The Annotators' Perspective on Co-authoring with Structured Annotations

ABSTRACT

Research has shown that grouping related annotations together can help those who review an annotated document by reducing their workload and raising the accuracy of their reviewing. Less is known about the impact on users who create these structured annotations – the annotators. The goals of the research reported in this paper were: (1) to better understand current annotation creation practices, (2) to explore how structuring can be used by annotators, both the structuring process and resulting types of structure, and (3) to evaluate the impact on annotators of having to create structured annotations. We conducted three studies to address each of these goals and learned that structured annotations are perceived to be worth the additional workload and that the bottom-up grouping approach complements the top-down approach in describing relationships amongst annotations in a document.

Author Keywords

Collaborative writing/authoring, usability study, workflow, tagging, information organization

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI)

INTRODUCTION

In asynchronous collaborative writing, annotations play an important role as a communication medium among co-authors. Most word processing systems, however, only support simple annotations (basic edits and comments), forcing valuable communication among group members to take place outside the shared document, most often in the bodies of emails between co-authors to which versions of the document are attached and sent. This results in communication being disembodied from the document, causing unnecessary overhead and inefficiencies [23]. Because of this, co-authors often need to provide explicit navigation statements such as "see page 2, paragraph 3", or they copy and paste some referenced text from the document into email messages to establish context. This separation of artifacts means that valuable information can easily be disregarded or misplaced. In our experience, these difficulties can increase dramatically after only a few reviewing cycles.

To address the shortcomings of current annotation tools, Zheng et al. developed an annotation model that unifies all document-related communication together within the document: *single annotations* are anchored at specific places in the document, *general comments* are anchored to the document as a whole, and *structured annotations* are a grouping of one or more single annotations or general comments [26]. Structured annotations have hierarchical structure (groups within groups); the structuring is intended to communicate meta-information, i.e., to act as meta-comments, relative to a group of annotations. Zheng et al. evaluated the effects of structured annotation on users reviewing an annotated document (the "recipients"), and found efficiency and accuracy benefits when compared to unstructured meta-comments written in email messages. The effects of structured annotation on "annotators," those who create the annotations, have not yet been explored. This is the focus of the work reported here.

Our target population is distributed groups collaborating asynchronously during the editing and reviewing stages of co-authoring, potentially creating a large volume of annotations to communicate document-related information. Our research was conducted in three phases. In Phase I, we used an observational study to gain a better understanding of annotators' workflow with existing annotation tools (both digital tools and traditional pen on paper markup), neither of which provide any explicit support for structuring annotations. We sought to understand if structuring was provided as an option to annotators, what process and types of structures annotators would use. Zheng et al.'s work assumed a top-down approach [26], but the recent explosion of tagging [12] suggests that a bottom-up approach to information organization may be preferable or at least complementary. In order to explore these and other issues independent of the tools to create structured relationships, in Phase II we conducted an exploratory study with a paper prototype. In Phase III, we conducted a

controlled experiment with an interactive prototype that supports structured annotation, including tagging, and compared it to an equivalent system that does not support structure. We investigated the impact of structured annotation on workload and amount of information communicated.

The observational study and the experiment (Phases I and III) were conducted with writing tutors at a local university, who professionally annotate documents to help students improve their writing skills. Having reviewed and annotated numerous documents of various lengths and types, tutors have experience communicating document-related information. We used experienced annotators because we wanted to understand how structuring annotations would be used to address different types of errors (e.g., syntax, semantics); less experienced writers tend to focus only on syntactic errors when reviewing the documents [14]. For Phase II, graduate students were used for recruiting efficiency. In that study, annotations did not need to be created, only organized; the level of experience required was relatively lower than in our other two studies.

Our research is the first to assess the impact of supporting structure on users who create annotations. We have learned that (1) bottom-up and top-down approaches to structuring annotations are complementary and that both should be supported, and (2) structured annotations are well received and perceived to be worth the additional effort. We have also contributed a lightweight implicit structuring approach based on tagging.

RELATED WORK

Collaborative writing has been examined since the early 1990's in the HCI community. The overall co-authoring process and the practices involved have been investigated [7, 15, 18, 19] and collaborative writing systems have been developed: SASSE [1], PREP [13] and Collaboratus [11]. In our review of the literature we focus primarily on document-related communication in the form of annotations among co-authors.

Studies of different communication mediums looking at both the annotation creation side [14] and the receiving side [3], have shown that more expressive and more interactive communication media are helpful to annotators as well as to recipients. Wojahn et al. studied the effects of annotation interfaces [25] on communication among co-authors, and found that difficulty in producing annotations often resulted in brief annotations with less elaboration. The Anchored Conversation tool supports real-time communication in the context of collaborative documents by allowing conversation scripts to be anchored into specific parts of a document [4]. Although the tool does merge the shared discussions and document artifacts, we suspect that verbosity of full conversations may overload authors.

Commercial systems (e.g., Microsoft Word, Adobe Acrobat) provide basic annotation features [26]. To enhance annotation support, richer annotation models have been developed [23, 26]. An activity-oriented annotation model was developed and implemented in a web-based tool in the context of co-authoring clinical trial protocols [24]. In that model, an annotation can be assigned to one of a set of pre-defined categories such as "question" or "reply", and can have properties assigned such as deadline or urgency [23]. Although the activity-oriented model extends basic annotation features, we suspect that pre-defined categories may be too rigid to capture many activities involved in co-authoring.

Structuring or grouping annotations is not entirely new; systems that support annotation grouping have been developed in other domains. The Knowledge Weasel system allows users to annotate and organize documents for capturing structural knowledge; annotations serve as links between the documents, and grouping annotations amounts to a hyperlink network of related information resources [10]. The Annotator, an annotation tool for taking notes on published HTML documents, supports annotation grouping across different documents by linking annotations together in "clumps" or annotation sets [16]. TagSEA, a collaborative annotation tool for software development projects, supports annotation grouping by associating related annotations with the same "tags" or keywords. The goal is to enhance navigation, coordination and knowledge capture among project members [22]. To our knowledge, no formal evaluations have been reported on any of these systems.

We have been inspired by the recent success of tagging, where users add meta-data or keywords to information resources, which later serve as navigational aids for finding and organizing information. Tagging facilitates a bottom-up organizational approach [17]. Browsing and searching shared tags encourages collaboration and cooperation, effectively promoting shared values and interests among collaborators [12]. Tagging is also claimed to require less cognitive workload from users than other information organization schemes [21].

PHASE I: OBSERVATIONAL STUDY TO UNDERSTAND THE ANNOTATION CREATION PROCESS

In order to better understand current annotating practices, we conducted a brief qualitative observational study with 5 writing tutors (4 females). The primary goal was to see what kinds of annotations are created and what processes are used. We observed the participants review and annotate a four-page essay-style document (1510 words) using their method of choice; 2 participants annotated with a pen on the printed document, and 3 used Microsoft Word with its track changes and commenting functions. A simulated email message window was also provided, allowing for additional document-related communication directed to a hypothetical recipient. Reviewing was followed by a semi-structured interview to probe the annotating practices observed. It required 2 hours for a participant to complete the study. We highlight the most salient behaviors and practices we observed in the remainder of this section.

Multiple Passes: All participants took at least two passes through the document while reviewing. They made annotations about syntax issues (e.g., grammar) on their first pass. Then, they took another pass or quick skim to check semantic issues (e.g., argument structure) and to achieve an overview of errors made and remaining work to be done. They then wrote comments on those issues at the end of the document or in an email message.

Justifications: Participants not only made suggested edit changes, but also occasionally added an explanatory comment, particularly when a problem was encountered for the first time. For example, an edit annotation suggesting a verb tense correction was accompanied by a comment explaining, "*Stay in the same tense as the rest of the sentence*". Participants revealed that explanation comments were added to help annotation recipients better understand their errors. When the same error was repeated in the document, participants did not add an explanation comment again, but expected recipients to refer back to the annotation for a previous occurrence of the problem. We note that this is one place where structure may be beneficial: all instances of the same problem can be linked together reducing any ambiguity.

Local versus Global Comments: Participants generally provided comments both at the "local" or sentence level, and at the "global" or document level. One of the participants explained: "*there are two different types of comments – comments that are specific for specific sentences … [and comments that] are about larger issues … [for example] whether or not the language is fitting to the general audience. So, it is more of a general comment, so maybe I want to highlight a few ideas that relate to that comment."* Current systems do not adequately support global-level feedback. Typically a comment such as "*Example of non-academic language – read through for this sort of language"* was inserted at the place in a document where the problem was first realized. With no explicit additional pointers, recipients would not necessarily be able to easily see all instances of a problem.

Tagging-like Behaviors: Participants used a keyword association technique to efficiently point out errors in the document. For example, one participant defined a keyword coding as "*WC=Word Choice*" and added the keyword "*WC*" to every place where she found a wording problem, instead of writing more verbose comments repeatedly.

Summary: This observational study confirmed the gap existing between current methods of annotating (pen on paper and MSWord) and annotators' needs. Annotators lack a way to describe relationships between annotations efficiently. It was clear that a keyword annotation feature would allow annotators to give feedback more efficiently in some situations.

PHASE II: PAPER PROTOTYPE STUDY TO UNDERSTAND ANNOTATION STRUCTURING

Having confirmed annotators' needs for more structured annotations, we sought to understand how annotators might go about structuring annotations. More specifically, we wanted to assess: (1) the semantics of the structures created, (2) the approaches taken to create structure (top-down, bottom-up, or otherwise), and (3) the complexity of the structuring created in terms of the size of annotation groups and whether hierarchies (e.g., groups within groups) might be used. In order to mitigate the impact of any particular tool (and its potential usability issues), we elected to do a qualitative exploratory study with a paper prototype where grouping annotations amounted to essentially making little piles of paper annotations.

Participants

Eight people (5 females) participated in the study. They were all graduate students at a local university, 1 from Zoology, 3 from Psychology, and 4 from Computer Science. A screening process ensured that all participants had co-authoring experience; 5 participants had co-authored more than ten documents, 2 had co-authored between five and ten documents, and 1 less than five documents.

Task and Materials

Participants were asked to perform the task of organizing annotations in a document. They were instructed to assume the role of a co-author collaboratively writing the given document with two other co-authors who had expertise in different areas. The participants' task was to organize *pre-existing* annotations in the document, ones they had hypothetically just created, so that their co-authors could review it efficiently and accurately.

The document consisted of 932 words and 42 annotations. Because we were interested in variability among users' grouping approaches and annotation groups, we provided the same document and annotations to all participants, who were asked not to add any new annotations. The scenario and annotations were designed with an assumption that different kinds of annotation groups could be created (e.g., based on types {edits, comments}, themes {tone, clarity}, or a targeted co-author). The document was on a topic in

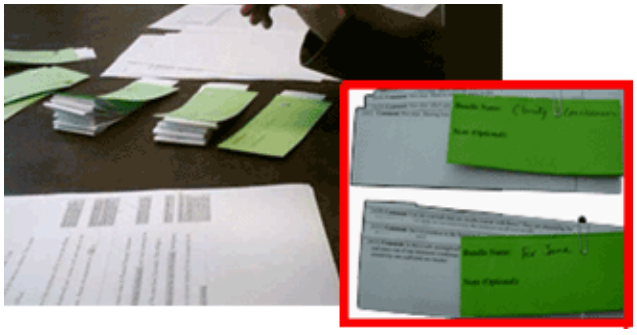

Figure 1: A participant performing the task with the paper prototype. Sample piles of annotation groups created are displayed on the right bottom corner.

Psychology (effects of music on memorization), but the content was general enough for participants to understand.

Participants were given a printed copy of the document with annotations displayed on the right margin with their text anchor highlighted. Separate identical copies of each annotation printed individually on small paper strips were made available so that participants could pile the strips together (and optionally paper-clip them) to make annotation groups. To identify a pile, a Post-it® sheet was placed on top for writing the annotation group's name and an optional note. Multiple copies of each annotation were available so that participants could place an annotation into more than one group. Each annotation group was allowed to be nested under other groups in any hierarchical structure. Figure 1 shows a participant performing the annotation-organization task, and some of the piles of annotation groups that were created.

Procedure

The study was designed for a single one and a half hour session. A demographic questionnaire was followed by an information session on general concepts such as collaborative writing and then a training session on how to use the paper materials in the task. Participants were then asked to read an annotation-free version of the task document, after which they were given the annotated version and told that they had to perform the annotation-organization task. For the first pass over the annotated document, participants were required to read the annotations in the order that they appeared in the document in order to simulate that they had themselves annotated the document in sequential order. They were, however, allowed to start grouping annotations at any point during the task. A short questionnaire and a debriefing session were used to gain better insight into the grouping behaviors and preferences that were observed.

Results

We coded all behaviours related to the simulated reviewing (i.e., participants reading pre-existing annotations) and annotation group creation. This allowed us to understand the temporal patterns of annotation grouping. We also collected the "piles" of annotation groups at the end of the task to analyze their structure. In addition, qualitative feedback provided during the debriefing was transcribed.

Temporal patterns for creating groups: Participants followed different temporal sequences for grouping annotations: four dominant patterns emerged (as shown in Table 1), distinguished by the number of passes made over the document and when, with respect to those passes, the annotation groups were created.

The only participant who used the *pre-review* approach mentioned that the co-authoring scenario informed him of the annotation groups he wanted to create. This may have been an artifact of our study design, or may simply represent a difference in style, as none of our other participants followed this approach. The 3 participants who used the *post-review* approach said that seeing all annotations in the document before grouping helped them make their groups more consistent and manageable. They mentioned that they took *mental notes* of annotations of interest so that they could relocate them for grouping later. Two of these participants externalized their mental notes by adding keywords or notes to annotations on the printed document during their first read through, and then grouped annotations based on those keywords. The 2 participants who used the *during-review* approach stated that they grouped annotations "naturally" as occurred to them without explicitly having to think about grouping. Lastly, the 2 participants who used the *hybrid* approach mentioned that they created "obvious" groups (such as typos, grammar) by using the pre-review approach and other groups by using other approaches. After grouping all annotations, 4 out of the 8 participants (2: *hybrid,* 1: *post-review,* 1: *during-review*) reshuffled some of their groups by merging or splitting.

	Description
Pre-review Num of pass: 1 (1 participant)	Participant formulated annotation groups prior to seeing annotations. Annotations that fit these pre-defined groups were later selected and associated with corresponding groups.
Post-review Num of pass: 2 (3 participants)	Participants read *all* the annotations prior to formulating any groups. Once groups were defined, relevant annotations were associated with the groups.
During-review Num of pass: 1 (2 participants)	Participant organized annotations into groups while reading annotations, created new groups when existing ones were not appropriate for a given annotation.
Hybrid Num of pass: 2 (2 participants)	Different groups of annotations were created using different approaches stated above (*pre-, post-* or *during- reviews*)

Table 1. Temporal patterns of structuring annotations. (N=8)

Semantics of annotation groups: Groups resulted from organizing disparate annotations throughout the document that were conceptually related. Most groups (82%) were *problem-based*: annotations were similar in the nature of problems that they addressed (e.g., the group named "Tone" had annotations that highlighted and discussed inconsistent tone throughout the document). The remaining groups (18%) were *recipient-based*: annotations that were to be reviewed by a particular co-author (e.g., the group named "For Jane" had annotations that solicited Jane's expert knowledge).

Structures of groups: The average number of groups created per participant was 8.3 (sd: 2.9, min: 5, max: 14). The average number of annotations per group was 8.2 (sd: 2.6, min: 1, max: 22). Overall, group structuring was not very complex; 45% of groups had a *flat structure* in that they

had no groups within them, and the remaining 55% of groups were structured in *hierarchical structures* (groups within groups). We analyzed the complexity of these hierarchies in terms of height (the length of path from the top-level group to the furthest sub-group), and found that the average height of *hierarchical structures* was only 1.4 (sd: 0.5, min: 1, max: 2).

Analysis on the groups within groups revealed that 44% of them were true subcomponents of their higher leveled groups (what we call a "proper hierarchy"), e.g., a subgroup named "Missing standard deviation" nested within a group "Missing Information" (since standard deviation is one of the information types presented in the document). The rest of the subgroups did not reflect such proper subset relationship; the hierarchy seemed to result from the individual participant's decision about the relative importance of attributes (what we call an "arbitrary hierarchy"). For example, a participant created a group "For Jane" with a nested sub-group "Clarifications" because she wanted to emphasize and make the recipient-based information more salient. At the same time, another user created the reversed structure: "Clarifications" with a nested sub-group "For Jane", in which the problem-based information was more emphasized.

Strong support for structuring: 7 out of 8 participants strongly agreed that they liked being able to organize annotations within a document. One of the participants commented that groups were "*infinitely easier than the current annotation format and [can be used] to delegate sections to different authors [which] reduces duplicate effort*". Another participant mentioned that she would like to use annotation groups not only to facilitate her co-authors' workflow but also to manage her own workflow.

Summary and additional comments: All participants created annotation groups during the task and perceived the benefits of supporting grouping in annotation tools. We found that participants used different temporal patterns to organize annotations, and we identified common semantics of annotation groups, but we did not observe very complex group structuring. It could be that the numbers of annotations and groups were not large enough to call for complex structures; additional research with larger documents and more iterative cycles of reviewing would be needed to assess the usefulness of complex structuring. Interestingly, the subject of the document appeared to have had an impact on the results. The 3 participants from Psychology created 8 or more groups in the task, while other participants from Computer Science and Zoology created fewer groups. One possible explanation for this result is derived from the social psychology literature: large numbers of categories were reflected by users' familiarity with the subject of the document and their thorough understanding about the subject [20]. Structures of some groups also resembled a "divide and conquer" problem decomposition approach in which annotation groups and subgroups correspond to components and subcomponents of the remaining work to be done in the document.

IMPLICATIONS AND MOTIVATIONS FOR STRUCTURING AND TAGGING DERIVED FROM PHASES I & II

The observational study and the paper prototype study revealed that the ability to organize annotations into groups could benefit co-authors in several ways, namely by facilitating communication, problem decomposition, and workflow management among annotators and co-authors.

The different temporal patterns observed for organizing annotations seem to reflect the top-down and bottom-up approaches, as suggested by the information processing literature [8]. We speculate that the *pre-review* approach is a form of *top-down* while *post-review* is more like to *bottom-up*. Reshuffling of annotation groups can be considered as *middle-out* processing. To support these different approaches, it should be possible to create annotation groups at any time, before, during or after single annotations are created. We assert that mechanisms for creating and managing annotation groups should be flexible and lightweight.

Consistent with findings reported by Neuwirth [14], we found that participants sometimes made mental notes on annotations to which they wanted to return for organizing. Hence, providing a lightweight means to externalize such mental notes and support for navigational aids should help reduce users' cognitive load.

We saw diversity in the degree of group structuring (proper and arbitrary hierarchies, and flat structures). Disagreement or conflicts among users in defining structures and hierarchies of annotation groups may cause ambiguities and inefficiencies. We argue that tagging is likely a good solution to this problem. With tagging, users do not need to agree on a particular hierarchy, instead they just need to have a shared understanding of a tag's meaning [12]. Hierarchies should, however, still be supported to recognize proper subset relationships among annotation groups when they exist.

Although we did not observe very complex group structuring, we imagine that complex structures might arise as the number of annotations or the size of a collaborative artifact grows over time. We realize that having complex structures might hinder the co-authoring process, due to the additional navigation time required to explore a highly nested annotation and the additional efforts to develop and agree among collaborators on hierarchical information. This affirms the importance of reducing complexity in the degree of structuring, and we note that tagging can be a good approach because it can allow for implicit structures.

Integrating Tagging and Structuring to Annotations

Based on the implications we drew from Phase I and II, we extended a previous structured annotation model [26] by adding tags as an optional annotation attribute (see Table 2). Tags in our model can serve to:

- efficiently associate a keyword with annotations,
- externalize mental notes or act as navigational aids,
- easily identify semantic concepts inherent in annotations,
- facilitate workflow by allowing for bottom-up annotation grouping, and
- simplify structures of annotation groups through *implicit* grouping.

Tags are treated as meta-information about annotations that users can easily add as they review and annotate a document. Tags can be used as navigational aids; by filtering annotations on a particular tag, users can easily jump between related annotations in the document. Tags allow flexible classification of annotations based on their semantic concepts. Tags provide implicit groups for annotations because co-authors can easily see relationships among annotations labeled with the same tags, even though the annotations may not be grouped together explicitly.

Extended Bundle Editor

We implemented an extended version of the *Bundle Editor*, a prototype that supports structured annotations [26]. The main interface to the Bundle Editor prototype consists of a document pane and a reviewing pane. The document pane serves as a document editor with basic functionalities such as insert, delete, comment, etc. The reviewing pane consists of multiple tabs, each of which displays a specific "bundle" or group of annotations.

The Bundle Editor facilitates multiple approaches to creating and managing bundles previously described. It supports *top-down* grouping of annotations by allowing users to create a bundle (Figure 2A) and then explicitly select annotations to be grouped into that or other bundles (Figure 2B). *Bottom-up* grouping is supported by allowing annotations to be tagged with one or more keywords (Figure 2C), and then filtered or selected based on their tags (Figure 2D). *Middle-out* grouping is supported by allowing bundles and annotations to be easily added to or removed from existing bundles (Figure 2E).

To achieve all the advantages of tags, as described above, we designed tagging to be pervasive throughout the system. Users can easily associate an annotation with one or more tags by simply typing into a textbox or selecting from a list of existing tags. The prevalence of a given tag is visible to users through the display of its frequency right next to the tag word. Users can filter annotations based on AND, OR combinations of tags and other attributes such as annotator and annotation type.

Annotation Model	
Mandatory Attributes:	*Optional Attributes:*
▪ Annotator	▪ Name ▪ Modification
▪ Timestamp	▪ Recipient ▪ Substructure
▪ Reviewing Status	▪ Comment ▪ Tag
▪ Anchor	▪ Priority

Table 2. Annotation Model from [27] with our added Tag attribute.

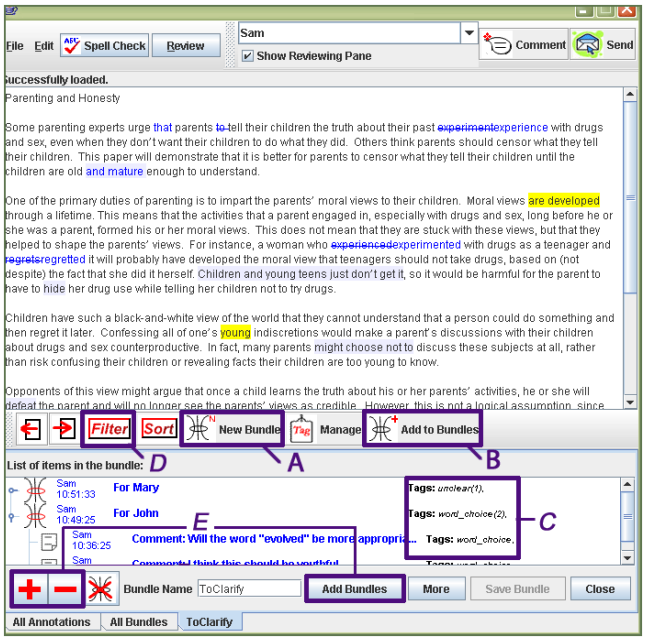

Figure 2: Bundle Editor with document and reviewing panes.

PHASE III: EXPERIMENT TO COMPARE THE IMPACT OF STRUCTURING ANNOTATIONS

In Phase III we conducted a controlled experiment to compare our extended Bundle Editor with a Simple Editor, an otherwise equivalent system that did not support annotation structure (Table 3 summarizes the differences between the two systems). Our main goals for the experiment were to investigate the impact of structuring annotations on annotator's workload and on the amount of information communicated to hypothetical co-authors. More specifically, we wanted to know (1) if under a controlled comparison users would still perceive the benefits of structuring their annotations (as they had with the paper prototype study); (2) if the overall workload for structuring annotations would be similar to users providing non-structured annotations but also having to communicate any meta information in a text email format (i.e., outside of the document); and (3) whether the overall amount of information communicated would differ with structured annotations.

Participants

As in the observational study, professional writing tutors were recruited. A total of 12 people (all females) participated in the study. All 12 participants used a word processor regularly (mainly Microsoft Word), although 6 had never used any annotation functions. Eleven participants used a word processor frequently (4 participants everyday and 7 every two to three days) and felt very confident in their usage. The remaining participant only used a word processor once a month and her confidence was relatively low. All participants had

reviewed documents more than 10 times. Seven participants had previously been involved in collaborative authoring, 3 more than five times, and 4 fewer than five times.

Tasks

Participants were asked to review and annotate two documents, one with each of the systems. They were asked to assume a role as a collaborator within a group of three co-authors. The given documents were assumed to be drafted by the other two co-authors with some sections explicitly noted as being jointly drafted by the two co-authors, and other sections drafted separately by one co-author.

Unlike in the paper prototype study, participants were expected to create annotations, both comments as well as direct edit changes to the document text. They were also requested to provide: (1) their general impression of the writing and (2) a brief summary of a review that provided an overall status of the document, which would help their co-authors skim the document quickly and prioritize the remaining work. The requested feedback was representative of common meta-comments communicated between annotators and recipients as found in our observational study.

The documents were manipulated to be isomorphic; they had the same number and types of problems planted at similar locations throughout the documents. Prior to the experiment, to test the manipulations, two independent raters with expertise in writing were asked to identify planted problems in the documents. Both raters were able to identify the majority of planted problems in each document (72% and 70% respectively). The problems that were not identified by either rater were removed from the task documents. In the end, each document had a total of 33 planted problems: 19 syntax errors and 14 semantic errors. A third document was used during practice sessions. Because that same document was used first by every participant in both conditions, we did not control the number or types of problems in the document.

	Bundle System	***Simple System***
Interface Components	Document panel, *multi-tabbed* reviewing panel	Document panel, *single pane* reviewing panel
Communication Support	Single annotations with *(optional) user-defined tags*, *general comments*, *structured annotations* embedded in the document and listed in the reviewing panel.	Single annotations embedded in the document and listed in the reviewing panel, a *simulated email message window*
Filtering Functions	AND, OR filtering on Author, Type, and *Tag*	AND, OR filtering on Author, and Type

Table 3: Functions of bundle system and simple system. Both systems were created by modifying our Bundle Editor so that they differed only in their communication supports in terms of annotation functions.

Design

The experiment used a within-subjects (system type) factorial design. Document-type was a within-subjects control variable, and both system and document presentation orders were between-subject controls. To minimize learning effects, we counterbalanced the order of presentation for both system type and document, resulting in four configurations.

Procedure

Each participant had a single four hour session. It began with a demographic questionnaire to obtain past computer, co-authoring and reviewing experience. Participants then saw a training video on general concepts such as collaborative writing and annotations, and how to use their first assigned system. To ensure that all participants would have a similar level of familiarity with the system functionalities, participants had 15 minutes to perform a set of practice guided annotation tasks with that system, where they were provided with a list of annotations to create in the practice document. Participants had an hour to perform the experiment task on the first document with the first system. A questionnaire followed to collect feedback on that system. Participants were given a short 10 minute break and were then shown a training video on how to use their second system, followed by a 15 minute practice session, then the experiment task on the second document with the second system. A final questionnaire was administered to gather feedback on that system. A short semi-structured interview to collect further information regarding preferences and perceived performance ended a session.

Measures

The amount of communication among co-authors was assessed by counting the number of (1) single annotations and (2) meta-comments. Single annotations were the same in both systems: edits and comments. Counting meta-comments differed between the two systems. In the Bundle system, each unique bundle and tag as well as each general comment counted as a meta-comment. In the Simple system, email content was analyzed to extract meta-information items; e.g., a statement saying "Try to use more academic words in the places I highlighted" was counted as one meta-comment. To ensure all meta-comments created were counted, single comments anchored at the beginning and the end of documents in both systems were also analyzed to see if they contained any meta-information. Self-reported measures from the two system-specific questionnaires were used to assess subjective workload measures associated with each task using the NASA-TLX

workload index [6]. During the interview, participants were asked to comment on the cost-benefit tradeoff of using each system.

Hypotheses

Communication Hypotheses: (a) Participants will create more meta-comments in the Bundle system than in the Simple system because creating structured annotations is an easier way to provide meta-commentary than doing so separately in the body of an email; (b) participants will create similar numbers of single annotations in both systems because both systems support single annotations identically (except for tags).

Workload Hypothesis: Reviewing with the Bundle system will not require significantly higher workload than reviewing with the Simple system because the added effort to group and tag annotations will not be greater than that required to compose a detailed email with the equivalent information.

Cost-Benefit Hypothesis: Participants will perceive the net gain (the amount by which the benefit exceeds the cost) to be higher in the Bundle system than in the Simple system.

Results

We report on the quantitative data along with the qualitative feedback provided during the interview. Before testing our hypotheses, we checked to make sure that there was no effect of document; a series of 2 documents x 2 order of systems x 2 order of documents ANOVA tests on our dependent measures showed no significant main or interaction effects of document on the data. We then ran a series of 2 systems x 2 order of systems x 2 order of documents ANVOA to test our hypotheses. Along with statistical significance, we report partial eta-squared (η^2), a measure of effect size, which is often more informative than statistical significance in applied human-computer interaction research [9]. To interpret this value, 0.01 is a small effect size, 0.06 is medium, and 0.14 is large [5].

Communication

Participants created an average of 64.5 single annotations in the Bundle system and 76.1 annotations in the Simple system, a difference that was not statistically significant ($F(1,8)=2.20$, $p=0.18$, $\eta^2=0.22$). Although, this was expected and supports our communication hypothesis for these types of annotations, we note that a large effect size was found indicating that it may be prudent to validate this finding with further research.

In terms of meta-comments, two main categories were observed: (1) *recipient-based* (e.g., a to-do list for Nick) and (2) *problem-based* (e.g., tone of the document). The categories were not exclusive; in some cases, the same annotation(s) were counted as both types, e.g., an annotation tagged with the content-related info "argument" and associated with a to-do bundle "For Mary", or a statement saying "Mary, you should watch out for unsupported claims in paragraph 3."

Participants created significantly more meta-comments ($F(1,8)=13.09$, $p=0.01$, $\eta^2=0.62$) when reviewing with the Bundle system (avg: 6.7, sd: 3.3) than reviewing with the Simple system (avg: 3.7, sd: 2.2), also supporting our communication hypothesis. Interestingly, we found a significant system order effect on the number of meta-comments ($F(1,8)=17.44$, $p<0.01$, $\eta^2=0.69$). Participants who were exposed to the Bundle system first included significantly more meta-comments across both systems (avg: 7.0) than those who used the Simple system first (avg: 3.3). One explanation is that the Bundle system facilitated meta-comments in the first reviewing task, leaving participants with the inclination to similarly provide more information in the second task.

Consistent with the quantitative data, many participants also said that they were able to provide a more comprehensive review using the Bundle system, e.g., *"[The Bundle system] maximizes the interaction between the writers", "[When] you need a more critical approach [it] gives you the exact tools",* and "*I could communicate more information [that] I think is important to get across. … I was not just correcting the problems; I had a chance to explain why… to justify it".*

Self-Assessed Workload

Perceived workload with the Bundle system, as measured by the TLX, was 69.8 (sd: 8.0) while that associated with the Simple system was 63.0 (sd: 11.8), a marginally significant difference ($F(1,8)=4.53$, $p=0.07$ and $\eta^2=0.36$). This finding was not consistent with our workload hypothesis and we reflect on this unexpected difference in the Discussion section.

Cost-Benefit Tradeoff

During the interview, participants were asked to comment on the cost-benefit tradeoff for using each system. Eleven of the 12 participants found both systems to be useful and have positive net gain (the benefit outweighed the cost). Among these 11 participants, eight said the net gain is higher in the Bundle system than in the Simple system, and that they would definitely use the Bundle system in their future annotating tasks. These participants acknowledged that even though the cost of using the Bundle system was higher than for the Simple system (as reflected in our workload measure), the Bundle system would return a much greater benefit, especially over iterative collaborations. For example, one participant explained that because she was able to provide the authors with a more comprehensive review using the Bundle system, *"going forward, if I am working with [the same co-authors] again, they would already know what kind of things I am looking for. So it's like I do the front heavy loading [by putting in extra effort for the first iteration] … As you front load the work, as you go through [over iterations], it only gets easier*."

The remaining three participants (out of 11) mentioned that although the Bundle system's benefit was higher than the Simple system, the cost associated with the Bundle system was much higher resulting in a lower net gain compared to

the Simple system. Nonetheless, they mentioned that they would use the added functionality in the Bundle system in some of their future annotating tasks where they needed to provide detailed and precise feedback on larger documents.

The remaining one participant believed that the returned benefit was not worth the cost for either system. When forced to choose between the two systems for her future use, she chose the Simple system because she preferred to give free-form non-structured feedback similar to the verbal feedback to which she was more accustomed. Thus, while the majority of participants thought that structuring annotations was worth the effort, there was clearly some diversity of opinion on this point.

Other Measures: Usage of Bundles and Tags

Twelve participants created bundles, and eleven created tags to communicate meta-information. Bundles were used to communicate both *problem-based* (59%) and *recipient-based* (41%) types of meta-information, while tags were used exclusively for *problem-based* information. Most problem-based bundles were created based on a single tag or a set of similar tags using a bottom-up approach. Most participants described that creating bundles using this approach was easier and less time-consuming. Recipient-based bundles were created using a top-down grouping approach, and had more diverse sets of tags associated with the grouped annotations. We believe that these recipient-based bundles were created to help the recipients manage their workflow. This was explained explicitly by one participant, who reflected on one of her previous collaborative experience and stated that:

> "*Bundles would have been useful for addressing the co-authors' problems individually. They didn't have the same […] errors. So, being able to separate them out, say you need to work on this, you need to work on that. But then also for the things that they were working together, [bundles would allow me] to be able to combine them as well. So it's a way of both separating them out but then making [them] more cohesive at the same time."*

Another participant explicitly explained that the workflow-related information communicated through bundles could help recipients review annotations efficiently because "*everything that requires a certain way of dealing with is together [in bundles]"* and hence, co-authors "*don't have to keep switching their mindset from one thing to another"*.

It was surprising that participants did not use tags *at all* to communicate recipient-based meta-comments; they used tags only to explain specific aspects of the problems that they were trying to address in the annotations. Participants said that the information communicated through tags brought "*awareness to patterns of problems in documents"*, and would allow co-authors to achieve a quick overview of the current document status, and also to see the strengths and weaknesses of their writing. One participant further stated that the information also allowed her to achieve "*a greater perspective on the reviewing process"* that she went through. We also found that tags were used as alternatives to long comments when addressing recurring problems. This was explained explicitly by one participant during the interview, "*One can comment the first time one runs into a problem. But after that, [tags] are like reminders, almost to go back to that comment*."

The average numbers of bundles and unique tags created per participant were 2.4 (sd: 1.4, min: 0, max: 5), and 4.8 (sd: 2.8, min: 1, max: 11) respectively. It was interesting to note that while a grater number of unique tags were created than bundles, a higher percentage of annotations were associated with bundles (30.9%) than with tags (23%); the number of annotations per bundle (avg: 7.7) was higher than that per tag (avg: 3.3). Moreover, we found that tags were used to label more comments than edits, while bundles were used to organize both comments and edits in similar proportions. Furthermore participants created more edit annotations in the documents than comments.

Feedback on Usability of the Bundle System

Six participants suggested that the Bundle system needed to be more intuitive and straightforward. The interaction technique for adding/removing annotations to/from bundles was a bit cumbersome: a separate tab for each bundle had to be opened and a few button clicks were required for each annotation added/removed. Improving the usability of the system would involve implementing more efficient interaction techniques for annotation organization, such as drag-and-drop.

Preference for structuring

We note that 4 of the 6 participants who were exposed to the Bundle system first said that while performing the second task with the Simple system, they wished it had some of the Bundle system functionalities (e.g., tagging or grouping annotations). They felt that they could not provide feedback "*as precise and thorough as in the Bundle system"* and they had difficulty "*explaining how problems [were] connected and uniting comments."*

Discussion

Structured annotations are worth the effort: The majority of participants thought that structured annotations offered higher net gain, and definitely would use them in the future. For the remaining participants, the perceived benefits did not sufficiently outweigh the additional workload associated with structured annotations; nevertheless they said that they would use structured annotations in certain contexts where they needed to annotate documents thoroughly. We believe that the high workload may be explained by usability issues uncovered with the Bundle system as discussed earlier, or by the fact that participants included more meta-comments using the Bundle system, thus requiring more work. Similarly, when communicating meta-comments in the email messages of the Simple system, participants did not provide detailed information such as explicit pointers or references to the document text, thus doing less work.

Top-down and bottom-up approaches complement each other: Although the bottom-up approach was considered to be less time consuming and more lightweight, it did not displace the top-down approach. Both approaches were used by participants (one participant who did not use bundles commented during the interview that she ran out of time to create bundles towards the end of the task). We suspect that recipient-based bundles were created using a top-down grouping approach because the information regarding intended co-author(s) was known *prior to* annotating the documents. Conversely, problem-based bundles were created bottom-up because the structure was formulated only after realizing relationships among single annotations. Hence the top-down and bottom-up approaches support structuring annotations in a complementary way.

Structured Annotations promote valuable communication: Participants created more meta-comments in the Bundle system than in the Simple system. Participants' remarks on the usefulness of bundles and tags as described earlier suggest that the meta-information communicated in the form of structured annotations was perceived to be valuable and beneficial for both recipients and annotators themselves, by allowing for a comprehensive review that goes beyond simple annotations, providing for a greater perspective on the reviewing process, supporting workflow management, and offering a quick overview of the document status.

CONCLUSION AND FUTURE WORK

We assessed the impact of supporting structured annotations on users who create annotations. The studies conducted in Phase I and II of our research revealed that structured annotations could benefit co-authors in facilitating communication, problem decomposition, and workflow management among annotators and co-authors. The experiment in Phase III compared the impact of structured annotations relative to unstructured annotations. Participants in that study perceived structured annotations to be worth the additional workload required. The study further suggested that the bottom-up grouping approach complements the top-down approach in describing relationships amongst annotations in a document. Further work is needed to improve the usability of the prototype, and to explore how the additional workload can be streamlined. The impact of supporting structured annotations in larger documents and iterative cycles of collaborative reviewing and editing needs to be examined with longitudinal field studies.

REFERENCES

1. Backer, R. M., Nastos, D. , Posner, I. R. and Mawby, K. L. The user-centered iterative design of collaborative writing software. In Proc. CHI 1993, ACM Press, pp. 399-405, 1993.
2. Bloehdorn, S., and Volkel, M. TagFS – tag semantics for hierarchical file systems. In Proc. I-Know 2006.
3. Chalfonte, B., Fish, R. S., and Kraut, R. E. A comparison of speech and text as media for revision. In Proc. of CHI 1992, ACM Press, pp. 21-26, 1992.
4. Churchill, E., Trevor, J., Bly, S., Nelson, L., and Cubranic, D. Anchored conversations: chatting in the context of a document. In Proc. CHI 2000, ACM Press, pp. 454-461, 2000.
5. Cohen, J. Eta-squared and partial eta-squared in communication science. In Human Communication Research, vol. 28, Oxford Journals, pp. 473-490, 1973.
6. Hart, S. G., and Staveland, L. E. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), Human Mental Workload, Amsterdam: Elsevier Science. pp. 139-183, 1988.
7. Kim, H., and Eklundh, K. S. Reviewing practices in collaborative writing. In Proc. CSCW 2001, ACM Press, pp. 247 - 259, 2001.
8. Kinchla, R. A., and Wolfe, J. M. The order of visual processing: top-down, bottom-up or middle-out. In Perception & Psychophysics, vol. 25, pp. 225-231, 1979.
9. Landauer, T. Chapter 9: Behavorial research methods in human-computer interaction, in M. G. Helander, T. K. Landauer, and P. V. Pranhu (Eds), *Handbook of Human-Computer Interaction*, 2nd Edition Amsterdam: Elsevier Science B.V, 1997, pp. 203-227.
10. Lawton, D., and Smith, I. E. The knowledge weasel hypermedia annotation system. In Proc. HT 1993, pp. 106 - 117, 1993.
11. Lowry, P. B., Albrecht, C. C., Lee, J. D., and Nunamaker, J. F. Users experiences in collaborative writing using Collaboratus, an Internet-based Collaborative Work. In Proc. of International Conference on System Sciences, 2002.
12. Mathes, A., Folksonomies - cooperative classification and communication through shared metadata. Accessed: April 12, 2007. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html
13. Neuwirth, C., Kaufer, D. S., Chandhok, R., Morris,J.H. Issues in the design of computer support for co-authoring and commenting. In Proc. CSCW, ACM Press, pp. 183-195, 1990.
14. Neuwirth, C., Chandhok, R., Charney, D., Wojahn, P., and Kim, L. Distributed collaborative writing: a comparison of spoken and written modalities for reviewing and revising documents. In Proc. CHI, ACM Press, pp. 51-57, 1994.
15. Noel, S., and Robert,J.M. Empirical study on collaborative writing: what do co-authors do, use and like? In Proc. CSCW, vol. 13, ACM Press, pp. 63-89, 2004.

16. Ovsiannikov, I., Arbib, M., and Mcneill, T. Annotation technology. In International Journal of Human-Computer Studies, vol. 50, pp. 329-362, 1999.
17. Porter, J., Folksonomies: A user-driven approach to organizing content. Accessed: March 14, 2007. http://www.uie.com/articles/folksonomies/.
18. Posner, I. R., and Baecker, R.M. How people write together. San Mateo, CA: Morgan Kaufmann, pp. 239-250, 1993.
19. Rimmershaw, R. Collaborative writing practices and writing support technologies. In Instructional Science, vol. 21, pp. 15-28, 1992.
20. Scott, W. A. Cognitive complexity and cognitive flexibility. In Journal Information for Sociometry, vol. 25, pp. 405-414, 1962.
21. Sinha, R. A cognitive analysis of tagging. Accessed: June 10, 2007. http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html
22. Storey, M., Cheng, L., Bull, I., and Rigby, P. Shared waypoints and social tagging to support collaboration in software development. In Proc. CSCW, ACM Press, pp. 195-198, 2006.
23. Weng, C., Gennari, J. Asynchronous collaborative writing through annotations, Notes. In Proc. CSCW, ACM Press, pp. 564-573, 2004.
24. Weng, C., McDonald, D. W., and Sparks, D. Participatory design of a collaborative clinical trial protocol writing system. In International Journal of Medical Informatics, pp. 245-251, 2006.
25. Wojahn, P., Neuwirth, C., and Bullock, B. Effects of interfaces for annotation on communication in a collaborative task. In Proc. CHI, ACM Press, pp. 456-463, 1998.
26. Zheng, Q., Booth, K. S., and McGrenere, J. Co-authoring with structured annotations. In Proc. CHI, ACM Press, pp. 131-140, 2006.