these stages, the author might have very different mental models of, or approaches towards, their goals and writing needs at any given point in time. For example, in drafting mode, their goal might be to simply write down thoughts, without worrying too much about spelling or grammar; whereas in copy-editing mode, they might want to go back and correct their typographical or grammatical errors.

While digital writing environments include a number of tools (such as grammar- and spell-checkers, anchored comments, change tracking, etc.) to support authors as they progress through the stages of writing a document, the surfacing of these tools through the user experience can sometimes be inopportune, causing the writer to break their focus. Despite the widespread use of context-aware intelligent systems in many domains such as conversational search [46], travel planners [8], and calendar-aware email recommendation [67], modern writing tools often ignore the notion of writers' "situated actions" [53] beyond the most recent changes to a document's text.

While better individual writing support may benefit any individual author, understanding document evolution becomes even more important, we would argue, as writing moves online, primarily because one of the primary benefits of these online platforms is their support for collaborative document authoring. Greer et al. [26] note that, if writing is decomposed into distinct rhetorical phases, this phased writing process helps structure writing, and can guide collaborators co-creating documents to appropriate subtasks during different phases of document creation [34, 55]. Understanding – and depicting – document phases can also help new collaborators get up-to-speed, while supporting efforts at proper collaboration etiquette [17, 27].

While there is broad awareness that many documents are created collaboratively [55], research on document evolution and stages of writing has been restricted to laboratory studies or theoretical investigations. Previous efforts include the analysis of writing captured on video [59], thinking-aloud methods [20], and retrospective interviews [6, 45]. Most of these previous approaches focus on specific kinds of writing (e.g., essay writing in classroom [63]), require that documents be of a certain quality (e.g., project reports [43] or featured articles in Wikipedia [61]), or are limited in scope – and are therefore ill-suited to supporting a general-purpose Web-scale collaborative digital writing environment used by millions of users for heterogeneous document authoring.

In this paper, we probe the question of whether and how authors divide their writing into stages and the link between authors' actions and the temporal lifespan of a document they create. While document lifespan can, in theory, be infinite, here we focus on the "active" lifespan of a document, i.e. the portion of the document's existence when it is being actively modified. We explore whether authors of collaboratively written documents perceive writing stages [21, 48] and whether these stages can be discriminated via differences in user action over different periods of a document's (active) lifespan. To address this research goal, we leverage a mixed-methods approach [13] that combines data from three studies: a qualitative study that explores the link between writing tasks and document lifespan, a quantitative study that analyzes user behaviors over document lifetimes, and a preliminary predictive analysis that examines whether different stages within a collaboratively authored document's lifespan can be inferred from user interaction log data.

Our qualitative study surveys 183 information workers at a large technology company about their writing habits and finds that authors frequently collaborate around document creation and writing, and that authors have mental models of their writing goals, and these broadly correspond to temporal stages of a document's lifespan. Authors also note that different authoring stages are associated with different types of activities on the writing canvas, and they perceive the benefits of stage-aware writing assistance that supports them in their situated actions and through the stages of their writing, particularly for collaborative writing tasks. They also offer insights into the types of support needed at different stages. To validate the link between document lifespan and

different writing activities, we perform a follow-on quantitative analysis on the user interaction logs of several million documents authored on a popular Web-scale digital writing platform. Our quantitative log analysis confirms that writers distribute their writing activities differently over temporal stages of document lifespan both in terms of magnitude and types of these activities. The triangulation of our qualitative insights on authors' perceptions and user actions provides a better understanding of how authors engage with the writing task at different stages and serves as a guide to design of intelligent writing assistants that incorporate stage-appropriate assistance for writing activities. Finally, as a first step towards demonstrating the feasibility of stage-appropriate writing assistance tools, we model the temporal stage of document evolution using the interaction logs from our quantitative study. Our results indicate that with a statistically significant margin, log-based interactions can be leveraged to predict these stages of document evolution more accurately than a strong baseline.

One implication of our qualitative survey was that, while our log analyses indicate an ability to predict document stage, survey respondents were almost evenly split in their desire for automatic inference of document lifespan, primarily due to past negative experiences with intelligent writing assistants (e.g. Clippy). In our discussion, we revisit this aspect of our work, and highlight that document-lifespan inferences can be leveraged in many ways: as intelligent tools with passive user involvement; as defaults with an ability to over-ride; as a mixed initiative adaptation [9, 31]; or as an awareness mechanism to support more effective collaboration [27].

2 RELATED WORK

There are several areas of research related to stages of document authoring and intelligent writing assistance. We review them in what follows and highlight differences to our work.

2.1 Writing Process

A number of papers have investigated the process of writing, both as an individual effort and as a collaborative endeavor. Some studies focus largely on the text itself, characterizing writing as an "outline-draft-edit" cycle. These works assume that "meaning" is first assembled in the author's mind and then articulated in text [38]. Other studies adopt the *process* view of writing, basing their research on cognitive process models [22] for problem solving. Broadly this involves sequences of activities, such as setting goals, planning, organizing, transcribing, and editing. Fitzgerald [19] presents a comprehensive overview of early approaches to the study of writing. She identifies three models of writing: the Stages Model [48], the Problem-solving Model [21], and the Social Interaction Model [41]. While there is no single comprehensive theory of the writing process [26], the Flower and Hayes model [20] is one of the more widely accepted theories. It considers writing as a series of cognitive processes. Through a think-aloud study they showed that the writing process involves three main sub-processes: planning, translating, and reviewing. Models of writing processes have been evaluated observationally in different ways, using – for example – video analysis [59], think-aloud methods [20], and retrospective interviews [45], mostly conducted at small-scale and in controlled settings. There has also been some effort on using keystroke logging to measure writing processes during typing [11, 52, 56].

Collaborative writing, on the other hand, builds on single author writing by involving multiple people, thus increasing the complexity of the writing process [24]. Although collaborative writing is a group effort, many activities are often divided and conducted on an individual basis [36, 54, 55]. One of the reasons for the amplified complexity is the increased need to coordinate between multiple viewpoints and work efforts [5], and the need to establish consensus [24, 27]. Over the last three decades different HCI and CSCW researchers have focused on people's various practices in collaborative writing and how tools can support these practices. While some earlier works

have largely involved improving technical capabilities of writing tools, such as updating revision histories [33] or users' permissions or access controls [16, 42], a different body of more recent work explores the adoption of collaborative writing tools long after they were developed and suggests that these tools or their collaborative capabilities are not necessarily used in practice [14, 40, 57, 58]. A range of social [7], personal [14] and privacy [58] factors remained important barriers to the adoption of these tools. The main takeaway from this line of work is that, while many technological advances in designing CSCW tools are supported by past research and are expected to provide utility for the users, understanding users' attitudes towards and perception of these advanced functionalities should be at the heart of designing these new collaborative writing tools. Similarly, our qualitative study also provides insights regarding the acceptance of stage-aware tools and intelligent assistants that aim at optimizing users' writing practices, and whether fully automated solutions are appreciated by the users.

2.2 Decomposing the Writing Process

In addition to research on the writing process itself, several previous studies have looked at the decomposition of the process into distinct stages. The cognitive process model of writing [20] divides up document authoring into distinctive and hierarchical phases, which enables writers to construct highly personal, idiosyncratic and complex workflows [55]. The complexity of these workflows requires authors to coordinate multiple and often competing writing activities [3, 4]. These processes compete for a writer's attention which can lead to cognitive overload [26, 34]. Especially high demand processes, such as planning and reviewing, require significant levels of attention control [11, 34]. Non-expert authors utilize strategies to reduce the cognitive load by *decomposing* the writing task into smaller, more manageable subtasks and stages [25, 26, 34, 35]. Effectively, this allows authors to scaffold their work, prompts thoughtful engagement and eases cognitive load [55].

2.3 Revision Research

Our work also relates to previous research on how revisions impact document evolution. In the education domain, analysis of revisions has typically relied on manual observation of students' essay writing practices in classrooms, and coding of writing behavior, as well as grading the textual products of revisions. The main motivation behind this line of work is to analyze the relationship between revision and either text quality [51] or their didactic effect [39]. For example, Faigley and Witte [18] show that expert writers revise in ways that are different from inexperienced writers. Differences between revisions contain valuable information for modeling document quality or extracting users' expertise and can additionally support various natural language processing (NLP) tasks [61]. The availability of vast amounts of data in Wikipedia revision histories has enabled modeling and applications in areas such as, sentence compression [60], lexical simplification [62], information retrieval [2], textual entailment recognition [65], language bias detection [47], and spelling error correction and paraphrasing [37, 66]. These tools, however, do not incorporate any awareness of the stage of writing or the state of document lifespan to provide more timely assistance and minimize interruptions to the users' workflow.

2.4 Contextual Recommendations

Broadly, the idea of stage-aware writing assistance also relates to research on contextual recommendation systems [1, 29, 50]. This work aims to leverage a user's context to improve the quality of personalized services. For example, previous approaches have tailored recommendations by integrating user context in conversational search [46], travel planners [8], intelligent email reminders [49], and calendar-aware email recommendation [67]. Relatively little prior research has focused on

contextual modeling for writing assistance, although Horvitz et al. [32] did explore using Bayesian inference to understand writers' intents and goals given their actions and queries.

Our work extends past research in two important ways. First, while past work examines the writing process [21, 22, 48] and decomposing this process [55], the link between stages of writing and document lifespan, particularly for multi-authored documents, is unclear. While it seems reasonable that writing stages would correlate with document lifespan [48], understanding authors' perspectives on the link between writing activities and document lifespan has not been clearly established in past research. Second, past document-based studies have often included text and word-level features of documents to understand writing stages. Our analysis is both larger in scale (millions versus thousands of documents) and leverages anonymized user interaction logs that do not include character or word-level features of a document.

3 STAGE-AWARE WRITING ASSISTANCE

Our work is motivated by a desire to understand, cognitively, how and whether authors perceive that there are stages in their writing process. Alongside this, we wished to explore how writing stages are linked to document lifespan and whether tools linked to document lifespan might better support collaborative writing tasks.

3.1 Survey Description and Analysis

To investigate whether providing writing assistance that is tailored to the current stage of a document under development is beneficial to the writers of that document, we need to better characterize two different aspects of *stage-aware writing assistance*:

- (1) **Characterizing the stage of a document:** How are the stages of document evolution perceived by the writers?
- (2) **Characterizing stage-appropriate writing assistance:**
What writing assistance is beneficial at different stages of document evolution?

To elicit people's perspectives on writing stages and whether stage-appropriate assistance might be beneficial, we distributed a survey within a large technology company. The survey included both closed and open-ended questions. We collected 183 completed responses to this survey (response rate: 10%, completion rate: 74%).

Our survey started with a section on demographics including job roles, number of current projects, and team size. The next section involved different questions regarding writing documents at the workplace. These questions helped us verify whether writing documents at the workplace is a common practice, what tools are frequently used for writing documents, and how often these documents involve other collaborators. Finally, the main segment of our survey asked participants to reflect on their perception of the utility of an intelligent writing assistant and how stage-aware writing assistance should be realized.

Respondents came from a diverse set of roles ranging from product managers and software developers to sales people and administrative assistants, the majority had been working at the company for more than 5 years (63%). Multi-tasking on different projects was a common trend among them, with 75% working on at least 3 projects in parallel at the time of completing the survey.

3.1.1 Analysis. The responses to closed-ended questions were aggregated to provide general statistics on participant demographics and writing habits. The responses to open-ended questions were analyzed using affinity diagramming. We use the K-J method for affinity diagramming (i.e. paper-note-quote based affinity diagramming) [28, 30]. As per Corbin and Strauss [12], we separate our coding into initial open coding (19 cluster) and subsequent axial coding where we identify

causal factors, connections, implications and other relationships between data to combine clusters into overall themes. We found that saturation for our qualitative data occurred approximately after 80 participants, and from this point no new clusters of information were identified. However, we continued to cluster responses for the remaining participants, particularly attuned to data that might expand our clusters or add nuance to our analysis.

3.2 Qualitative Findings

We found that 87% of respondents mentioned writing documents as a part of their work routine. They use a wide range of writing tools including Microsoft Word's Desktop client (89%) and Online (Web-based) client (49%), Notepad / Notepad++ (19.3%), OneNote (16%), Overleaf / Sharelatex (11.1%), Google Docs (8.2%), and others.

Collaboration on these documents was a common theme, with the majority of participants (74%) collaborating on at least 25% of their documents. A closer examination of the qualitative responses (discussed below) indicated that the need for tailored support in different stages of document authoring is associated with both Single Author Writing (SAW) and Collaborative Writing (CW) scenarios.

"in an ideal scenario I would just do some sort of content dump to the system and the system structures this text. So that I don't have to think about paragraphs, sections, where to put what etc. And then in later stages I want help with how to nicely formulate things. That's not only in a collaborative setting though." [P10]

To understand whether there is a need for different types of support at different stages of document evolution, the participants were asked *"Should an intelligent writing assistant provide different types of support at different stages of writing in the document?"* with a binary Yes / No response type.

Over 70% of participants expressed a need for different types of support at different stages of writing. To elicit the rationales behind these responses, the participants were presented with a series of follow-up questions that provided an opportunity for them to elaborate on their responses and helped with decoupling our participants' attitudes towards stage-aware assistance versus intelligent writing tools.

Participants who responded Yes to the above question were prompted to elaborate on their response by the following three questions:

[Q1] Why do you think different types of support are needed at different stages of document authoring? Can you think of some scenarios where you needed different types of support at different times?"

[Q2] "Should the AI infer your current phase of writing (e.g. drafting vs reviewing vs editing) or would you like to specify it yourself?"

[Q3-a] [AI should detect it] "How do you expect the Intelligent Assistant to customize your experience based on your inferred stage?"

[Q3-b] [I'd like to specify it] "Why do you prefer to be in charge of specifying your current phase of writing to the system?"

Participants who responded No to the initial question were asked *"Why do you think different types of support are NOT needed at different stages of writing?"*

A closer examination of these responses to the follow-up questions indicated that while 70% of all participants were in favor of stage-aware writing assistance they were mixed between a desire for the AI system to detect this stage (40%), specifying the stage themselves (40%) and Mixed-initiative (20%). Over 68% of participants who responded No to the initial question as well as 65% of participants who responded Yes to this question but mentioned they would like to specify the stage themselves cited bad experiences with the performance of AI systems and a lack of trust

in the ability of intelligent systems to get this right as the main rationale behind their reluctance to the idea of intelligent stage-aware writing assistants. Cumulatively, we find participants separated into three groups: approximately half of participants viewed some form of intelligent, stage-aware writing assistance as valuable; an additional 40% were in favor of stage-aware writing tools, but wished to specify themselves, which tool sets to use; finally, 10% of participants were indifferent.

In coding our responses to these questions, one theme that emerged from our data was a link between temporal stages of document authoring and temporal stages of writing. While Teevan et al. [55] break collaborative writing tasks down into microtasks to support the idiosyncratic nature of workflows, one aspect that has been unexplored by past work is this link between temporal stages of writing and temporal evolution of a document. To probe this issue in additional detail, the qualitative analysis of the follow up open-ended questions resulted in four broad themes, corresponding to different notions of the stage of a document as perceived by our participants and the utility of different types of writing assistance based on these stages.

The following three subsections elaborate on different perceptions of a document's stage and their corresponding writing activities, which has implications for designing stage-aware writing assistants. The final subsection provides a different perspective on the utility of document stage awareness, which is essential for more effective collaborative writing scenarios.

3.2.1 Temporal Stages of Document Authoring. The Stages Model of Writing [48] focuses on the evolution of the text as an external by-product of the writing where the main three stages of pre-writing, writing, and rewriting are separated in time. This notion of document stages resonated with many of our participants, who referred to them as early versus later stages of document writing. Qualitative analysis of the respondent's descriptions of these temporal stages revealed how they are conceptualized and what writing phases and activities are associated with each stage.

The early stages of writing are perceived to be highly focused and many participants expressed a desire for fewer system suggestions and notifications during these stages, particularly about minor fixes and improvements.

"I want fewer interruptions when i am focusing on adding content early in the document writing process." [P11]

"As someone who primarily writes documents in LaTeX, I like the ability to focus on writing the content without having to worry about formatting. Then only once the content is put down do I go back and tweak the look of the document. This allows me to focus on different aspects of the documents at different stages." [P12]

Later stages, on the other hand, are described as the time when most of the content has been added to the document and the writers can now focus on improving different aspects of the document during Reviewing, Copy-Editing and Finalizing phases.

"Document creation and document perfection are fundamentally different. In my opinion, the best way to create a collaborative document is to get a rough draft made as quickly and dirtily as possible and then begin to improve it; different forms of assistance are required for each scenario." [P237]

At later stages the document is more "mature" with most of the content already included and structured. This quality has at least two different implications for what support can and should be provided at earlier versus later stages of writing. First, in later stages the system can provide more end to end and global feedback as opposed to local suggestions (e.g., identifying missing topics to be included versus spelling errors to be fixed). Second, with more content in place and a longer history of edits, intelligent writing assistants can be more reliable in predicting what support is needed.

“When I’m actively working on creating the document I want suggestions around better writing style, but focused on what I’m doing at that moment. Once the document is more mature I’d want feedback on the end to end document, not a specific word here or there.” [P106]

“During the draft phase, there will be a lot of changes that just won’t stick. At reviewing phase it would be helpful to objectively view the whole document to suggest improvements.” [P207]

“Once a lot of content is in a document, a ‘bot’ may be able to warn that you are going to change something that was just edited.” [P77]

“Intelligence should improve as the document matures. The intent of the document, the subject matter, and the style of the document will become more clear as it evolves.” [P78]

Finally, while supporting the activities that are associated with a certain stage was appreciated by survey respondents, untimely system support can be perceived as *distracting*, *annoying*, or *unnecessary*. In particular, minor fixes, proofing and grammar fixes, collaborative activities such as commenting and change tracking are perceived as unhelpful during early stages of document evolution.

“Worrying about complete sentences and grammar while brainstorming would slow things down considerably ... but necessary for a final release.” [P227]

“Initial document writing is mainly non-collaborative, whereas later it’s a matter of tweaking and improving with the feedback of others.” [P178]

The following three design implications are our main takeaways from the preceding analysis:

- (1) Intelligent support should be provided based on the activities associated with the current stage of document evolution (i.e., early versus later stages).
- (2) Authors are in a focused mode during earlier stages of writing and system interruptions should be minimized.
- (3) Authors can benefit from a higher level of control over the writing process in earlier stages with outlining, structuring, or searching functionalities that are provided on demand. During later stages, intelligent systems can be more proactive in providing feedback and suggestions to improve the overall document quality.

3.2.2 Phases of Writing. Different models of the writing process [21, 36] state that writers go through different phases throughout the document authoring process. These phases are characterized as brainstorming, outlining, drafting, reviewing, revising and copy editing [36] or more broadly as planning, translating, and reviewing [21]. Our participants’ responses reveal that the idea of going through different phases during the process of writing documents resonates with them and that they perceive the stage of the document based on these phases of writing. In fact these phases were described in a variety of terms: *“from document creation to document perfection”* [P237], *“starting from nothing phase”* [P98], *“Rough draft vs final draft”* [P79], *“Draft, Editing, Polishing”* [P93, P220], *“Creative or Brainstorming, ‘culling the first main draft’, draft editing, reviewing, finalizing”* [P113] *“Outlining, Writing, Consulting”* [P158], *“Schedule (content requirements), Reminders & goals, Refresh”* [P135], *“early drafting, end of drafting, reviewing”* [P224], etc.

In probing the ways our participants described these phases of writing we noted an overarching temporal ordering to these phases as well as a variety of activities that are associated with each phase.

*“Assistance for spellcheck/alignment is much better suited to the end of **drafting**. Reminders to share or tag people are better suited for **reviewing**.”* [P224]

*“**Outlining** should help you structure your document before you begin. **Writing** should check for spelling and grammar. **Consulting** should make sure information is accurate.”*

[P158]

*“The assistant could provide different services at different stages of the doc, such as during **editing** would provide with content suggestions; while **reviewing** it could provide with hints to include the right audience; on **publishing** it could help with Permission settings and privacy/compliance reviews”* [P50]

A second round of qualitative coding of these responses resulted in identifying different activities that are associated with different phases of writing as perceived by our participants:

- **Planning/Brainstorming Phase:** Outlining/Diagramming [P36], section naming and setup [P46], help with structuring the document [P158]
- **Drafting Phase:** Getting content or wording help [P44], structuring support to offload the mind [P63], support for sharing [P167]
- **Reviewing / Consulting Phase:** notifications for formatting, grammar and spelling errors [P40], look of the document [P12], proofing [P75, P224], repetitive language usage [P167], highlighting unaddressed comments [P37], hints to include the right audience [P50, P133, P167], reminders to share or tag people [P224], fact checking [P123, P158], adding required references and sources [P133]
- **Finalizing / Publishing Phase:** checking for broken paragraphs or tables before printing [P42], help with permission settings [P78], help with requesting required privacy and compliance reviews [P50].

3.2.3 Modes of Writing. Writing documents usually spans multiple sessions of activity that are distributed over document lifetimes. When a writer attends to a document (representing the workspace), to do a particular task, they bring with them a general understanding of the current stage of a document (more clearly visible in SAW scenarios) and a basic idea of what to look for or what activity to do (see also Gutwin and Greenberg’s [27] descriptive framework of workspace awareness for real-time groupware).

Our participants referred to different goals they have when they attend to a document at different times, and that based on this goal they are seeking different types of support from the system. A system can provide support that matches the intended activity: For example, hiding comments and change tracking annotations when the user is not in “review mode” or disabling editing when the user is in “reading mode” can help simplify the user’s interaction with the writing environment and minimize accidental changes.

However, once we move to multi-author scenarios there is a need for a shared awareness of the document as the workspace, and the intent of a collaborator is impacted by this awareness.

“An ideal collaborative writing environment would allow users the choice of whether collaborators can view only or make edits to a document because the document should not be editable in certain situations.” [P206]

3.2.4 Awareness of Document’s Stage. So far, the three themes discussed above highlighted the importance of the detection of the current stage of a document or its corresponding writing phases to be able to provide writing support that is tailored to this stage. A distinct and interesting theme that was discovered through the qualitative coding of our survey responses was that the awareness of the stage of document’s evolution is not only crucial for the system, i.e., an intelligent writing assistant, but is also crucial for collaborating authors who join a shared document at different stages of its lifetime. In single author scenarios an author who attends to a document at some point in time generally has a good understanding of the current stage of a document and what the next steps should be, whereas in collaborative writing scenarios developing and maintaining this workspace awareness is challenging [27]. Gutwin and Greenberg [27], likewise, motivate the importance of

workplace awareness for collaborative scenarios which can reduce effort, increase efficiency, and reduce errors for the activities of collaboration. Assisting others with their local tasks is an integral part of collaboration and one that also benefits from workspace awareness. Awareness is useful because it helps people determine what assistance is *required* and what is *appropriate*.

Some of our participants were proactive in suggesting ways that an intelligent system can detect the current stage of a document and share that information with other collaborators who attend to the document at different times:

“[The system can track] the number of people with the document opened for edit. Tracking the “shares” of the document to know when it’s sent for review or edit. the intelligent assistant that provided me the prompts to share it for review or editing would know [the stage of the document] from the selections I make.” [P172]

3.2.5 Summary of Findings. Our survey data provided a range of insights regarding how authors perceive the stages of a document evolution and what types of writing assistance is expected at these stages of document authoring. The qualitative analysis of survey responses indicated that the writers in our population perceive a document stage in three different, yet related ways: (1) temporal stages of document evolution; (2) phases of writing, and (3) modes of writing.

While phases of writing have been identified in past work, one thing that past work has struggled with is the link between phases of writing and document lifespan, i.e. the evolution of the document. Teevan et al. [55] note that writing phases and writing workflow is idiosyncratic, and frequently interleaved; their solution is to break the writing process down into microtasks to support more effective collaboration. However, given that authors of multi-author documents (in our sample) perceive a link between document lifespan/document evolution and writing phases, one additional question that we can ask is whether or not different patterns of writing interactions can be identified over document lifespan such that differences in user behavior (e.g. commands used) could be associated with the stage of document evolution. To explore this question, in the next section we perform a log-based analysis that seeks to establish this link between the temporal stage of document evolution and differences in writer behavior.

4 CHARACTERIZING TEMPORAL STAGES OF DOCUMENT EVOLUTION

Ideally, one would be able to infer writing phases *a priori* and through logs of interactions, but to the best of our knowledge there is no reliable way to directly measure authors’ intended actions while interacting with a writing application. However, given the perceived link in our survey responses between phases of writing and document lifespan, in this section, we examine the temporal stages of document evolution through a different lens, by characterizing differences in user interaction logs of a commercial writing application. Our main goal with this quantitative analysis is to verify our qualitative findings at Web-scale and through a record of actual users’ interactions as they are writing documents.

In particular, we are interested in investigating three main research questions corresponding to the first three themes emerged from our qualitative data:

- [RQ1] Do authors interact with the writing application differently at different stages of document lifetime?
- [RQ2] Are there noticeable differences in the likelihood of observing certain activities at different points during the document lifetime?
- [RQ3] Do different authors attend to a shared document with different intentions as manifested in their first activities?

4.1 Log Data Analysis

We analyzed a sample of the anonymous interaction logs from users of Microsoft Word over a three-month period from February 1 to April 30, 2019. We note that these interaction logs were available through an existing telemetry system. They were not specifically designed to understand user behavior but rather were designed to measure feature usage and performance for the corresponding writing application¹.

To create a dataset that is suitable for investigating our main research questions specified in Section 4, we collected a subset of document candidates in English from the US market that were created within a 2 week period from February 1 to February 15. For these documents we collected logs of user interactions with different functionalities available through the application UI.² Our sample consists of approximately 16 million document candidates and approximately 16.6 million unique writers who accessed these documents.

Note that the interaction logs are devoid of textual data and writer data. The only exception to this was a system generated ID assigned to an individual writer that served to uniquely map the set of interactions of that writer on a document. This ID was not linked to any other identifying user information at any point in time, ensuring that the researchers had no information that could be traced back to an individual user. These interaction logs are highly granular and include approximately 2000 unique commands. For a subset of these documents (96% of initial documents), we generated numeric statistics on document content (e.g., #words, #sentences, #charts, etc), logged as static snapshots at different time points. Finally, for a much smaller subset of documents (about 2% of all docs), we have access to some statistics on typing traces that includes timing information on the duration of typing by different users with respect to a document³. We do not have access to keystroke-level data or the actual characters that were typed [7].

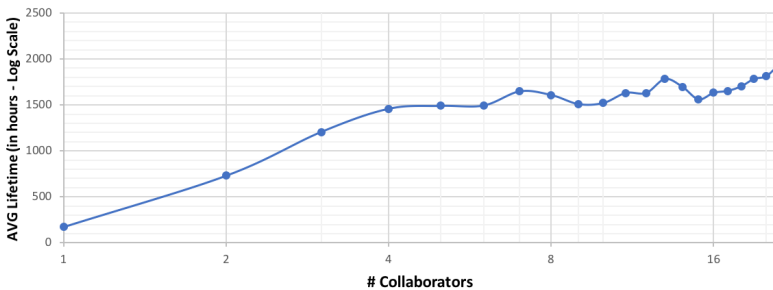


Fig. 1. Distribution of average document lifetime across different number of collaborators.

4.1.1 Lifetime of Documents. Because a document evolves over its lifetime, we defined the document lifetime as the time interval from its creation to the last user interaction with the document during our three-month window. We find that the majority of documents in our collection have a short lifetime (less than 24 hours - See Figure 2 - gold/left). In addition, there is a correlation between the number of collaborators involved in a document and the lifetime of the document.

¹Users consented to log collection as a part of the user agreement

²While our analysis is based on the documents that are written in Microsoft Word, we note that this modern, widely used word processing application provides standard functionalities including copy-paste, formatting, commenting, change tracking, etc., which are available through other modern writing applications as well.

³Since only a small subset of documents have typing statistics we only use this information when available.

Figure 1 shows that the lifetime of documents has a non-decreasing trend as the number of authors increases ($r = 0.78, n = 20, p < 0.0001$).

4.1.2 Pre-processing. We take two pre-processing steps to address limitations in our log data and prepare it for more robust analysis. Firstly, logged interactions are at a very fine-grained command level: we have over 2000 unique commands in our logs, each of which can be relatively sparse in its occurrence. To address this, we need to provide a mapping between low level user actions to higher level activities, which are interpretable and correspond to meaningful writing behavior. Initially, we looked to related past work [7] to develop a taxonomy of commands; however, on inspection, we found these past taxonomies (e.g. insertion/deletion or major/minor edits) insufficient to fully describe the potential categories of commands in our dataset. To develop a richer set of initial categories of activities, a subset of all commands were reviewed by three annotators independently to identify the impact of applying those commands to the writing environment. The impact of a command was investigated based on widely accepted operations that change the content of a document (e.g., adding or editing) as well as commands that have no impact on the document content (e.g., viewing, navigating, communicating, etc).

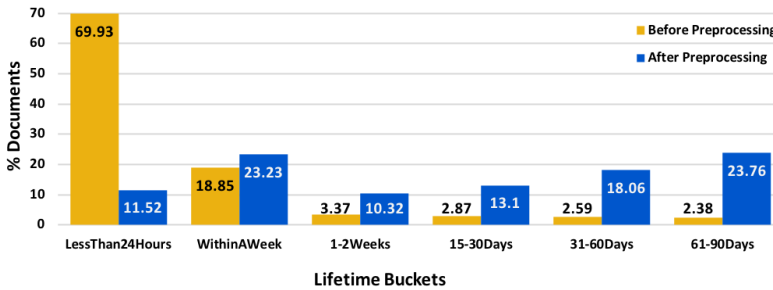


Fig. 2. Distributions of document lifetime before (gold - left) and after (blue - right) pre-processing

The annotators then engaged in an iterative vetting and refining process – described by Campbell et al. as *negotiated agreement* [10] – to group commands based on how they impact the document and the writing environment. For example, all commands that result in adding some content (text, shapes, citations, etc.) to the document are grouped together as *Adding*, while commands that are only used for navigating throughout the document (e.g., NextHeader, GoToPreviousPage, etc.) are grouped into the *Navigation* category. Initial annotator agreement was moderate [10], at 67%⁴ majority agreement between the three annotators. To enhance agreement through negotiation [10], command categories were then revised iteratively for this subset of commands until all conflicts were resolved and the agreement was 100%. This resulted in a finalized schema for grouping commands which was used by one of the authors to complete the groupings for the remaining commands in our logs while all authors would meet periodically to check for consistency. This analysis resulted in a set of 27 command categories. These categories were then classified into 10 higher level categories that more closely correspond to writing activities. Figure 3 shows a subset of the higher levels of this taxonomy. These high level groupings are sufficiently general and are supported by many modern writing tools. For more details and examples of the categories of commands in our dataset please see Appendix A.

The second pre-processing step is used to filter out all documents with a single author throughout their lifetime. While writers in individually authored documents do progress through a similar yet

⁴Note the use of percentages to describe inter-coder agreement, as per Campbell et al. [10].

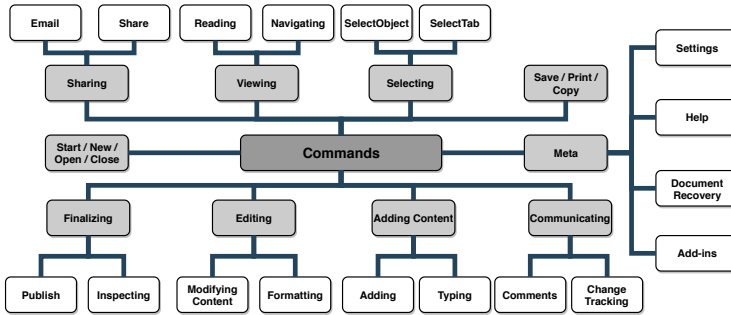


Fig. 3. Taxonomy of writing commands in our dataset. Note that only the top two levels are shown.

simplified subset of writing activities (as characterized by Lowry et al. [36]), we are focused on characterizing stages of document evolution where multiple authors write collaboratively. Thus we limit most of our analyses to documents that log interactions from 2 to 10 users⁵. Note that our methodology is, in general, agnostic to the number of users, and throughout the rest of the paper we will show some results that equally apply to individually authored documents.

Finally, in order to reduce noise in the dataset, we eliminate documents that have a lifetime shorter than 1 hour. Our final dataset comprises approximately 170k documents where the majority has 2 or 3 collaborators involved (78% and 12% respectively). The lifetime of documents in this final set are higher and more evenly distributed overall (Figure 2 - blue/right).

4.1.3 Analysis. In our quantitative analysis we focus on quantifying how authors interact with the writing application over temporal stages of a document lifetime. In order to facilitate the analysis of interactions with the writing application over time, we divide the lifetime of a document into equally-sized segments that correspond to 10-percent buckets of the document’s lifetime. We note that a linear segmentation of the document’s lifetime does not imply writing phases can be discretized as fixed-size, non-overlapping buckets. Rather, the goal of our quantitative analysis is to understand changes in the patterns of interactions with the application over time. Past research coupled with our qualitative findings suggest that authors perceive a temporal ordering to certain phases of writing (e.g. planning, drafting, copy-editing) and associate some writing activities to early stages of document lifetime and other activities to later stages of document evolution. Overall we found that writers distribute their writing activities differently over stages of document authoring both in terms of magnitude as well as types of these activities. Further, through an analysis of the distribution of authors’ first activities we highlight differences in the authors’ modes of writing. The following three subsections describe these differences in detail.

4.2 Magnitude of Writing Activities Over Time

We first characterize a document’s temporal stages by tracking the magnitude of author interactions with the writing application over a document’s lifetime, hereinafter referred to as Contributions Across Time (CAT). To this end, we segment the lifetime of each document into 10 equally sized buckets and look at the magnitude of interactions in each of these segments.

Figure 4 shows these CAT distributions for different types of documents. These activities are normalized by the total frequency of activities at the end of the lifetime of the documents.⁶ We can

⁵Our survey confirms (see Section 3.2) that 98% of collaboratively authored documents have no more than 10 users.

⁶The x-axis corresponds to the 10 buckets over the lifespan of the document plus two additional buckets (pre and post) for before the document is created (0-operations) and after document editing ends (0-operations), yielding 12 points in Figure-4.

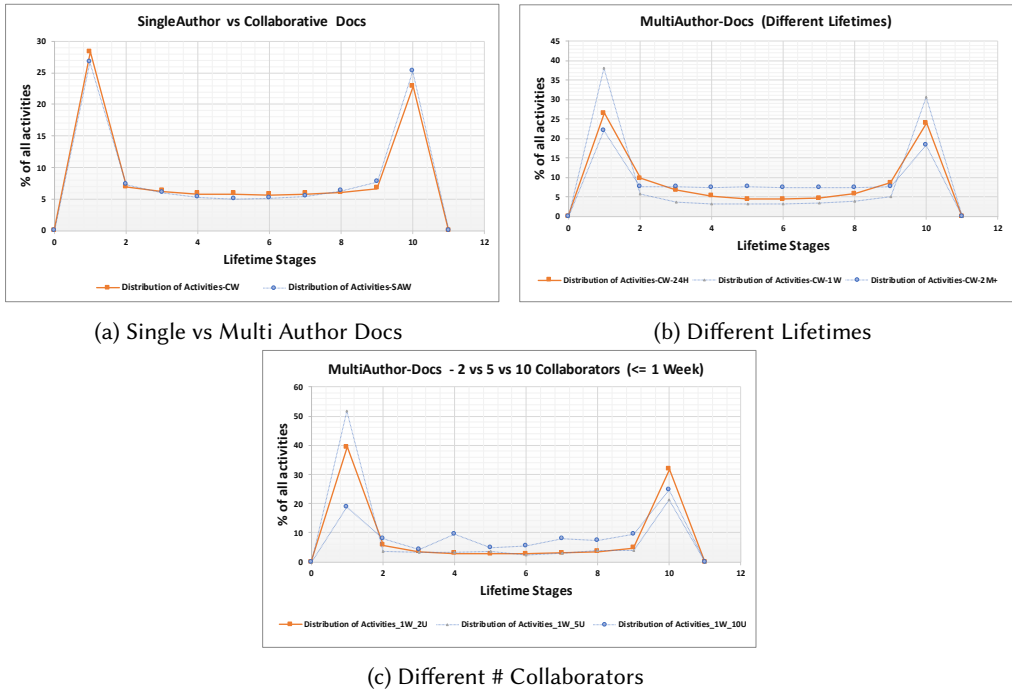


Fig. 4. Distribution of frequency of writing activities over stages of document evolution (a) single vs multi author documents, (b) multi author documents with different lifetimes, and (c) multi author documents with different #collaborators.

see an interesting pattern where there is a high level of activities at the very early and very late stages of document writing, while the middle stages show a much lower level of interaction with the writing application overall.

Interestingly, even when we move away from single-author documents to multi-author documents with different numbers of collaborators (See Figure 4a), the overall pattern does not change. We also experimented with documents of different lifetimes (i.e., short [between 1 and 24 hours]; medium [within a week]; long [more than 2 months] - Figure 4b), as well as different number of collaborators with the same lifetime (Figure 4c) and noticed the same pattern persisted throughout.

The progression of these writing activities over the document lifetime is interesting; there is a notable difference at early and later stages of document authoring that could be attributed to the task management aspects of document writing and how authors distribute the workload over the course of the writing process. Another implication is in predicting the lifetime of documents by observing the changes in the rate of interactions between the users of a writing application and the writing environment. We revisit the problem of predicting the document's lifetime in Section 5.

Please note that these CAT distributions can only be observed at the end of the document lifetime where we can compute the percentage of activities at different stages of the document evolution. That is, throughout the document evolution we can only capture the rate at which the interaction level is changing and compare it against previous snapshots of the document.

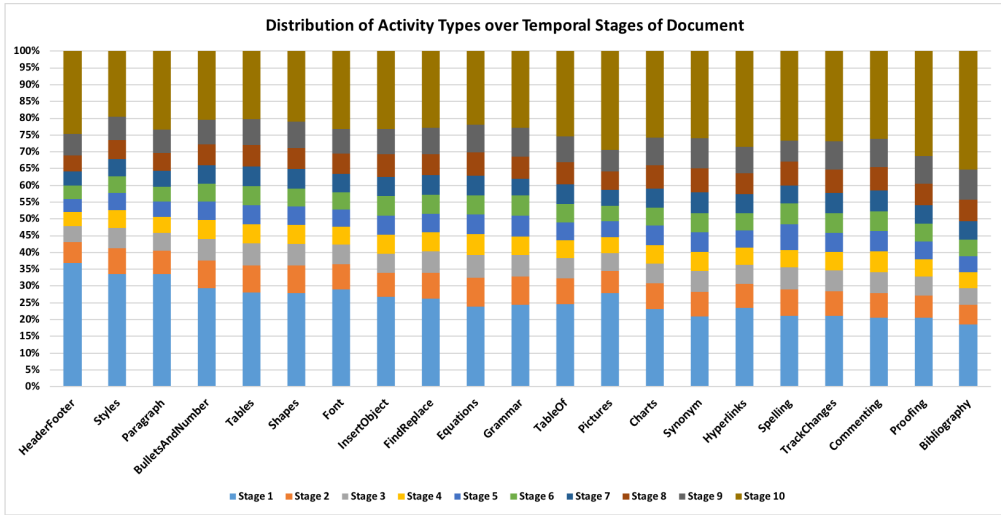


Fig. 5. Distribution of types of writing activities over stages of document evolution.

4.3 Types of Activities in Early vs Late Stages

The next research question we investigated was whether our qualitative observations regarding the types of activities that writers perform in earlier versus later stages can be observed in our large scale dataset of user interactions. Recall that based on our survey responses, outlining, structuring the document and drafting were among the early activities, while proofing, collaborative activities, referencing were identified as later stage activities.

Reusing our lifetime segmentation of authors' contributions we looked more closely at the distribution of different types of writing activities across these 10 segments. A subset of activities that correspond to the ones mentioned by our survey participants was used for this analysis. Please note that this set of activities correspond directly to the ways that different functionalities are grouped by the writing application UI. These groupings provide a different categorization of users commands that is complementary to the categories derived iteratively and exhaustively from our commands as discussed in Section 4.1.2. Figure 5 demonstrates the distribution of these activities over these 10 stages of document evolution. These activities are ordered from left to right based on the difference between their shares in Stages 1 and 10 to reflect how early they are likely to be observed during the document evolution.

The distribution of each of these activities over the 10 stages of document lifetime is normalized by their overall frequency. Note that for each of the activity types, stages 1 and 10 are the most prominent stages in which they occur, which reflects the results from the CAT plots. However, when comparing the likelihood of a writing activity between the first and the last stage we can see variations across activities. For example, activities such as interacting with HeaderFooter, Styles, Paragraph, and BulletsAndNumbers are more likely to be observed during the earliest stage of document authoring than the latest stages (e.g., 37% vs 24% for HeaderFooter; 34% vs 23% for styles and paragraphs). These activities correspond to setting up the overall structure of the document, outlining and early stages of drafting which involves creating and modifying lists and working them into paragraphs. A variety of advanced objects including Tables, Shapes, Equations and Charts are added and manipulated throughout the middle stages, with Tables and Shapes preceding Equations and Charts.

Among the later stage activities we can see improving the writing style (e.g. finding Synonyms) is more likely to be a later stage activity (26% vs 20%). Collaborative activities such as TrackChanges and Commenting are also likely to happen towards the end of a document's lifetime. Interestingly interacting with Proofing and Spelling functionalities are 11% and 6% more likely to happen at the latest stage, while Grammar related interactions seem to be middle stage activities. Finally, once we examine specific proofing commands we see some evidence of *copy-editing mode* that seems to be more likely in later stages. For example, 28% of all occurrences of NextMisspelling command, representing an author sequentially navigating through spelling suggestions by the writing tool, was observed at the latest stage (vs 22% during the first stage).

4.4 Distribution of First Activities

For each document in our dataset we also looked at how activities differ between authors that start working on the document at an earlier or later stage. For this purpose, we ordered the authors of a shared document by the time of their first activity in the document and we collected the type of their first activity as well as the set of all their activities. Contrasting the distribution of activities by the first and second authors of a shared document we do not see a high variation. Overall, Formatting, Adding and Modifying Content accounts for 22%, 15% and 15% for both first and second authors. The main two distinctions are for Typing and Initiating Communication (e.g., adding a new comment) that accounts for 5% and 0.7% of second author's overall activities but only to 2% and 0.4% of first author's overall activities.

Once we shift our focus to the first activities of the first and second authors at the time they first attend to a shared document, we start noticing differences in their intent and modes of writing. Our log based analysis indicated that the second authors are 7× more likely to start their interaction with typing and 2.5× more likely to start with adding non textual content (e.g., advanced objects) compared with the first authors (14% vs 2% and 10% vs 4% respectively). A similar trend is observed with second authors being two times more likely to start with editing activities (i.e., modifying content and formatting - 10% vs 5%). Finally, not only do second authors initiate communication more than the first authors throughout the document lifetime, they are also three times more likely to attend to a shared document for the first time with a communication intent (i.e., commenting and change tracking activities).

4.5 Summary of Findings

From our qualitative data, we find that authors have a perception of stages or phases in the writing process and that these stages represent an ordering in the creation of a document – from ideation through creation to refinement. Our quantitative data reflects this partially ordered approach to document creation through changes in both the magnitude of and types of interactions that occur over a document's lifetime. Further, authors in our qualitative study welcome the idea of stage-appropriate writing assistance. They highlight varied goals and activities to which assistance can be tailored at different stages. Alongside the potential for stage-specific assistance, previous work has noted the importance of stage-appropriate behavior during collaborative tasks [27], and therefore stage-aware tools have the potential to foster improved collaboration.

In summary, implications of the synthesis of our qualitative and quantitative data is that (1) there exists a perceived benefit to stage-specific writing assistance, (2) we observe different activities at different stages of the document lifespan, and (3) stage awareness mechanisms can communicate to collaborators the current state of the document, which can aid in on-boarding of new collaborators and in appropriate behaviors during collaboration [27]. However, to effectively design interfaces to support stage-specific writing tasks and to communicate stages, one open question remains: Can we predict the stage of document completion given observation of user actions? To explore

this last research question, we present a study in the next section assessing the feasibility of stage-identification.

5 PREDICTING THE STAGE OF A DOCUMENT

Our qualitative analysis has shown that users would appreciate a more situated writing assistant that takes the stage of a document into account. To this end, a system that could predict the stage of a document automatically would be of high value. From our log analysis we also know that document evolution is reflected in the signals we find in user interaction logs. In this section, we conduct a preliminary investigation into how useful log signals can be for predicting a document's stage. Specifically, given a telemetric snapshot of a document – at a uniformly and randomly sampled point in its lifetime – we attempt to predict the quartile of the document's lifetime to which the snapshot belongs. We chose to represent the document lifetime in only four stages for this preliminary investigation to simplify the task. Note that despite this simplification, this is a non-trivial prediction task, since the lifetimes of individual documents vary greatly in our data, as do the magnitudes and distributions of their logs. We note that being able to effectively identify boundaries between phases of writing or detect transitions between these stages would be a more interesting goal. However, to the best of our knowledge the most successful attempts at inferring phases of writing (e.g. planning versus drafting versus reviewing) was achieved by analyzing keystroke logs and used the time intervals between different keystrokes to infer changes in the users' cognitive load and correlate them with phases of writing they are transitioning through (e.g. [11]). Given issues of data privacy, our goal was to leverage interaction logs that are devoid of any text and keystroke information.

We sample the longitudinal logs of 60k documents from our collection of 170k documents (see Section 4.1.2). We use 50k of these documents for training and reserve the rest for testing. For each document, we sample five distinct *snapshots*, each of which is an aggregate of telemetry signals starting with the document's creation and ending at a uniformly and randomly sampled point t in the document's lifetime. We represent t as a float number in the range $[0:1]$ where 0 corresponds to the beginning of the lifetime and 1 corresponds to the end of the lifetime. Each document's lifetime is divided into 4 equally sized buckets with the snapshot boundaries 0, 0.25, 0.75, 1.0. Given this, any snapshot can be labeled by its corresponding quartile by mapping the float number to the quarter bucket it belongs to. Given this data we train and evaluate four one-versus-all boosted decision forest classifiers⁷ [23] – one each for the four lifetime quartile labels in our dataset. The model is trained with telemetry features from the lifespan of the document up to the timestamp the snapshot is sampled from. In other words, the model has no access to information beyond this point or any features from the entire lifespan.

Our model incorporates features from the user interaction logs that can broadly be grouped into the following five classes: (1) commands (2) command categories (3) high-level command categories (4) advanced functionalities (5) content statistics (see Sections 4.1 and 4.3 as well as Appendix B for more detail). The first four feature classes are normalized – that is, individual feature values in each class represent the percentage of contribution of a feature within its class, at a given snapshot of a document. For example, the command category *Navigation* will have as its value the percentage of navigation commands amongst all command categories observed in the snapshot. We supplement these signals with two sets of features designed to reflect the balance of collaboration in a document. Our qualitative data suggests that not all stages of document authoring are collaborative and therefore these features can provide some signal in predicting the stage of a

⁷We experimented with several types of classifiers, including logistic regression, SVM, and multi-layer perceptron, and found boosted decision forest to perform the best.

document. These collaborative features measure the contribution of different collaborators to a shared document across 4 dimensions: adding content, editing, communicating, and finalizing (see Figure 3).

The first metric, Contribution Balance Score (CBS) broadly considers any signal logged to one of the 4 categories as *content contribution*, and computes the share of *content contributing* activity performed by each collaborator of a document. The second metric, Role Balance Score (RBS), separates contributions into one of the 4 dimensions of activity, in order to determine whether collaborators took on dedicated roles (e.g., Writer, Editor, etc.) over the course of document authoring. We adapted the method used in [44] to measure the contribution of different collaborators using their share of basic text-based operations to our interaction logs and computed the CBS and RBS features as follows:

$$CBS(d_i) = \frac{1}{\log(u)} \sum_{j=1}^u CC(ij) \cdot \log\left(\frac{1}{CC(ij)}\right) \quad (1)$$

$$RBS(d_i, a_k) = \frac{1}{\log(u)} \sum_{j=1}^u AC(ijk) \cdot \log\left(\frac{1}{AC(ijk)}\right) \quad (2)$$

where u is the number of authors observed in a document snapshot, CC the number of general content contribution signals, and AC the number of specific activity contribution signals to one of the 4 dimensions $a_k \in \{\text{Adding, Editing, Communicating, Finalizing}\}$.

We compare our model against a baseline that uses the sole feature of elapsed time since document creation. Note that this signal provides a strong baseline, especially for the last quartile of a document's lifetime. In particular, given that our data longitudinally tracks documents over a maximum period of three months, if the elapsed time exceeds, e.g., two months, it is highly likely that the snapshot under consideration belongs to the 4th quartile label.

Comparing the baseline against a model trained on all features, we note that the baseline achieves a macro-averaged accuracy of 41.77% and the model with all features outperforms this score with 44.80%. Differences are significant at $p < 0.01$ (Approximate Randomization Test [15]). Figure 6 shows the precision-recall characteristics of our models for each of the four predicted stages. Precision and recall are computed per predicted quartile to illustrate differences in the model performance across quartiles. The chart in Figure 6 contains a total of 8 precision recall curves, for the combination of two settings (using lifetime features only versus using all features) and four labels (1st, 2nd, 3rd, and 4th quartile). Overall, the results demonstrate that user interaction logs do contribute positively towards the prediction of a document's temporal stage. Interestingly, the most significant gains are observed for the first and last quartile labels in documents' lifetimes. These seem to correspond to the CAT distributions that we observe for magnitudes of writing activities over a document's life span (see Section 4.2, Figure 4). We hypothesize that when user interaction logs are recorded in volume they can provide a valuable additional signal to predict the temporal stage of documents.

We also perform manual feature analysis to identify the most predictive features across classifiers. Not surprisingly, elapsed lifetime is at the top of the list, followed by the CBS and RBS collaboration features and then a mix of the other signals, from individual commands to activities and activity categories.

In summary, our first attempt at modeling a document's temporal stage has shown promising results with signal from user interaction logs. This is nevertheless, a preliminary foray into a task that could almost certainly benefit from more complex modeling approaches. For example, a model capable of capturing sequential dependencies (such as an LSTM) will likely be better able to leverage

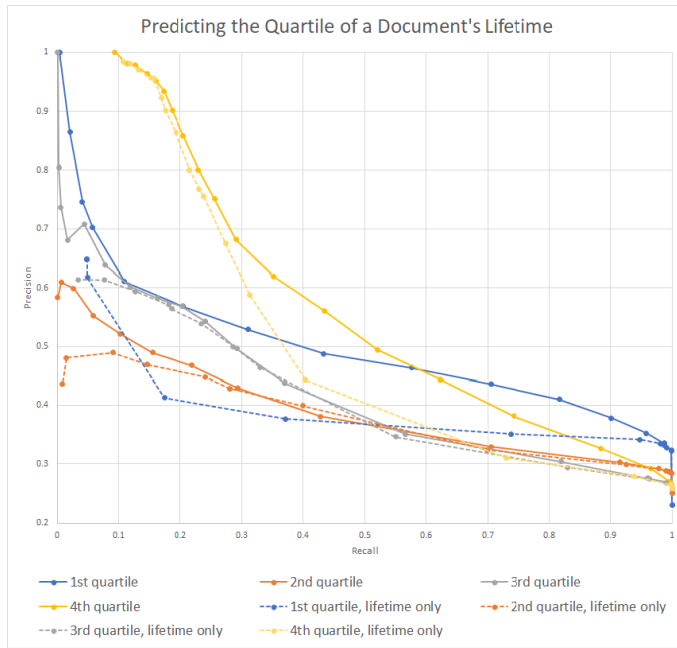


Fig. 6. Precision and recall curves for predicting the quartile of a document’s lifetime. Curves are shown per model, per quartile. Dashed lines = time elapsed baseline, Solid line = learned model.

the temporal changes in a document’s evolution than our current approach, which aggregates the document’s history into a snapshot. We leave these efforts to future work.

6 DISCUSSION AND DESIGN IMPLICATIONS

Previous work [19, 21, 22] on supporting authors who write documents individually or with others, have characterized the process of writing as a series of stages – planning, drafting, reviewing and finalizing – and that the decomposition of these stages can be leveraged by the authors as a means of scaffolding their work [25], reducing their cognitive load [26, 34], and supporting more effective coordination and collaboration [55]. While techniques to support collaborative writing (e.g., microtasking [55], effective communication and coordination mechanisms [24], or automated spelling and grammar support in modern writing tools such as Google Docs or Microsoft Word) appear promising, we would argue that any technique that seeks to leverage stages of writing requires, *a priori*, an understanding of whether authors perceive stages in writing, of whether those stages generalize across a broad range of writing tasks, and of whether or not the stages of writing can be automatically inferred.

In our work, we specifically target this notion of staged document writing. The three previous sections provide evidence that authors perceive stages and associate specific activities to them, that the magnitude of writing activities in both collaborative and individual writing follows a similar pattern in these stages, and that these stages can be inferred from privacy-preserving data logs of document interaction. Together, these results provide support for the feasibility of stage-aware writing assistance. More specifically, our qualitative findings provide a range of insights about different characteristics of temporal stages of document authoring: Earlier stages of writing were described as more focused and less structured, suggesting the necessity of minimizing system

interruption and affording more control to the user; Later stages, on the other hand were perceived as more mature and more stable, where the system can be more proactively involved in providing local and global feedback and suggestions to improve the overall document quality.

One significant and interesting observation from our log analysis was the similarity in ways that authors in both individually authored and multi-authored documents distribute their work load across stages of document evolution (see Section 4.2). While initially unexpected, this similarity is important in the design of document writing assistants because, if these intelligent assistants are destined to guide author behavior through a determination of stages, then regardless of whether one is writing alone or collaborating, the overall goal of any intelligent writing assistant would be to support the writer(s), making them a default integration into digital writing platforms. In hindsight it may be unsurprising that individual and group document editing share some similar aspects: some past views [48] of the stages of document creation have specified stages of pre-writing, writing, and re-writing, i.e., that individual authorship is a collaboration with oneself through the mechanisms of communication and distributed cognition [64]. Another important implication of this observation is in the ability to infer the stage of a document evolution by simply tracking the rates of interactions between the authors and the writing tool.

In collaborative scenarios, our analysis of distribution of first activities (Section 4.4) provided quantitative evidence for our qualitative characterization of modes of writing and indicated that first and second authors of a shared document enter this joint workspace with noticeably different intents and modes of writing. For example, we find that second authors of a shared document are 3X more likely to appear on the shared space with communication intent and 7X more likely to start their collaboration with typing new content. This finding has implications for designing more customized toolbars for the writing applications that can match the intended mode of writing (e.g. reviewing versus reading mode), which helps simplify the user's interaction with the tool and minimizes accidental changes.

Every study comes with different limitations. In our work, we also acknowledge at least two potential confounds: (1) our survey population is limited to information workers in a large technological company. We made an attempt to mitigate this limitation by recruiting across many different departments with a diverse set of job roles and demographics; our log analysis seeks to address this limitation by considering a much broader population of authors. (2) the prediction accuracy of our trained model is currently at an acceptable level, but it can be improved by leveraging candidate sequence learning methods (e.g. LSTM) which can capture sequential dependencies we observed in our data (i.e. CAT distributions). We still argue that our preliminary attempt towards inferring the stage of a document can serve as a promising starting point for designing stage-aware writing assistants.

While our work on linking the stages of document evolution to writing activities is important to the design of digital collaborative document authoring assistants, it is clear that much work remains to be done to make these tools a reality. For example, while we showed that the current stage of a document under development can be inferred and utilized to provide timely writing assistance that is tailored to this stage, how we realize tools to support stage aware writing assistance remains an open problem. For example, should tools automatically infer writing stage and provide varied assistance based on this inference? Or should users pre-specify writing stage? Intelligent adaptation of an interface is something that researchers have extensively probed [9, 31], and the reluctance to automatically trust adaptive tools [9] was echoed in our initial survey, where over one third of participants were reluctant to have systems infer writing stage. How to incorporate tool adaptivity into any interface that leverages inference to improve productivity remains an on-going area of research. On the other hand, echoing past work in collaborative systems [27] and in collaborative writing [55], appropriate writing etiquette is also an important part of effective collaboration.

Therefore, another issue to probe with future work is that awareness of document evolution does not, necessarily, need to adapt the interface to different phases of writing, but can instead be leveraged to foster more effective collaboration through enhanced task awareness [57]. In this way, instead of enforcing modified behavior on the writer, these tools can instead, serve to promote improved context awareness for collaborators in computer mediated collaborative writing activities.

7 CONCLUSION

In this paper, we have presented the first large-scale study of the stages of writing, specifically as they relate to the temporal evolution of documents. Using a mixed-methods approach that includes qualitative, quantitative, and predictive analyses, this paper presents a comprehensive picture of how authors perceive stages of document authoring, how these perceptions correlate with actual user interaction logs, and how these logs can, in turn, be used to predict the temporal stage of a document. More specifically, a survey of 183 participants concluded that writers do have mental models of definite stages while authoring documents; moreover, they associate different goals and types of activities with each of these stages. We support the findings of this qualitative survey with a quantitative analysis of writers' longitudinal user interaction logs on a popular Web-scale document authoring platform. These logs are notable in two important ways. First, they cover a set of several million documents – thereby surpassing the scale of any previous study of document evolution by many orders of magnitude. Secondly, these logs contain no actual text – thereby overcoming some of the assumptions about document topic or quality embedded in previous approaches. Our findings confirm that writers' activities are distributed differently through different stages in a document's evolution, both in magnitude and in the types of activities.

Finally, as a first step towards enabling stage-aware intelligent writing assistance, we conduct a predictive experiment to model the temporal stage of a document by featurizing the user interaction logs. The results of our experiments suggest that these logs provide predictive power over a strong baseline. These findings show promise for future work on intelligent writing assistance.

ACKNOWLEDGMENTS

We would like to thank Shamsi Iqbal, Justin Cranshaw, Mark Encarnación, Michael Bentley, and Tomasz Religa for their valuable input throughout this work. We also thank the reviewers for their feedback; and finally, all of our study participants for their help in making this research possible.

REFERENCES

- [1] Gediminas Adomavicius and Alexander Tuzhilin. 2011. Context-aware recommender systems. In *Recommender systems handbook*. Springer, 217–253.
- [2] Ablimit Aji, Yu Wang, Eugene Agichtein, and Evgeniy Gabrilovich. 2010. Using the past to score the present: Extending term weighting models through revision history analysis. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 629–638.
- [3] Denis Alamargot and Lucile Chanquoy. 2001. *Through the models of writing: with commentaries by Ronald T. Kellogg & John R. Hayes*. Vol. 9. Springer Science & Business Media.
- [4] Denis Alamargot and Michel Fayol. 2009. Modelling the development of written composition. *The SAGE handbook of writing development (2009)*, 23–47.
- [5] Ronald M Baecker, Dimitrios Nastos, Ilona R Posner, and Kelly L Mawby. 1995. The user-centred iterative design of collaborative writing software. In *Readings in Human-Computer Interaction*. Elsevier, 775–782.
- [6] Jeremy Birnholtz and Steven Ibara. 2012. Tracking changes in collaborative writing: edits, visibility and group maintenance. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 809–818.
- [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] Joan Borràs, Antonio Moreno, and Aida Valls. 2014. Intelligent tourism recommender systems: A survey. *Expert Systems with Applications* 41, 16 (2014), 7370–7389.
- [9] Andrea Bunt, Cristina Conati, and Joanna McGrenere. 2007. Supporting interface customization using a mixed-initiative approach. In *Proceedings of the 12th international conference on Intelligent user interfaces*. 92–101.
- [10] John L. Campbell, Charles Quincy, Jordan Osserman, and Ove K. Pedersen. 2013. Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement. *Sociological Methods & Research* 42, 3 (2013), 294–320. <https://doi.org/10.1177/0049124113500475> arXiv:<https://doi.org/10.1177/0049124113500475>
- [11] Rianne Conijn, Jens Roeser, and Menno van Zaenen. 2019. Understanding the keystroke log: the effect of writing task on keystroke features. *Reading and Writing* (2019), 1–22.
- [12] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.
- [13] John W Creswell. 2013. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- [14] Gabriele d’Angelo, Angelo Di Iorio, and Stefano Zacchiroli. 2018. Spacetime Characterization of Real-Time Collaborative Editing. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.
- [15] Eugene S Edgington. 1969. Approximate randomization tests. *The Journal of Psychology* 72, 2 (1969), 143–149.
- [16] Clarence A Ellis, Simon J Gibbs, and Gail Rein. 1991. Groupware: some issues and experiences. *Commun. ACM* 34, 1 (1991), 39–58.
- [17] Mica R Endsley. 2017. Toward a theory of situation awareness in dynamic systems. In *Situational Awareness*. Routledge, 9–42.
- [18] Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College composition and communication* 32, 4 (1981), 400–414.
- [19] Jill Fitzgerald. 1992. Towards knowledge in writing: Illustrations from revision studies. (1992).
- [20] Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication* 32, 4 (1981), 365–387.
- [21] Linda Flower and John R Hayes. 1984. Images, plans, and prose: The representation of meaning in writing. *Written communication* 1, 1 (1984), 120–160.
- [22] Sarah Warshauer Freedman et al. 1987. Research in Writing: Past, Present, and Future. (1987).
- [23] Jerome H. Friedman. 2000. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29 (2000), 1189–1232.
- [24] Jolene Galegher and Robert E Kraut. 1994. Computer-mediated communication for intellectual teamwork: An experiment in group writing. *Information systems research* 5, 2 (1994), 110–138.
- [25] Stephen Graham and Karen Harris. 2012. The role of strategies, knowledge, will, and skills in a 30-year program of writing research (with Homage to Hayes, Fayol, and Boscolo). In *Past, present, and future contributions of cognitive writing research to cognitive psychology*. Taylor and Francis, 177–196.
- [26] Nick Greer, Jaime Teevan, and Shamsi T Iqbal. 2016. *An introduction to technological support for writing*. Technical Report. Technical Report. Microsoft Research Tech Report MSR-TR-2016-001.
- [27] Carl Gutwin and Saul Greenberg. 2002. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work (CSCW)* 11, 3-4 (2002), 411–446.
- [28] Gunnar Harboe and Elaine M Huang. 2015. Real-world affinity diagramming practices: Bridging the paper-digital gap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 95–104.

- [29] Negar Hariri, Bamshad Mobasher, and Robin Burke. 2013. Query-driven context aware recommendation. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 9–16.
- [30] Karen Holtzblatt and Hugh Beyer. 1997. *Contextual design: defining customer-centered systems*. Elsevier.
- [31] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.
- [32] Eric Horvitz, Jack Breese, David Heckerman, David Hovel, and Koos Rommelse. 1998. The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 256–265.
- [33] Claudia-Lavinia Ignat and Moira C Norrie. 2008. Multi-level editing of hierarchical documents. *Computer Supported Cooperative Work (CSCW)* 17, 5-6 (2008), 423–468.
- [34] Ronald T Kellogg. 1988. Attentional overload and writing performance: Effects of rough draft and outline strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14, 2 (1988), 355.
- [35] Teresa Limpo and Rui A Alves. 2013. Teaching planning or sentence-combining strategies: Effective SRSD interventions at different levels of written composition. *Contemporary Educational Psychology* 38, 4 (2013), 328–341.
- [36] Paul Benjamin Lowry, Aaron Curtis, and Michelle René Lowry. 2004. Building a taxonomy and nomenclature of collaborative writing to improve interdisciplinary research and practice. *The Journal of Business Communication (1973)* 41, 1 (2004), 66–99.
- [37] Aurélien Max and Guillaume Wisniewski. 2010. Mining Naturally-occurring Corrections and Paraphrases from Wikipedia’s Revision History.. In *LREC*.
- [38] Alex Mitchell. 1996. *Communication and shared understanding in collaborative writing*. University of Toronto Toronto.
- [39] Donald M Murray. 1980. *Writing as process: How writing finds its own meaning*. Technical Report. TR Donovan & BW McClelland (Eds.), Eight approaches to teaching composition.
- [40] Sylvie Noël and Jean-Marc Robert. 2004. Empirical study on collaborative writing: What do co-authors do, use, and like? *Computer Supported Cooperative Work (CSCW)* 13, 1 (2004), 63–89.
- [41] Martin Nystrand. 1989. A social-interactive model of writing. *Written communication* 6, 1 (1989), 66–85.
- [42] Judith S Olson, Gary M Olson, Marianne Storrøsten, and Mark Carter. 1993. Groupwork close up: A comparison of the group design process with and without a simple group editor. *ACM Transactions on Information Systems (TOIS)* 11, 4 (1993), 321–348.
- [43] 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), 4.
- [44] Adrian Pace, Louis Baligand, Stian Haklev, Jennifer Kaitlyn Olsen, Nore de Grez, and Bram De Wever. 2018. Quantifying Collaboration in Synchronous Document Editing. In *Proceedings of the 3rd Swiss Text Analytics Conference (Swiss-Text 2018)*.
- [45] Ilona R Posner and Ronald M Baecker. 1992. How people write together (groupware). In *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, Vol. 4. IEEE, 127–138.
- [46] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. ACM, 117–126.
- [47] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1650–1659.
- [48] D Gordon Rohman. 1965. Pre-writing the stage of discovery in the writing process. *College composition and communication* 16, 2 (1965), 106–112.
- [49] Bahareh Sarrafzadeh, Ahmed Hassan Awadallah, Christopher H Lin, Chia-Jung Lee, Milad Shokouhi, and Susan T Dumais. 2019. Characterizing and predicting email deferral behavior. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 627–635.
- [50] Yue Shi, Alexandros Karatzoglou, Linas Baltrunas, Martha Larson, and Alan Hanjalic. 2014. Cars2: Learning context-aware representations for context-aware recommendations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 291–300.
- [51] Nancy Sommers. 1994. Revision strategies of student writers and experienced adult writers. *Landmark essays on writing process* (1994), 75–84.
- [52] Fred Steck, Mahlet Lewis, Theresa Richter, Paul Deane, Amanda Roth, Anna Litz, and Vishal Goswami. 2018. Behavioral Differences Between Retyping, Drafting, and Editing: A Writing Process Analysis. (2018).
- [53] Lucy A Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.
- [54] Susan G. Tammaro, Jane N Mosier, Nancy C Goodwin, and Gabriel Spitz. 1997. Collaborative writing is hard to support: A field study of collaborative writing. *Computer Supported Cooperative Work (CSCW)* 6, 1 (1997), 19–51.

- [55] Jaime Teevan, Shamsi T Iqbal, and Curtis Von Veh. 2016. Supporting collaborative writing with microtasks. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 2657–2668.
- [56] Marion Tillema, Huub Van Den Bergh, Gert Rijlaarsdam, and Ted Sanders. 2011. Relating self reports of writing behaviour and online task execution using a temporal model. *Metacognition and Learning* 6, 3 (2011), 229–253.
- [57] April Yi Wang, Anant Mittal, Christopher Brooks, and Steve Oney. 2019. How Data Scientists Use Computational Notebooks for Real-Time Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
- [58] 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.
- [59] Cuiqin Xu and Yanren Ding. 2014. An exploratory study of pauses in computer-assisted EFL writing. *Language Learning & Technology* 18, 3 (2014), 80–96.
- [60] Elif Yamangil and Rani Nelken. 2008. Mining Wikipedia revision histories for improving sentence compression. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, 137–140.
- [61] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2000–2010.
- [62] Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 365–368.
- [63] Soobin Yim, Dakuo Wang, Judith Olson, Viet Vu, and Mark Warschauer. 2017. Synchronous Collaborative Writing in the Classroom: Undergraduates’ Collaboration Practices and their Impact on Writing Style, Quality, and Quantity. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 468–479.
- [64] Vivian Zamel. 1982. Writing: The process of discovering meaning. *TESOL quarterly* 16, 2 (1982), 195–209.
- [65] Fabio Massimo Zanzotto and Marco Pennacchiotti. 2010. Expanding textual entailment corpora from wikipedia using co-training. In *Proceedings of the 2nd Workshop on The People’s Web Meets NLP: Collaboratively Constructed Semantic Resources*. 28–36.
- [66] Torsten Zesch. 2012. Measuring contextual fitness using error contexts extracted from the wikipedia revision history. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 529–538.
- [67] Qian Zhao, Paul N Bennett, Adam Fourney, Anne Loomis Thompson, Shane Williams, Adam D Troy, and Susan T Dumais. 2018. Calendar-aware proactive email recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 655–664.

A DESCRIPTION OF MAIN COMMAND CATEGORIES

Table 1. Description of Commands Categories and Sample Commands per Category.

Name	Description
Adding Content	Commands in this category add to the content of a document by either inserting an object or an element to an object (e.g. Tables, Shapes, etc) or text (e.g. by pasting or typing).
SubCategories	Sample Commands
Adding	InsertCitation InsertBibliography InsertFile Paste; PasteTextOnly; ... TableInsert
Typing	Typing is captured as system's events and we add these events to the list of commands that are logged based on the interaction with the tools UI to incorporate the periods of typing into our dataset
Editing	Commands in this category result in changes either to the content or the formatting / styling characteristics of existing content in the document.
SubCategories	Sample Commands
ModifyingContent	AutoCorrect ChartDelete Replace EditCitation JoinList
Formatting	listPosition SetTransparentColor Recolor BibCitationToText Bold
Viewing	Commands in this category do not result in any changes in the content or formatting of the document. These commands are tracking users navigational or viewing interaction with the writing environment.
SubCategories	Sample Commands
Navigating	NextField / PreviousField NextHeader / PreviousHeader GoToHeader / GoToFooter NextMisspelling
Reading	ViewReadingMode ReadingModePageview ExpandAllHeadings ToggleReadingMode
Viewing	BibliographyFilterLanguages DocInfoFileName ShowDocumentText ViewHeader

continues on next page

	Find / FindNext
Selecting	Similar to the Viewing category, commands in this category do not result in any changes in the content or formatting of the document. These commands are tracking users' selection of objects/text in the document as well as menu items and UI's toolbars.
SubCategories	Sample Commands
SelectObject	SelectTable / SelectRow / SelectCell SelectAll Pointer DrawSelectNext SelectionMode
SelectTab	RefPaneSelection TabReview / TabHome MergeToolbar EndOfLineExtend EquationScriptGallery
Sharing	Commands in this category capture different interactions with sharing the created document either via email or through writing application's built-in sharing features.
SubCategories	Sample Commands
Email	MergeToEMail Envelope / EnvelopeSend TabMailing MailMergeCreateList AddressBook
Share	Share FileShareMenu SendForReview SharedWithPersonaCanView
Communicating	Commands in this category capture different ways of collaboration and communication between co-authors of a shared document, either through comment based communication or tracking and responding to changes.
SubCategories	Sample Commands
Comments	NewComment DeleteComment ReplyToComment MarkCommentDone
ChangeTracking	NextRevision AcceptRevision RejectAllChangesShown RevisionsHighlightChanges AcceptConflict

continues on next page

Finalizing	Commands in this category are mainly applied to a completed document or as preparation for sharing or publishing. These commands include adding restrictions to the document or the audience who can access the file as well as different inspections that can ensure the document is now ready for sharing or publishing.
SubCategories	Sample Commands
Publish	DocEncryption MarkAsReadOnly FileFinalizeMenu RestrictFormatting
Inspecting	WordCompatChkr StatusHasPolicy DocInspector AccessibilityChecker DocProtectionManage
Start / New / Open / Close	These commands simply indicate a new document is created or the writing tool is launched.
SubCategories	Sample Commands
Start	OfficeStart
New	FileNewDialog New NewDefault
Open / Close	These commands simply indicate a document is opened or closed through the menu.
Save / Print / Copy	These commands simply indicate the user has saved or printed the document or copied some content from this document. These commands could potentially indicate some stage of (partial) completion during the document's lifetime.
SubCategories	Sample Commands
Save	Save / SaveAs / SaveAll SaveToWebSignIn DocExport ConvertDoc
Print	PrintPreviewAndPrint Print FileNewPrint ClosePreview
Copy	Copy CopyText WebCopyHyperlink
Meta	These commands include different functionalities of a modern writing environment that is beyond document creation, authoring or viewing.
SubCategories	Sample Commands
Settings	StatusBarConfigMenu Custom TrackChangesAdvancedOptions

continues on next page

	Ruler11 EquationLinearAll
Help	TabHelp Insights HelpContactUs WhatsNewRecentUpdates
DocRecovery	StatusDocRecovery ShowRepairs OpenDrp / ViewDrp
Macros	StatusRecordMacro PauseRecorder TabDeveloperTools Play
AddIns	OpenLinkedIn OfficeExtensionsGallery TabAddins

B OVERVIEW OF FEATURES USED FOR PREDICTION

Table 2. Feature Categories used for temporal stage of document evolution prediction.

Name	Description
Command Features	
See Appendix A	The lowest level of taxonomy described in Section 4.1
Command Category Features	
See Appendix A	The second level of taxonomy described in Section 4.1
High Level Command Category Features	
See Appendix A	The highest level of taxonomy described in Section 4.1
Advanced Functionality Features	
BulletsAndNumbering	num. lists + whether there is a list (i.e. binary feature)
Canvas	whether the drawing canvas is used
HeaderFooter	whether HeaderFooter is enabled
Shapes	num. shapes + whether there is a shape (i.e. binary feature)
Bibliography	whether Bibliography is inserted
Hyperlinks	num. hyperlinks + whether there is a hyperlink (i.e. binary feature)
...	These features correspond to the analysis discussed in Section 4.3
Collaboration Features	
# Collaborators	Collaborators are users who contributed to the document in at least one of adding, editing, communicating or finalizing
CBS	Contribution Balance Score reflects how balanced the contribution was across all collaborators regardless of the type of contribution.
RBS-adding	Whether all collaborators contributed equally to adding content
RBS-editing	Whether all collaborators contributed equally to editing
RBS-communicating	Whether all collaborators contributed equally to communicating
RBS-finalizing	Whether all collaborators contributed equally to finalizing
...	RBS scores can be calculated for each of command categories using the formula introduced in Section 5
Content Features	
PageCount	num. pages in the document at a given timestamp
SectionsCount	num. sections in the document at a given timestamp
ParagraphCount	num. paragraphs in the document at a given timestamp
LineCount	num. lines in the document at a given timestamp
WordCount	num. words in the document at a given timestamp
CharacterCount	num. characters in the document at a given timestamp
Lifetime Features	
TimeElapsed	amount of time in milliseconds that has passed since document creation
TimeElapsedBucket	A coarse grained bucketization of TimeElapsed as numDays, numWeeks, etc.