Creating Better Action Plans for Writing Tasks via Vocabulary-Based Planning

HARMANPREET KAUR, University of Michigan, USA
ALEX C. WILLIAMS, University of Waterloo, Canada
ANNE LOOMIS THOMPSON, Microsoft Research, USA
WALTER S. LASECKI, University of Michigan, USA
SHAMSI T. IQBAL, Microsoft Research, USA
JAIME TEEVAN, Microsoft Research, USA

While having a step-by-step breakdown for a task—*an action plan*—helps people complete tasks, prior work has shown that people prefer not to make action plans for their own tasks. Getting planning support from others could be beneficial, but it is limited by how much domain knowledge people have about the task and how available they are. Our goal is to incorporate the benefits of having action plans in the complex domain of writing, while mitigating the time and effort costs of creating plans. To mitigate these costs, we introduce a vocabulary—a finite set of functions pertaining to writing tasks—as a cognitive scaffold that enables people with necessary context (e.g. collaborators) to generate action plans for others. We develop this vocabulary by analyzing 264 comments, and compare plans created using it with those created without any aid, in an online study with 768 comments ($N = 145$) and a lab study with 96 comments ($N = 8$). We show that using a vocabulary reduces planning time and effort and improves plan quality compared to unstructured planning, and opens the door for automation and task sharing for complex tasks.

CCS Concepts: • **Human-centered computing** → *Empirical studies in HCI*; Collaborative and social computing systems and tools;

Keywords: Action plans; task decomposition; crowdsourcing

ACM Reference Format:
Harmanpreet Kaur, Alex C. Williams, Anne Loomis Thompson, Walter S. Lasecki, Shamsi T. Iqbal, and Jaime Teevan. 2018. Creating Better Action Plans for Writing Tasks via Vocabulary-Based Planning. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2, CSCW, Article 86 (November 2018). ACM, New York, NY. 22 pages. https://doi.org/10.1145/3274355

1 INTRODUCTION

Action plans—step-by-step breakdowns of larger tasks—help people get started on and accomplish tasks [9, 20, 35]. While making plans for one's tasks is considered a helpful form of contemplation, people prefer to have automatically created plans at their disposal over making them. Prior work shows that people complete more tasks when they have access to pre-generated personalized plans for their tasks [29]. We also see evidence of this in the rise of industry applications that provide

Authors' addresses: Harmanpreet Kaur, University of Michigan, USA, harmank@umich.edu; Alex C. Williams, University of Waterloo, Canada, alex.williams@uwaterloo.ca; Anne Loomis Thompson, Microsoft Research, USA, annelo@microsoft.com; Walter S. Lasecki, University of Michigan, USA, wlasecki@umich.edu; Shamsi T. Iqbal, Microsoft Research, USA, shamsi@microsoft.com; Jaime Teevan, Microsoft Research, USA, teevan@microsoft.com.

2573-0142/2018/11-ART86 $15.00
https://doi.org/10.1145/3274355

planning and task breakdown services: RunKeeper plans people's workouts based on a health goal [1], Cook Smarts helps plan their meals [2], and Mint helps plan finances [4].

While having personalized, pre-generated plans is beneficial and preferred, it is hard to outsource plan creation for all kinds of tasks. Fully-automated approaches cannot currently break down a task into subtasks because of the lack of natural language understanding, or because they are missing process-level structural information about complex domains. Prior work has navigated this issue by employing crowd workers, friends, and collaborators to make plans for people, but this works primarily for context-free, independent tasks (e.g. planning a trip from an airport to a hotel) because these can be planned without additional information [29, 32]. Even then, people sometimes find it awkward to ask their friends or collaborators to help with planning because of the social costs of sharing potentially private information or burdening others with their tasks [8, 38].

Outsourcing the generation of action plans is particularly challenging for context-embedded tasks—those that require additional contextual information, such as comments or to-dos in a written document or codebase, for health goals, etc. Crowd workers or, more generally, people who do not have the required context and domain knowledge, cannot plan for context-embedded tasks [29, 39]. Even people who already have the required domain knowledge in these situations (e.g. collaborators, friends) are often not a feasible option because of the information differential between the person(s) doing the task and those collaborating on it. This information differential requires considerable time and effort in planning for someone else's task, and often there is tacit information about the task that is not communicated between those doing the task and those collaborating on it [40].

In this paper, we present and evaluate a cognitive scaffold that enables individuals to generate step-by-step action plans in support of helping others accomplish context-embedded tasks. Our cognitive scaffold is a vocabulary—a set of generic functions that define the processes involved in performing larger tasks in a domain. The vocabulary provides a language for decomposing a larger task within a particular contextual domain into a set of actionable subtasks. Within the context of creating action plans, the utility of a vocabulary is grounded in its ability to reduce the costs associated with the time and effort of planning, since the potential steps of a plan are provided as functions within the vocabulary. Similarly, having a set of consistent, generic functions to choose from reduces the number of instances where all basic steps are not listed. Further, as a cognitive scaffold, a vocabulary may reduce the necessary amount of in-depth domain knowledge required for providing both helpful feedback and a plan for a particular task.

We focus on the domain of writing, exploring how writing tasks specified in comments left in a document can be transformed into a plan. We choose comments in a document as our initial way of accessing actionable, interdependent, context-rich tasks embedded in a document, that need to be accomplished to improve the document's writing. Generally, comments are anchored to a location in the document, making it easier to contextualize the tasks, and they reliably present both the collaborators' and the author's anticipated tasks for different parts of the document. As a domain, writing varies greatly in the amount of domain knowledge required to contribute [45]. We thus evaluate our approach for two different levels of domain knowledge required to provide comments on a document's writing: (i) low-domain knowledge situations where people with minimum to no domain knowledge (e.g., most crowd workers) can both comment on and plan toward the task of improving a document's writing even if they cannot, and do not, take part in the writing process, and (ii) high-domain knowledge situations where individuals with domain knowledge (e.g., experts or collaborators) primarily provide comments on a document's writing and help generate plans to resolve these comments.

Our project has three phases (Figure 1). In Phase 1, we use a data-driven process to qualitatively derive a vocabulary of 18 functions for accomplishing writing tasks by using a set of 264 comments left on Wikipedia articles and academic papers. In Phase 2, we explore how plans are created using the vocabulary, and compare the experience of creating vocabulary-based plans to creating unstructured plans (i.e., plans that are made without aids) in two different settings: (i) *an online study* where 145 Mechanical Turk workers create plans for 768 comments on Wikipedia articles—representing a low-domain knowledge setting; and (ii) *a lab study* with 8 intern-mentor pairs at a large technology company, where mentors comment and plan for interns' project documents (total 96 comments)—representing a high-domain knowledge setting. We measure differences in time and effort between the vocabulary-based and unstructured planning conditions, and also evaluate plan quality based on the granularity of each step, and how well the original task is broken down. Finally, in Phase 3, we evaluate the perceptions around the usefulness of the plans by asking people for their preferred method of accomplishing a task (the original comment, the unstructured plan, or the vocabulary-based plan), as a proxy for user satisfaction.

Our online study results indicate that in low domain knowledge settings, people spend significantly less time and effort on making vocabulary-based plans compared to unstructured plans, and they make vocabulary-plans with more granular, basic steps. Additionally, our lab study results show that for high domain knowledge settings, the stage of a document (i.e., how close it is to completion) has important implications for the use of a vocabulary: later-stage documents are easier to plan for using a vocabulary compared to early-stage documents due to the large number of non-action, reflection-oriented comments left in the latter. When asked to evaluate, people preferred action plans over using the original comment to complete a task, but were divided in their choice between unstructured and vocabulary-based plans. Overall, our vocabulary accomplishes its goal of being a cognitive scaffold for both of the domain knowledge settings that we test. We discuss the limitations in the high domain knowledge case, which result from the esoteric processes involved in academic writing and the different stages of writing. We further note the application of the structure and low-level breakdown of tasks afforded by this vocabulary in task sharing (via crowd- and/or self-sourcing), and automating plan generation for writing tasks.

2 RELATED WORK

2.1 Benefits of Action Plans

Action plans are a step-by-step breakdown of how to accomplish an overarching task [49, 50]. The benefits of having an action plan have been extensively outlined in HCI and cognitive psychology literatures. Most people perceive their daily activities as action steps because this structure enables faster processing of what needs to be done [50]. Similarly, people find it easier to get started on large macrotasks if there is a step-by-step plan for accomplishing the task [9, 13]. Plans also lend themselves to identifying task boundaries, which in turn enables better interruptibility [13, 24]. On comparing the outcomes of completing a macrotask with completing a list of microtasks for the same macrotask, Cheng et al. [13] found that microtasking results in higher quality work, and is more resilient to interruptions.

Breaking tasks down not only makes it easier to get started and accomplish these tasks, but also enables crowdsourcing and self-sourcing support while doing so. If plans for tasks include independent, context-free steps, crowd workers can accomplish these tasks without needing the context of the entire document [14]. Not only that, people could also self-source these tasks and complete them in micromoments rather than blocking time for writing. For example, the author could rewrite a sentence on their phone while waiting in line for coffee if they knew exactly what changes were needed [44]. This crowd- and self-sourcing of tasks not only helps people accomplish their tasks, but also provides motivation to continue working on the larger goal [43, 44].

2.2 Costs of Action Plans

While having an action plan is beneficial for the person responsible for completing the task, plan creation is challenging. On one hand, there are studies that show that planning for your own tasks helps concretize the steps, making them appear easier [9, 19, 34]. On the other hand, research shows that people often do not want to make plans for their own tasks because of the additional time and effort costs [31, 42]. Rather, they prefer plans made by others because they can provide diverse and serendipitous ways for completing the tasks [7, 8, 29].

As a potential solution, researchers have used crowd workers to make plans for other people's tasks [8, 29, 32, 33]. This has shown positive results for independent, high-level tasks such as, "Find things to do for trip to Atlantic Beach, NC" [29] because the person planning does not require additional contextual information to plan for these tasks. However, asking crowd workers (or people without any context) to make step-by-step plans is considerably more challenging for interdependent tasks that are embedded in a larger context (e.g. comments or to-dos in a document, to-dos in a codebase, or long-term health goals) [29, 39]. For these tasks, prior work has explored friend-sourcing as an alternative, wherein, given a to-do or health goal, people's friends make plans for how these can be accomplished [8, 36, 38]. This allows greater context and personalization as friends can craft their plans based on their personal knowledge about the individual, which can yield gains in user satisfaction [7]. While friend-sourcing has shown potential for both health behavior change planning [8] and information seeking [36], it comes with social costs: people do not like disclosing personal health-related information to friends, or they do not want their friends to complete the hardest aspect of a task [8, 38]. Additionally, asking people with context to make plans can result in plans that are not broken down to basic steps since friends or collaborators can assume that the original user has the same contextual information (tacit knowledge issues [40]). In general, current research suggests that individuals are generally satisfied with the quality of outsourced action plans whether they're created by a crowd [7, 8, 29, 51] or by a machine [48].

Our study aims to translate the benefits of having a plan for accomplishing tasks into interdependent task domains (such as writing, coding, and behavior change), while mitigating some of the costs and quality concerns associated with planning for these domains. Recent work on planning and decomposing tasks has used writing as the application domain because of the inherent interdependency of tasks in writing a document. Most to-dos or tasks being written in a document are context-embedded and require having awareness of the entire document to be accomplished. Therefore, we also choose writing as our test domain for making actionable plans for tasks.

2.3 Decomposing Writing Tasks

While there are various models of writing tasks, Flower and Hayes's cognitive process model [17] characterizes writing as distinctive cognitive processes that correspond to various actions that the author needs to perform when writing. These are often hierarchical in nature, and there can be several similar hierarchical processes for the different stages of writing.

We see evidence of this kind of decomposition in prior work on microtasking that enables crowd workers to help write documents. Kittur et al. [28] present a Partition-Map-Reduce framework to break down article writing, decision-making, and science journalism tasks by: (i) partitioning the original task into a series of subtasks, (ii) mapping each subtask to crowd workers, and (iii) reducing subtasks by aggregating the results of various subtasks to complete the original task. Soylent [11] similarly breaks down open-ended writing tasks by using a Find-Fix-Verify pattern: (i) workers are asked to find the issues in the text provided for the task, (ii) workers revise the text to fix the most popular issue identified in the first step, and (iii) workers verify the fix done in the second step.

WearWrite [37] leverages this task decomposition and helps authors capture new writing ideas via their smartwatch while employing crowd workers to jot down and summarize them in a document.

The above vein of prior work focuses on decomposing writing tasks to enable crowd workers to help authors (i.e., people responsible for accomplishing the task). In doing so, it puts a burden on the authors because they have to create microtasks for crowd workers by breaking down the bigger task [28]. Much like this prior work, our goal is to leverage the context that the authors and collaborators of a document have, and use it to generate actionable plans for the writing tasks. Additionally, we aim to reduce the time, effort, and social costs associated with doing this planning while improving plan quality. In this paper, we present a potential solution to solving this problem.

3 APPROACH

Our goal is to outsource the generation of action plans for writing tasks to people with necessary context (e.g. collaborators on an academic paper, readers of a Wikipedia article), while reducing the time and effort of planning and improving plan quality. We know from prior work (e.g. [10, 12]) that people take less time selecting from a list of options than they would on creating the list themselves. Intuitively, selection would also require less effort than creating the list. Thus, to reduce the time and effort costs of planning, one approach could be to ask people who are making the action plan to select from a list of steps, rather than generate the list on their own.

Action plans are most effective when they are broken down to behavioral primitives [50], especially when planning for someone who does not have the same knowledge about the task as the planner. However, when asked to list the steps of performing a task, people often do not list them to the most primitive task level. One reason for this is that some of the tasks are so routine to people that they become tacit knowledge [40]. Thus, asking people to come up with action steps without any guidance could lead to inadequate breakdown.

Our proposed approach is to generate a list of functions a priori—this list can then be used by planners to come up with their step-by-step plans. We call this list of functions a vocabulary: it is a finite set of writing functions (i.e., primitives) that are used to complete writing tasks. It serves as a cognitive scaffold that supports structured interaction between people with necessary context (in both high and low domain knowledge settings) and people who are responsible for accomplishing a task. The general list of functions makes it easy for people making the plan and people using it to communicate via a consistent, structured list of potential subtasks.

Our vocabulary is similar to a domain-specific language (DSL)—DSLs provide "a notation tailored towards an application domain and are based on the relevant concepts and features of that domain [46]." DSLs are often constructed using Unified Modeling Language [30, 41], which explains the role of each class (i.e., feature) of the application domain using variables and functions, and also includes the hierarchical relationships among these classes. Similarly, our vocabulary has some core classes corresponding to different types of writing tasks, and each class contains functions that can be applied to accomplish the task. For example, the "Adding Content" class contains functions such as, "Add a sentence about [this]," "List a few things about [this]," and, "Add a Table/Figure for [this]." Our vocabulary classes do not include any variables since the text highlighted per comment serves this purpose. In this way, our planning approach and methodology can be applied to other domains where such DSLs or vocabularies can be created. Prior research has proposed similar approaches in alternative domains with success, e.g. bolstering narrative structure in video-based storytelling via a taxonomy of common expert patterns [26]. The primary differences between [26] and ours lie in the intricacies of the domain and the purpose of the research: Kim et. al [26] focus on using expert patterns to enable novices to create strong narrative structures in video-based storytelling, while our work focuses on using common writing functions to construct a varied range of actionable writing plans with the goal of writing or improving a document.

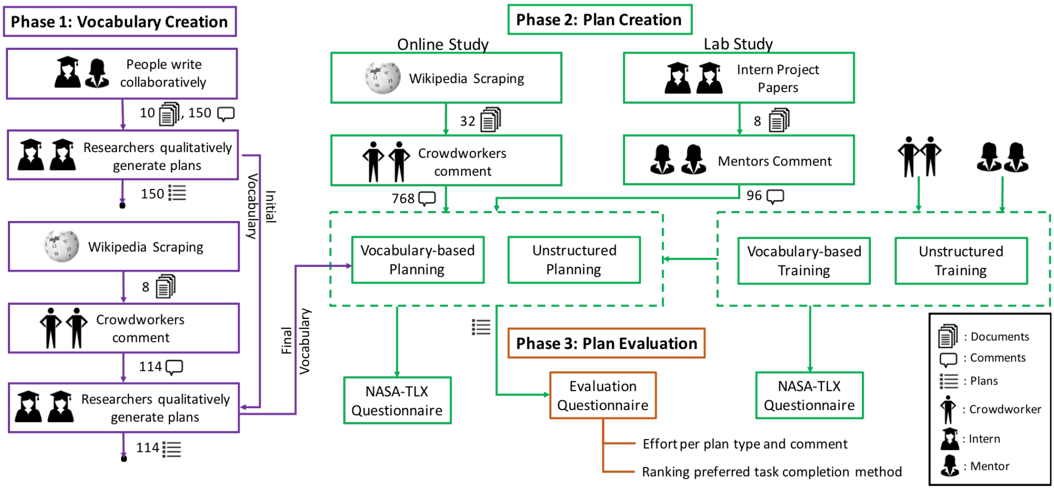


Fig. 1. Overview of all phases of our study.

3.1 Measures of Success

We compare our vocabulary-based planning approach to an unstructured planning process, wherein plans are made with no oversight or defined process, using five measures: *time, effort, granularity, atomicity,* and *sourcing potential.*

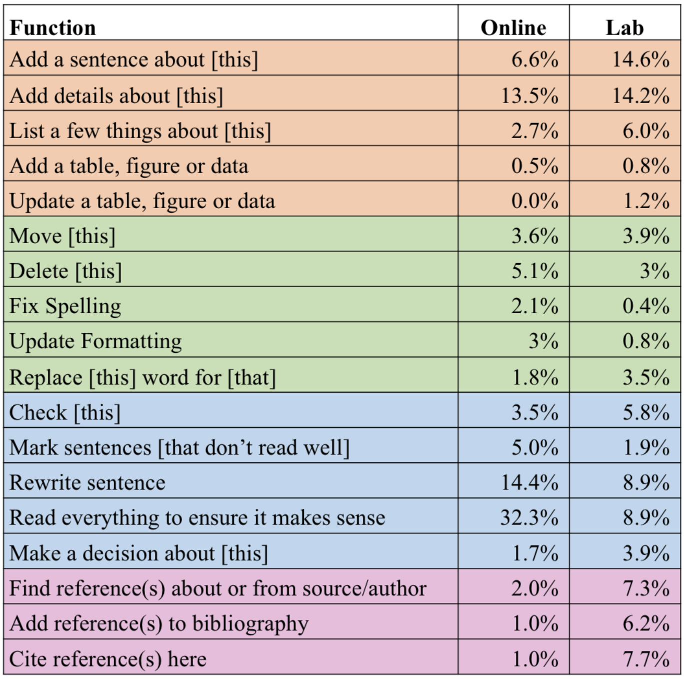
To evaluate whether we can successfully make planning less time-consuming, we measure the time it takes for people to make plans and compare it for conditions with and without vocabulary. To measure planning effort, we use the NASA-Task Load Index (TLX) questionnaire [6, 22], which includes six Likert scale questions (range 1-10, Very Low - Very High) about a task, pertaining to: (i) mental demand, (ii) physical demand, (iii) pace of the task (how hurried or rushed it felt), (iv) task performer's perception of success, (v) level of hard work required, and (vi) level of insecurity, discouragement or annoyance with the task. We do not ask the physical demand question in our study, as our tasks do not include any physical components.

For a plan to be good, reducing the time and effort necessary to create must not come at a loss of quality. Since the primary benefits of having an action plan are getting started on and accomplishing macrotasks, one way to measure plan quality is by checking how well the original task is broken down: the more broken down a task, the easier it should be to get started on it and accomplish it, and it should be easier for people other than the author to do the task. We use three proxies for this: (i) the number of steps (i.e., microtasks) generated per plan (granularity); (ii) the number of steps of the plan that are mechanical, i.e., can be done with little cognitive effort and need for context (atomicity); and (iii) the number of steps that can be crowd- or self-sourced, i.e., can be done in micromoments (sourcing potential). More granular and atomic plans with high sourcing potential can benefit people in ways highlighted in the Related Work section above.

Below, we describe the three phases of our project that implement our approach: (i) vocabulary creation, (ii) plan creation, and (iii) plan evaluation (Figure 1).

4 PHASE 1: VOCABULARY CREATION

Our first step is to create a vocabulary—our cognitive scaffold for writing tasks, which consists of the most basic functions used during writing tasks similar to a DSL. We followed an inductive, data-driven qualitative coding process to generate this vocabulary.



Function	Online	Lab
Add a sentence about [this]	6.6%	14.6%
Add details about [this]	13.5%	14.2%
List a few things about [this]	2.7%	6.0%
Add a table, figure or data	0.5%	0.8%
Update a table, figure or data	0.0%	1.2%
Move [this]	3.6%	3.9%
Delete [this]	5.1%	3%
Fix Spelling	2.1%	0.4%
Update Formatting	3%	0.8%
Replace [this] word for [that]	1.8%	3.5%
Check [this]	3.5%	5.8%
Mark sentences [that don't read well]	5.0%	1.9%
Rewrite sentence	14.4%	8.9%
Read everything to ensure it makes sense	32.3%	8.9%
Make a decision about [this]	1.7%	3.9%
Find reference(s) about or from source/author	2.0%	7.3%
Add reference(s) to bibliography	1.0%	6.2%
Cite reference(s) here	1.0%	7.7%

Table 1. Our vocabulary of writing functions and the percentage of times each function is used in our two studies. The functions are grouped by purpose: adding content (top), surface-level issues (second), editing content (third), and references (bottom).

We collected 10 documents written for academic audiences: five summer project descriptions written by interns at a large technology company, and five nearly complete drafts of papers being submitted to various HCI conferences. For the five intern project descriptions, we asked the project mentors to provide feedback in the form of comments on these documents, whereas for the nearly complete drafts this feedback was already included in the drafts collected. Our dataset for this part comprised of 150 comments.

To generate the vocabulary, we made step-by-step action plans for the tasks specified in the 150 comments in our dataset. Per Zacks et al.'s [50] suggestion, we recursively broke down each step of a plan until it was the most primitive function we could identify (i.e., no further non-trivial breakdown was possible). We created a list of functions used for each step, and iterated over it after making plans for each document, to remove redundant functions and break functions down to more primitive forms. After each iteration, we updated the plans per document in accordance with the new list. We followed this inductive, iterative qualitative process until we had a list of functions that could make plans for comments in 9 out of 10 documents. We used the final document as a test document which we also used to conduct an inter-rater reliability test for making plans using our function list (i.e., our vocabulary). The first author and a collaborator made plans for comments in the final document using our vocabulary, and calculated inter-rater reliability on these plans. We found that the plans have substantial agreement (Cohen's Kappa of 0.72), thus corroborating the use of our vocabulary for making plans for these high domain knowledge academic writing tasks.

4.1 Vocabulary Expansion

To account for the differences in various types of writing, we expanded our vocabulary using more general writing instances: Wikipedia articles. We picked the top five most popular Wikipedia categories (Popular Culture, Geography, Arts, History, and Current Events) [47] and queried Wikipedia databases to get articles belonging to these categories. We only picked articles graded as "Start" (incomplete articles still in development phase) or "C" (substantial articles, but missing important content and containing irrelevant material) according to Wikipedia's grading scheme because these are the two longest stages in the life cycle of a Wikipedia article [3, 5]. Additionally, we ensured that the selected articles were at least one page long to get a meaningful amount of text. We selected two articles per topical category (one Start and one C class), but were unable to find any that overcame the constraints in the Current Events category. Thus, we used eight articles (one Start and one C class) from the remaining four categories to expand our vocabulary.

For each article selected, we recruited four Amazon Mechanical Turk workers to provide comments (a total of 4x8 = 32 workers, pay = $1.50 at $11/hour). We instructed workers to leave at least two comments with editing tasks pertaining to each of the following categories: (i) mechanics, or surface-level details, such as grammar, spelling, or presentation of repetitive ideas; (ii) organization, or how the content is structured into various sections and paragraphs; and (iii) semantics, or meaning-making, ensuring that the content makes sense, explains the topic, and has no missing details. These categories and definitions are borrowed from the rhetorical writing categories identified by Greer et al. [21]. This setup ensured that the comments on these Wikipedia articles were balanced across the different types of writing tasks.

The Mechanical Turk setup above provided a dataset of 114 comments after filtering out duplicates and poor-quality comments (such as those with poor sentence structure and grammatical errors, comments that do not have actionable editing tasks, etc.). Following the same inductive, iterative qualitative process as above, we made plans for each of these comments, looking to expand our vocabulary for this general writing setting. After this process, only one new function was added to our vocabulary—"Fix Spelling", making our final vocabulary 18 functions long (Table 1).

5 PHASE 2: PLAN CREATION

Now that we have a vocabulary of functions for writing tasks, we compare plans made using this vocabulary with unstructured plans via two studies: (i) an online study using Wikipedia articles with Amazon Mechanical Turk workers as people with necessary content (general writing, low domain knowledge setting), and (ii) a lab study using project documents from intern-mentor pairs at a large technology company (academic writing, high domain knowledge setting). Both studies contain three steps: obtaining documents, leaving comments on documents, and making plans for the tasks specified in the comments.

5.1 Online Study

Obtaining documents. As before, we used Wikipedia articles for commenting and planning purposes. However, we only used Start class articles because in the Vocabulary Expansion study we found C class articles to be too well-written for generating comments using people who did not have domain knowledge. We picked the top 16 topical categories and two Start class articles for each category (total 32 articles).

Leaving comments. We asked crowd workers to read and leave comments on the selected Wikipedia articles (pay = $1.50). To get a consistent number of comments throughout the article, we divided each article into two sections. Each section was assigned to two crowd workers, and each crowd worker was asked to leave six comments (as before, two comments each for issues related to mechanics, organization, and semantics). This resulted in a total of 768 comments generated by 128 crowd workers (24 comments per article).

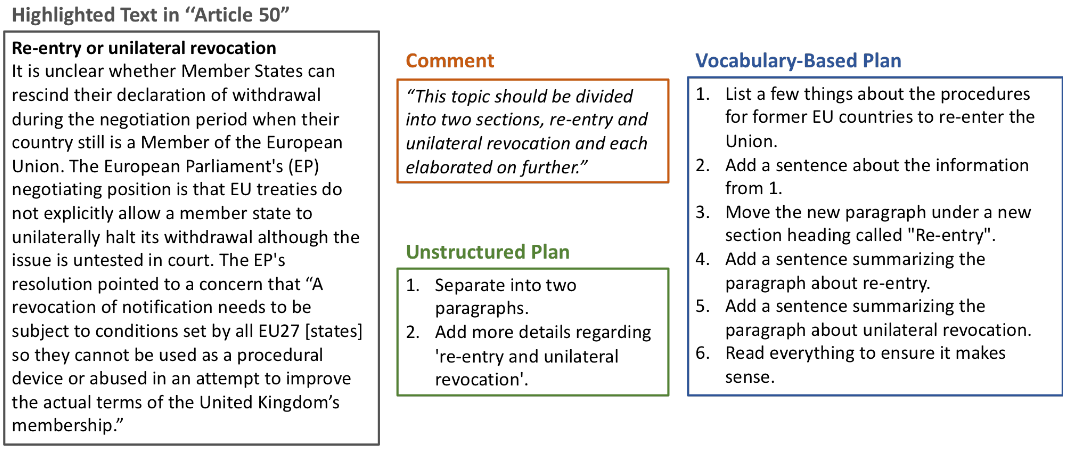
Making plans. We asked a different set of crowd workers to make plans for addressing each comment per Wikipedia article. Contributing to Wikipedia is possible even with low domain expertise: a large number of contributions on Wikipedia are a result of the "wisdom of crowds" [27]. We thus considered these crowd workers as proxy Wikipedia editors/collaborators. The task was presented to crowd workers in a survey format, built using the SurveyGizmo tool. We assigned workers to one of two conditions: unstructured planning or vocabulary-based planning. Both conditions had the same steps, modified slightly based on the condition. First, we asked crowd workers to skim the article in 10 minutes to get some context about the topic and the article itself. This was done to further establish their role as collaborators of a Wikipedia article. Second, crowd workers in both conditions went through a training to make plans. Our training setup was inspired by Doroudi et al.'s work [16] on training crowd workers to accomplish complex tasks. For the vocabulary condition, we trained crowd workers by providing our vocabulary functions and their descriptions, and showing them example plans made using the vocabulary. We also asked them to make a sample plan, and showed them our plan for the same comment to validate their understanding of the use of our vocabulary. For the unstructured condition, we had no aid for the crowd workers, but we showed them the same example plans as those in the vocabulary condition. In both conditions, crowd workers stayed in this training setup for 1.5 minutes—the ”Next” button was disabled for this time. After training, crowd workers answered the NASA-TLX questionnaire for the training task. Data from these questionnaires was used to measure effort per condition.

Next, each crowd worker was assigned to one out of two sections per article, and was asked to make plans for the 12 comments left by other crowd workers in that section. While in the vocabulary condition, the crowd workers had access to the vocabulary function list as they were making plans; in the unstructured condition, the interface simply asked them to make a step-by-step plan for each comment. After planning, crowd workers filled out another NASA-TLX questionnaire—this time about the planning task. We added one more question in the vocabulary condition: Imagine you already knew the vocabulary (i.e., you did not have to learn it), how mentally demanding do you think making plans would be in that case? Finally, crowd workers in both conditions answered some open-text questions about the entire process, and were compensated with $6 for the entire study (pay rate of $11/hour).

Dataset. Each article was planned for by four crowd workers (two per section; one for unstructured and the other for vocabulary-based planning). Workers were randomly assigned to an article section and a planning condition. Due to a skew in the random assignment process of our survey tool, SurveryGizmo, we generated 840 unstructured plans and 900 vocabulary-based plans. In comparing the two conditions, we ensured that the pairwise comparison was done for each unstructured-vocabulary plan pair, giving us a total of 1014 comparison points.

5.2 Results: Online Study

Figure 2 presents an example comment from our online study, and the unstructured and vocabulary-based plans generated for it. In general, the use of vocabulary functions follows a long tail distribution when planning for Wikipedia articles: there is a subset of functions that is used more frequently than the others (see Table 1 for numbers). Checking the flow of a paragraph, rewriting sentences, and adding details are the most popularly used vocabulary functions for the online study. On the other hand, adding references and tables are used rarely for these articles. Given that the Wikipedia articles selected are all graded as Start class (i.e., they lack the required level of details about the topic and are not well-organized), this distribution of the vocabulary usage is not unexpected. To understand how vocabulary based planning compared to unstructured planning, we



Highlighted Text in "Article 50"

Re-entry or unilateral revocation
It is unclear whether Member States can rescind their declaration of withdrawal during the negotiation period when their country still is a Member of the European Union. The European Parliament's (EP) negotiating position is that EU treaties do not explicitly allow a member state to unilaterally halt its withdrawal although the issue is untested in court. The EP's resolution pointed to a concern that "A revocation of notification needs to be subject to conditions set by all EU27 [states] so they cannot be used as a procedural device or abused in an attempt to improve the actual terms of the United Kingdom's membership."

Comment

"This topic should be divided into two sections, re-entry and unilateral revocation and each elaborated on further."

Unstructured Plan

1. Separate into two paragraphs.
2. Add more details regarding 're-entry and unilateral revocation'.

Vocabulary-Based Plan

1. List a few things about the procedures for former EU countries to re-enter the Union.
2. Add a sentence about the information from 1.
3. Move the new paragraph under a new section heading called "Re-entry".
4. Add a sentence summarizing the paragraph about re-entry.
5. Add a sentence summarizing the paragraph about unilateral revocation.
6. Read everything to ensure it makes sense.

Fig. 2. An example case from our study. The Wikipedia document is called "Article 50 of the Treaty on European Union." We present the highlighted text, comment, and the unstructured and vocabulary-based plans created for this case in our online study.

use our five measures of success (defined in previous sections): time, effort, granularity, atomicity, and sourcing potential.

5.2.1 ***Time***. People take 120 seconds to make an unstructured plan and 87 seconds to make a vocabulary-based plan, on average (s.d. unstructured = 210, vocabulary = 113). The median times for these plans are 83 seconds and 57 seconds, respectively.

The results of a linear mixed-effects model indicate that vocabulary-based planning takes significantly less time compared to unstructured planning ($p < 0.01$). We created the model with time as the dependent variable, unstructured vs. vocabulary condition as the independent variable (fixed effects), and participants as the random effects. Participants were selected as random effects since each participant provided 12 data-points in our dataset—each planned for 12 comments. This implies that our vocabulary benefits planners by reducing the time it takes to make plans.

In addition to comparing the time taken for unstructured vs. vocabulary planning, we are also interested in learning whether planning gets easier over time. This is particularly relevant in the vocabulary condition since there is no longitudinal training with the vocabulary—we ask crowd workers to use it immediately after being introduced to it for the first time. Figure 3 presents a time-series for Plans 1-12 for all individuals per condition. This figure shows that: (i) vocabulary-based planning consistently takes less time than unstructured planning; and (ii) there is a decreasing trend in planning time for the vocabulary condition, whereas the unstructured condition is more variable. Our hypothesis is that using the vocabulary gets easier over time due to a learning effect. This is supported by some participant quotes:

> "I think I got the hang of the step-by-step process for addressing comments. It got easier as I went on." (P7)
>
> "I liked having the guide [vocabulary] to make the instructions from. It took a while, but I got used to it eventually." (P108)

5.2.2 ***Effort***. We use the NASA-TLX questionnaire data to compare the effort required to train and plan in the unstructured vs. vocabulary conditions. To make these comparisons, we conduct Mann-Whitney's U tests on each of our NASA-TLX questions for both training and planning tasks.

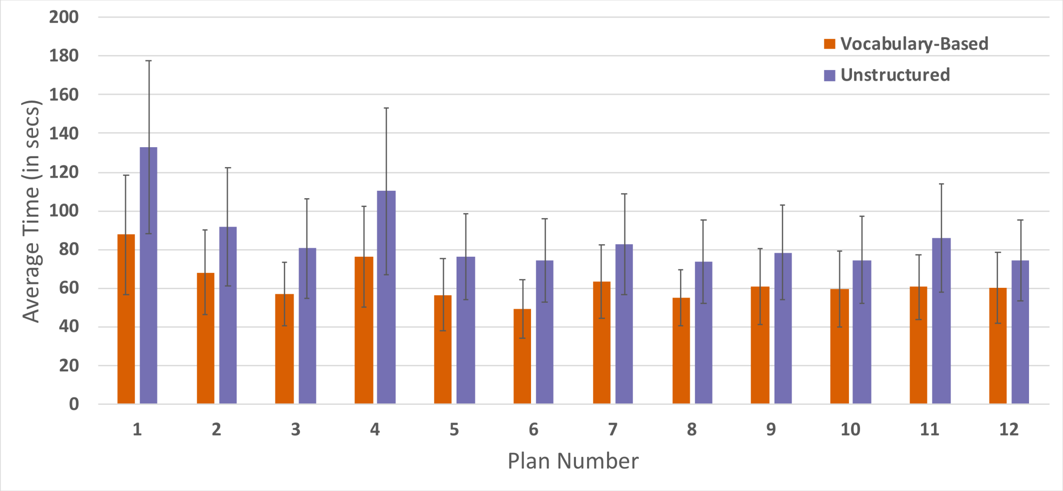


Fig. 3. Time taken for planning from Plan 1 to Plan 12. Time spent decreases as planners go from Plan 1 to 12. Vocabulary-based planning takes less time than Unstructured planning at all points.

For the training task, people felt significantly more hurried by the pace of the unstructured training compared to the vocabulary training (p-value < 0.005), and significantly more insecure and annoyed during the unstructured training (p < 0.05). We did not expect this since the unstructured training did not involve as many components as the vocabulary training. However, we hypothesize that this outcome is due to differences in the level of engagement in the two trainings: the unstructured training simply presented some example plans and asked people to read through these carefully, whereas the vocabulary training not only provided the example plans, but also included a training exercise. People made a sample plan for an example comment and were able to see our solution for the same comment once they submitted their plan. This interactivity in vocabulary training might have caused the training task to seem less rushed, since people progressed through it with explicit feedback from us.

For the planning task, there were two significant differences out of the five NASA-TLX questions: (i) people felt that they had to do significantly more hard work for making unstructured plans than vocabulary-based plans (p < 0.05), and (ii) people felt significantly more insecure and annoyed when making unstructured plans than vocabulary-based plans (p < 0.01). We also found a significant difference in the vocabulary condition-only question: people felt that if they already knew the vocabulary (i.e., they did not have to train for it or learn it), planning would be significantly less mentally demanding (p < 0.01). These results indicate that vocabulary-based planning requires either comparable or less effort than unstructured planning, thus providing evidence for our claim that a vocabulary can help reduce the effort costs of planning.

5.2.3 ***Granularity.*** We measure granularity as the number of steps (i.e., microtasks) per plan, and use it as a proxy for plan quality. People make unstructured plans with 2 steps and vocabulary-based plans with 2.5 steps, on average (s.d. unstructured=1, vocabulary=1). The median granularity for both unstructured and vocabulary-based plans is 2 steps.

The results of a linear mixed-effects model indicate that vocabulary-based plans are significantly more granular, i.e., have significantly more steps than unstructured plans (p < 0.005). We created the model with granularity as the dependent variable, unstructured vs. vocabulary condition as the independent variable (fixed effects), and participants as the random effects. Once again, participants were selected as random effects since each participant provided 12 data-points in our dataset—each

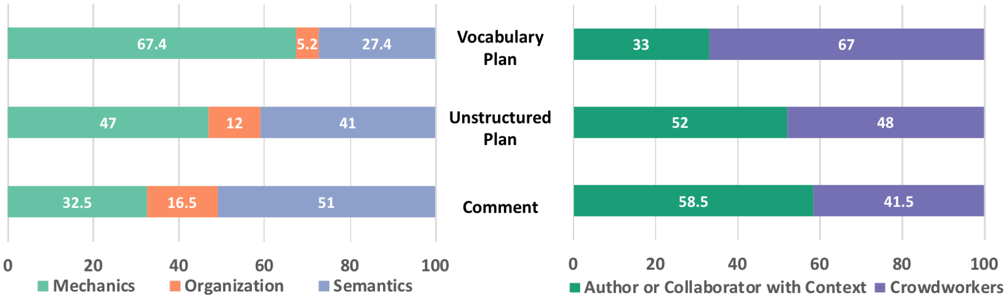


Fig. 4. Atomicity (left) and Sourcing Potential (right) for Online Study.

participant planned for 12 comments. This implies that a vocabulary improves plan quality in terms of generating plans that are broken down into more steps than unstructured plans.

5.2.4 ***Atomicity***. We qualitatively code a random sample of 200 comments, unstructured plans and vocabulary-based plans for whether each comment or step is related to writing mechanics, organization or semantics. This qualitative coding is done by the first and fifth author; inter-rater reliability is calculated using additional 20 comments and plans—Cohen's Kappa is 0.77 (significant agreement). To measure atomicity, we calculate the number of steps in unstructured and vocabulary-based plans that are mechanical, organizational, and semantic: a higher number of mechanical steps means higher atomicity.

Figure 4 shows the percentage of comments and steps of unstructured and vocabulary-based plans that belong to each category mentioned above. There is a ~35% increase in mechanical steps between vocabulary-based plans and comments, and a ~20% increase between vocabulary-based and unstructured plans (values depict absolute improvement). These results imply that there is strong qualitative evidence in favor of the claim that vocabulary-based plans are more atomic than unstructured plans. The vocabulary helps people generate more atomic steps by breaking tasks down to the more basic functions.

5.2.5 ***Sourcing Potential***. We qualitatively code the same random sample of 200 comments, unstructured plans, and vocabulary-based plans used above for whether each comment or step can be completed only by someone with domain knowledge about the topic (i.e., author or collaborator) or by a crowdworker with minimal-to-no domain knowledge. Once again, we calculate inter-rater reliability between the first and fifth author using the same additional set of 20 comments and plans; Cohen's Kappa is 0.84, showing significant agreement.

Figure 4 shows the percentage of comments and plan steps that belong to the categories mentioned above. There is a ~26% increase in crowdsourcing potential between comment and vocabulary-based planning, and a ~19% increase between unstructured and vocabulary-based planning. This implies that having a vocabulary when planning leads to planners using steps that can be crowdsourced. The author or a person with domain knowledge about the topic need only accomplish a small percentage of the steps in a vocabulary-based plan, the rest can be crowdsourced.

5.3 Lab Study

Research has demonstrated the challenge of collaborative writing for academic papers [45]. While our online study setup is a proxy for a low domain knowledge setting that involves a collaboratively written document, we still lack an understanding of the high domain knowledge situation where planning needs to be done by people with considerably more topical expertise about the artifact being written. We conduct an exploratory lab study with 8 intern-mentor pairs to get this insight. The lab study setup mirrors the online study setup, but is conducted in person with authors and collaborators of an academic document. It provides more contextualized observational data about planning with people who have high domain knowledge about the project, and the use of our vocabulary. We compensated interns and mentors with $20 gift cards each.

Obtaining documents. We collected documents from interns at a large technology company: they wrote an initial draft of a paper about their project for submission to a conference.

Leaving comments. We asked each intern's mentor to leave comments on their project draft. We provided similar instructions as the online study for leaving comments—that the mentors leave a total of at least 12 comments, mostly semantic, some organizational, and very few mechanical.

Making plans. We asked mentors to make plans for addressing each of their comments in a lab setting. The lab study differs from the online study in that mentors do both unstructured and vocabulary-based planning. To avoid biasing the mentors to our vocabulary functions, they are made to participate in unstructured planning at least one day prior to vocabulary-based planning, both in a lab setting. The lab study mirrors the online study in that there is a training step, a NASA-TLX questionnaire about the training step, a planning step where mentors make unstructured and vocabulary-based plans for their comments, a NASA-TLX questionnaire about the planning step, and an open-text questionnaire about the process. We also collect observational data throughout the planning process, and ask semi-structured questions based on this observational data.

Dataset. Our dataset included 96 comments (8 papers x 12 comments), thus resulting in 96 unstructured and 96 vocabulary-based plans.

5.4 Results: Lab Study

The usage of vocabulary functions in the lab study is similar to that of the online study: the top four functions used are the same for both (see Table 1 for numbers). The main difference in the lab study is that more functions pertaining to references are used, and more of the functions are used. Below, we compare vocabulary-based plans and unstructured plans, using the same analyses for the lab study as we did for the online study, unless specified otherwise.

5.4.1 ***Time.*** People take 96 seconds to make an unstructured plan, and 91 seconds to make a vocabulary-based plan, on average (s.d. unstructured=80, vocabulary=72). The median times for unstructured and vocabulary-based plans are 74 seconds and 72 seconds, respectively. The results of a linear mixed-effects model with time as the dependent variable, unstructured vs. vocabulary condition as the independent variable (fixed effects), and participants as the random effects indicate that, on average, vocabulary-based plans take less time to make than unstructured plans, but this difference is not significant.

5.4.2 ***Effort.*** We used the NASA-TLX questionnaire to compare the effort required to plan in unstructured vs. vocabulary conditions. Since our unstructured and vocabulary-based planning effort data comes from the same participants for the lab study, we conducted Wilcoxon Signed Rank tests for each NASA-TLX question to compare these two conditions. We find no significant difference in any of the NASA-TLX questions for unstructured vs. vocabulary-based planning, i.e., the effort required to make either type of plan is not significantly different. We do find significant difference in the additional vocabulary condition-only question: people felt that if they already knew the vocabulary (i.e., they did not have to train for it or learn it), the planning task would be significantly less mentally demanding (p-value « 0.05).

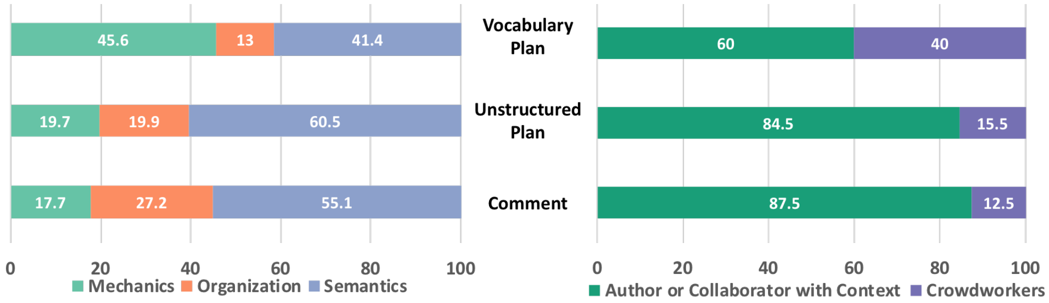


Fig. 5. Atomicity (left) and Sourcing Potential (right) for the Lab Study.

5.4.3 ***Granularity***. We measure granularity as the number of steps (i.e., microtasks) per plan, and use it as a proxy for plan quality. People make unstructured plans with 2 steps and vocabulary-based plans with 3 steps, on average (s.d. unstructured=vocabulary=1.5). The median granularity for unstructured and vocabulary-based plans is 2 and 3 steps, respectively. The results of a linear mixed-effects model with granularity as the dependent variable, and unstructured versus vocabulary condition as the independent variable (fixed effects), and participants as the random effects show that, on average, vocabulary-based plans have more steps than unstructured plans, but this difference is not significant.

5.4.4 ***Atomicity***. We qualitatively coded the 96 comments, unstructured plans and vocabulary-based plans from our lab study for whether they were related to mechanics, organization, or semantics. The higher the number of mechanical steps, the higher the atomicity, since mechanical steps are more atomic than organizational or semantic steps. Figure 5 shows the percentage of comments and the steps of unstructured and vocabulary-based plans that belong to each category above. There is a ∼28% increase in mechanical steps between vocabulary-based plans and comments, and a ∼26% increase between vocabulary-based and unstructured plans (values depict absolute improvement). Similar to the online study, the results of the lab study indicate that the vocabulary helps people break tasks down to more basic functions.

5.4.5 ***Sourcing Potential***. We qualitatively coded the 96 comments, unstructured plans and vocabulary-based plans for whether each comment or step can be completed only by someone with domain knowledge about the topic (i.e., author or collaborator) or by a crowdworker with minimal to no domain knowledge. Figure 5 shows the percentage of comments, and the steps of unstructured and vocabulary-based plans that belong to the categories mentioned above. There is a ∼28% increase in crowdsourcing potential between comment and vocabulary-based planning, and a ∼25% increase between unstructured and vocabulary-based planning. This implies that having a vocabulary when planning leads to planners using more steps that can be crowdsourced. The author or a person with domain knowledge about the topic need only accomplish a smaller percentage of the steps in a vocabulary-based plan.

5.5 Summary

Overall, we show that using the vocabulary significantly reduces planning time and effort in our online study, whereas for the lab study, time and effort remain comparable to unstructured planning. Additionally, vocabulary-based plans are more atomic and have higher sourcing potential in both studies. Plan granularity (i.e., number of steps per plan) is significantly higher when using the vocabulary for our online study, and comparable for the lab study. These findings imply that vocabulary-based planning is comparable or better than unstructured planning for our five measures

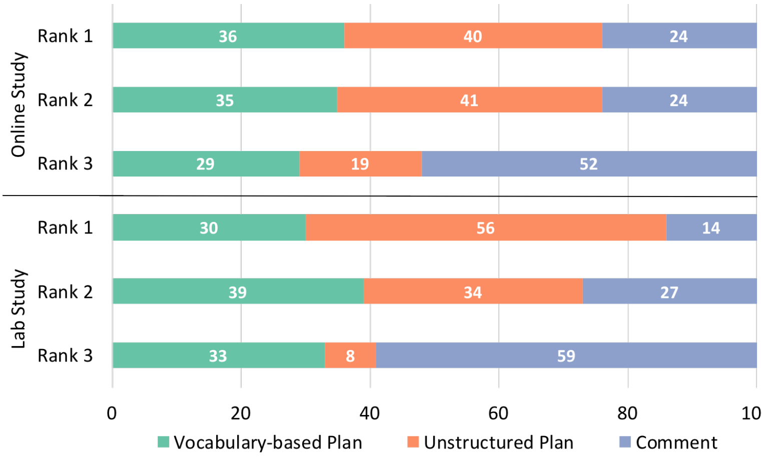


Fig. 6. Preference ranking for comments and plans for both the online and the lab study.

of success (Table 2). The vocabulary succeeds as a cognitive scaffold, both in terms of reducing costs and in terms of providing a structured form of interaction between people with necessary domain knowledge and people who are responsible for accomplishing the original task being considered.

6 PHASE 3: PLAN EVALUATION

6.1 Methods

Thus far, our studies have focused on comparing the creation of unstructured and vocabulary-based plans from the perspective of the person doing the planning. In this section, we present an evaluation of plans generated with the proposed vocabulary from the perspective of the person responsible for addressing the comments and writing, often the author of the document. Like the prior section, our evaluation is conducted in both an online setting and a lab setting.

For the online study, we once again use crowd workers as a proxy for Wikipedia editors since anybody can edit Wikipedia articles [27]. We randomly sample 100 comments from our dataset of 768, and ask crowd workers to rank the comment, and the unstructured and vocabulary-based plans in order of preference; and explain their rankings. We get three different crowd workers to rank each of the 100 comments (300 crowd workers total, paid $0.70 per task: a rate of $11/hour).

For the lab study, we ask interns to rank the comments and plans left by their mentors. We survey interns using the same set of questions asked in the online study plan evaluation above: a preference ranking of options (comment, unstructured plan, or vocabulary-based plan); and an explanation of their rankings. Instead of only asking these questions for one comment (which was the case in the online study evaluation above), we ask interns to answer these questions for all 12 comments left on their document. This gives us insight into the preferred method of addressing the task according to the authors of a document, in addition to the crowd worker feedback used as a proxy in the online study.

6.2 Results

Figure 6 shows the results of this preference ranking for both the online study and the lab study. One prominent theme in this figure is that planning of any sort is preferred over completing the task using the original comment, which was most frequently ranked third (least preferred). Between unstructured and vocabulary-based plans, unstructured plans are often better-ranked, especially for early-stage documents that have more comments about deliberating on the topic.

Vocabulary-based plans have consistent percentages for all three ranks. This could mean that there is higher variability in people's preferences when it comes to these plans.

One potential reason for the variability in rankings for vocabulary-based plans could be that not all planners edit the steps of vocabulary-based plans to add additional context to the vocabulary functions used. For example, a plan-step could be about adding a sentence. Some crowd workers simply used the vocabulary function and used "Add a sentence here" as the plan step, whereas others included contextual information, such as "Add a sentence about the remittance of Biscari Massacre to fit it into the summary." Differences like these could have caused some vocabulary-based plans to be highly preferred, whereas others to be less preferred because of the lack of meaningful information. We see evidence of this variability in people's explanations:

> "The original comment for this was a bit too broad and I would have liked more details. Vocabulary plan gives me a solid plan to write what my commenter might be looking for. This makes the task much easier and more manageable." (P7, lab)

While P7 appreciated the breakdown of a vocabulary-based plan, P3 felt it was too rigid for an early stage comment:

> "The amount of work here is quite large, and vocabulary plan instructions are much more rigid and are not exactly reflecting the comment's actions. E.g. 'we *might* fold Related Work into Intro' does not mean 'Delete the related work section."' (P3, lab)

Comments were ranked lower than either of the plans by most people. People explained their low ranking as follows:

> "Original comment was very vague." (P45, online)

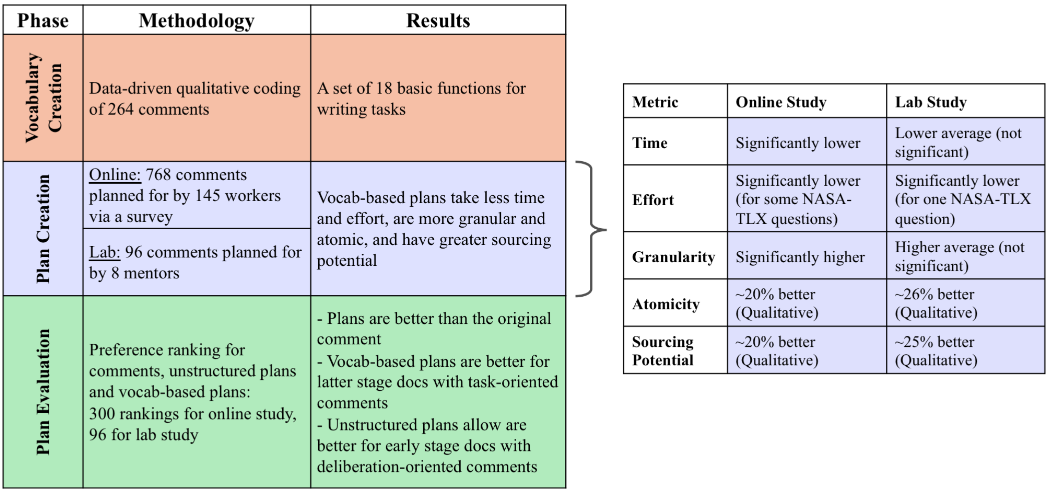
> "This comment is another pretty ambiguous one and thus any plan is helpful. The main thing with comments I find ambiguous is that I would like to ask the commenter a clarification question." (P7, lab)

6.3 Summary

Overall, people in charge of accomplishing tasks find action plans to be more helpful than the original comments. For early-stage documents, unstructured plans appear to be preferred more than vocabulary-based plans as they enable flexibility in creating plans for comments that are more reflection- or deliberation-oriented rather than task-oriented. For comments that are more concrete, vocabulary-based plans have the advantage of providing granular steps, some of which can be carried out by people other than the author (summary in Table 2).

7 DISCUSSION

Our investigation into creating action plans for writing tasks embedded in comments showed that having a predetermined vocabulary as a cognitive scaffold saves time and effort, and allows for the creation of plans that are comprised of more atomic steps. The vocabulary we used was generated using two types of documents, academic papers and Wikipedia articles, but our approach generalizes to other types of writing where we expect this vocabulary to be useful with little to no modification. Furthermore, it could be generalized to other domains where tasks are perceived in a similar, structured fashion [26]. In this regard, our vocabulary is similar to a domain-specific language (DSL) that provides a structured notation for an application domain. Our findings indicate that people prefer to address tasks using plans generated from comments instead of just the comments themselves. While we validate the effectiveness of the plans using preference rankings, in future work we would like to evaluate the quality of the outcome of following the plans in practice.



Phase	Methodology	Results
Vocabulary Creation	Data-driven qualitative coding of 264 comments	A set of 18 basic functions for writing tasks
Plan Creation	Online: 768 comments planned for by 145 workers via a survey	Vocab-based plans take less time and effort, are more granular and atomic, and have greater sourcing potential
	Lab: 96 comments planned for by 8 mentors	
Plan Evaluation	Preference ranking for comments, unstructured plans and vocab-based plans: 300 rankings for online study, 96 for lab study	- Plans are better than the original comment - Vocab-based plans are better for latter stage docs with task-oriented comments - Unstructured plans allow are better for early stage docs with deliberation-oriented comments

Metric	Online Study	Lab Study
Time	Significantly lower	Lower average (not significant)
Effort	Significantly lower (for some NASA-TLX questions)	Significantly lower (for one NASA-TLX question)
Granularity	Significantly higher	Higher average (not significant)
Atomicity	~20% better (Qualitative)	~26% better (Qualitative)
Sourcing Potential	~20% better (Qualitative)	~25% better (Qualitative)

Table 2. Left: Summary of results for all three phases of our project. Right: Results from the 5 metrics tested during Phase 2 (Plan Creation), highlighted from the perspective of vocabulary-based planning, i.e., "significantly lower" = Significantly lower for vocabulary-based planning compared to unstructured planning.

We used comments as a means of identifying tasks embedded in a document. This is valuable especially when considering collaborative writing practices, as collaborators may often leave comments instead of directly editing the content. Comments provide an easy access point to the tasks that need to be completed to improve a document—the tasks embedded in a comment are more actionable and accessible because of the highlighted text anchor provided along with the comment. However, as with any mechanism, this excludes some document- or meta-level tasks. That our vocabulary is created using data from tasks embedded in comments aligns with our results on the vocabulary being more useful for actionable tasks than tasks that require process-level deliberation. Process-level writing tasks are also more prevalent in the early stages of a document, which further corroborates our lab study results—that vocabulary-based, highly actionable comments are not as preferred in the initial stages of a document as open-ended plans because they do not capture that aspect of early-stage writing.

Given the contextual nature of writing and the unique differences in how different kinds of writing are done, we know that our vocabulary cannot be equally helpful to all the different kinds of writing tasks. Below, we discuss these different considerations for anybody trying to use our vocabulary in a generalized writing setting, and then propose a hybrid planning approach that generalizes the outcomes from this work.

7.1 Factors Affecting Vocabulary Use

While our vocabulary serves as a cognitive scaffold that reduces the time and effort for planning, several new patterns and hypotheses emerge for future considerations and testing. Our contrasting studies (Wikipedia vs. academic writing) are a unique setup for observing these differences and bringing them to light.

7.1.1 Stage of the Document. Mentors in our lab study found it particularly hard to plan for comments left on intern papers that were in the early stages of writing. Writing happens over several stages, usually going from ideation and organization / outlining, to writing paragraphs based on the outline [43, 45]. We hypothesize that the difficulty and context requirements for making action plans is higher when documents are in the early stages of writing, than when they are in the core writing and editing stages. The early stages often require deliberation to make progress, rather than active writing or editing tasks, making them less suitable for structured, vocabulary-based planning. In our lab study, 6 out of 8 documents were still in the earlier stages; mentors left comments that required further thinking about content and organization rather than actively writing. This, in turn, made it hard for them to convert their comments into action plans using our vocabulary—no significant improvement was found in time and effort spent in the unstructured vs. vocabulary-based planning conditions. Additionally, interns preferred the open-ended nature of the unstructured plans in some of these cases because those had steps to help with thinking through the problem, rather than simply focusing on the writing actions required as was the case with vocabulary-based plans. Thus, we believe that the stage of a document is an important consideration when thinking about the application of vocabulary-based planning.

7.1.2 Domain Knowledge. Our studies showed more promising results for writing that required low domain knowledge (Wikipedia articles) than high (academic articles). One potential reason for this could be that the level of domain-specific knowledge required for academic writing does not lend itself to creating consistent action plans. Aside from a template or common overall paper sections, most articles written for academic purposes have unique narrative and content requirements. Further, as has been studied before, while Wikipedia articles can be written by the crowd, the same is not true for academic articles. Even collaborators may not always have the same knowledge about a paper as the lead author. Given the high costs of engagement required to help with writing an academic paper, it seems plausible that the time and effort costs were not as significantly diminished as the low domain knowledge case, even when using a vocabulary.

7.1.3 Interpersonal Relationships. There has been a significant amount of prior work that highlights the importance of interpersonal relationships in collaborative tasks including writing (e.g. distributed authority [45]). Indeed, these differences affect the artifact being written more so in academic writing than Wikipedia: while Wikipedia editing is a more democratized and decentralized setup, academic writing often relies on an apprenticeship model. The goal of the latter is dynamic; depending on how experienced the author is, there are situations when the writing collaboration is an opportunity for the mentor to teach the mentee how to write in these settings. Certainly, this changes the kind of comments and tasks highlighted in the document and, consequently, the action plans made for these types of comments. In our study, we see cases where the comments and unstructured plans highlighted process-level considerations such as guiding the student to reflect on a certain problem before doing a writing task, or guided thinking process steps, but the vocabulary-based plans—being based on a small list of functions for active writing—were unable to capture this nuance.

In combination, these three factors suggest, perhaps intuitively, that making action plans using a vocabulary supports only a subset of the different kinds of writing tasks. Documents in the earlier stages of writing (e.g. ideation and outlining) require more deliberation and process-level comments, which do not lend themselves to "action" planning. However, once the documents reach a stage where concrete actions can be taken, our vocabulary is a significant cognitive scaffold that serves to reduce the time and effort costs of planning, while ensuring plan quality. The kind of collaboration this supports is when writing tasks are being generated considering all collaborators of the document as equal—the nuance of teaching and learning via writing can be lost with a task-based focus. For future work, we hope to extend this cognitive scaffolding approach to the thinking processes of writing.

7.2 Hybrid Planning Approach

Looking forward, we believe that vocabulary-based plans offer a pathway to automation by introducing structure and consistency into planning. In our studies, participants planned for 864 comments in general and academic documents using a vocabulary of only 18 functions. We saw that having a vocabulary helped break down a task into more mechanical steps, enabling more self- and crowd-sourcing. Once a comment was broken down into a list of multiple shorter steps using the vocabulary, many of the component steps required so little context that they could be accomplished by crowd workers. When crowdsourcing is not an option because of contextual needs, this high atomicity of steps enables authors to use micromoments to complete these steps via selfsourcing [44], or assign them to collaborators without having to communicate the big picture of the document [43]. It is likely that many of the steps could also be automated, especially because a significant amount of training data can be collected for each of the limited number of vocabulary functions as people use them. Additionally, it may be possible to automatically generate the underlying plans from the original comment. For example, Kokkalis et al. [29] introduce and test a natural language-based similarity algorithm for recycling plans made for high-level tasks. Our hope is that future work can enable this automation for complex, interdependent tasks such as comments in written documents.

Writing tasks obtained from comments in a document range from being about the mechanics and organization to the semantics of a document. Semantic tasks, especially those in early-stage documents, often involve deliberation and reflection. In these instances, comments are used as a means of initiating communication about the writing task rather than breaking it down. To handle this variability in the use of comments, we envision a hybrid planning approach that takes advantage of automation while keeping the human in the loop. With a hybrid approach, the planner has the autonomy to modify the automatically generated plans as need be, based on the task in the comment. In this way, the hybrid approach would allow people to be flexible and unconstrained, while still saving time and effort, and making plans that are more atomic and have more sourcing potential. A similar hybrid approach—assisted microtasking—has been successfully applied by Calendar.Help for scheduling via email [15].

A tool based on this hybrid approach would fit into people's mental model of writing [17, 49] because it maps directly to human cognitive processes. People perceive the world and make decisions based on two primary cognitive processes: automatic processes that do not require long-form, logical reasoning before taking an action; and deliberate processes that involve careful deductive reasoning and cognitive effort before an action is taken [18, 23]. Kahneman calls these two modes System 1 and System 2, respectively, and presents situations that exemplify the need for the automaticity afforded by System 1 (automatic processing), and the rational thinking of System 2 (deliberate processing) [25]. Depending on the task at hand, people switch between using System 1 and 2 without any conscious thought: System 1 is used when past instances of being in a similar situation or having learnt about it can be instantly recalled, and System 2 is used when no such past heuristics exist. Our proposed approach would play a similar role: it would use prior examples of plans created to automatically generate new plans, but allows planners to deliberate and create a new plan on their own depending on the task identified in the comment.

Overall, alongside prior work [26], our work reinforces the generalizability of a vocabulary-based approach for writing. By extracting structure from complex domains, such as writing, we enable task breakdown and planning in a more structured and efficient manner. Our work is a initial step towards easier planning for writing tasks, but there remain several unexplored complex domains where vocabularies could potentially be helpful, but none currently exist. While we hope our approach can help guide the creation of vocabularies in these other domains, the intricacies and considerations of these domains can vary. An important challenge for future research is generating vocabularies for these other context-rich domains: it falls upon researchers to generate these DSLs, vocabularies, or taxonomies, and validate them with studies like ours, so that others can build on this work and benefit from the structure extracted in these complex domains. We believe that domains where these DSLs can be generated, such as storytelling, programming, scheduling, or behavior change, can take advantage of our overall approach.

8 CONCLUSION

In this paper, we explore a mechanism for outsourcing plan creation for context-embedded tasks to people who have some context (e.g. collaborators of a written document). To facilitate the plan creation process, we develop a vocabulary of 18 basic functions that can be used to create plans. Compared to unstructured plans, we find that plans created using our vocabulary require less time and effort to be created, have more atomic steps and more steps that could be assigned to other people, thereby creating better plans that reduce the burden on the person who is tasked with executing the plan. However, we also learn that vocabulary plans seem less flexible compared to unstructured plans, especially for writing tasks pertaining to early-stage documents. We discuss the implications of our work in studying the use of a hybrid approach that incorporates the flexibility of unstructured planning and the structure of vocabulary-based planning, and automatically generates steps based on prior plans.

9 ACKNOWLEDGMENTS

We would like to thank Elena Agapie, Cliff Lampe, and Eric Gilbert for their valuable input throughout this work. We also thank the reviewers for their feedback; and finally, all of our study participants—crowd workers, and interns and mentors at Microsoft Research—for their help in making this research possible.

REFERENCES

[1] [n. d.]. Everyone. Every run. ([n. d.]). https://runkeeper.com/

[2] [n. d.]. Kitchen inspiration and weekly meal plan service. ([n. d.]). https://www.cooksmarts.com/

[3] [n. d.]. Life of an Article. ([n. d.]). https://outreach.wikimedia.org/wiki/Life_of_an_Article

[4] [n. d.]. Money, Bill Pay, Credit Score Investing. ([n. d.]). https://www.mint.com/

[5] [n. d.]. Template:Grading scheme. ([n. d.]). https://en.wikipedia.org/w/index.php?title=Template%3AGrading_scheme&oldid=800253353

[6] [n. d.]. TLX @ NASA Ames - NASA TLX Paper/Pencil Version. ([n. d.]). https://humansystems.arc.nasa.gov/groups/tlx/tlxpaperpencil.php

[7] Elena Agapie, Bonnie Chinh, Laura R. Pina, Diana Oviedo, Molly C. Welsh, Gary Hsieh, and Sean Munson. 2018. Crowdsourcing Exercise Plans Aligned with Expert Guidelines and Everyday Constraints. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 324, 13 pages. https://doi.org/10.1145/3173574.3173898

[8] Elena Agapie, Lucas Colusso, Sean A Munson, and Gary Hsieh. 2016. Plansourcing: Generating behavior change plans with friends and crowds. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 119–133.

[9] David Allen. 2015. *Getting things done: The art of stress-free productivity*. Penguin.

[10] John R Anderson, Michael Matessa, and Christian Lebiere. 1997. ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction* 12, 4 (1997), 439–462.

[11] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2015. Soylent: a word processor with a crowd inside. *Commun. ACM* 58, 8 (2015), 85–94.

[12] Justin Cheng, Jaime Teevan, and Michael S Bernstein. 2015. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1365–1374.

[13] Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. 2015. Break it down: A comparison of macro- and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 4061–4064.

[14] Lydia B Chilton, Greg Little, Darren Edge, Daniel S Weld, and James A Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1999–2008.

[15] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar. help: Designing a Workflow-Based Scheduling Agent with Humans in the Loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2382–2393.

[16] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2623–2634.

[17] Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication* 32, 4 (1981), 365–387.

[18] Andreas Glöckner and Tilmann Betsch. 2008. Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. (2008).

[19] Peter M Gollwitzer. 1999. Implementation intentions: Strong effects of simple plans. *American psychologist* 54, 7 (1999), 493.

[20] Peter M Gollwitzer and John A Bargh. 1996. *The psychology of action: Linking cognition and motivation to behavior.* Guilford Press.

[21] 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.

[22] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.

[23] Lynn Hasher and Rose T Zacks. 1979. Automatic and effortful processes in memory. *Journal of experimental psychology: General* 108, 3 (1979), 356.

[24] Shamsi T Iqbal and Brian P Bailey. 2006. Leveraging characteristics of task structure to predict the cost of interruption. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 741–750.

[25] Daniel Kahneman. 2011. *Thinking, fast and slow.* Macmillan.

[26] Joy Kim, Mira Dontcheva, Wilmot Li, Michael S Bernstein, and Daniela Steinsapir. 2015. Motif: Supporting novice creativity through expert patterns. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1211–1220.

[27] Aniket Kittur, Ed Chi, Bryan A Pendleton, Bongwon Suh, and Todd Mytkowicz. 2007. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web* 1, 2 (2007), 19.

[28] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 43–52.

[29] Nicolas Kokkalis, Thomas Köhn, Johannes Huebner, Moontae Lee, Florian Schulze, and Scott R Klemmer. 2013. Taskgenies: Automatically providing action plans helps people complete tasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 5 (2013), 27.

[30] Holger Krahn, Bernhard Rumpe, and Steven Völkel. 2008. MontiCore: Modular Development of Textual Domain Specific Languages. *TOOLS (46)* 11 (2008), 297–315.

[31] Michel Krieger, Emily Margarete Stark, and Scott R Klemmer. 2009. Coordinating tasks on the commons: designing for personal goals, expertise and serendipity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1485–1494.

[32] Anand Kulkarni, Matthew Can, and Björn Hartmann. 2012. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the acm 2012 conference on computer supported cooperative work*. ACM, 1003–1012.

[33] Walter S Lasecki, Jeffrey P Bigham, James F Allen, and George Ferguson. 2012. Real-Time Collaborative Planning with the Crowd.. In *AAAI*.

[34] Gary P Latham and Edwin A Locke. 1991. Self-regulation through goal setting. *Organizational behavior and human decision processes* 50, 2 (1991), 212–247.

[35] Howard Leventhal, Robert Singer, and Susan Jones. 1965. Effects of fear and specificity of recommendation upon attitudes and behavior. *Journal of Personality and Social Psychology* 2, 1 (1965), 20.

[36] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. 2010. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1739–1748.

[37] Michael Nebeling, Alexandra To, Anhong Guo, Adrian A de Freitas, Jaime Teevan, Steven P Dow, and Jeffrey P Bigham. 2016. WearWrite: Crowd-assisted writing from smartwatches. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3834–3846.

[38] Jeffrey M Rzeszotarski and Meredith Ringel Morris. 2014. Estimating the social costs of friendsourcing. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2735–2744.

[39] Niloufar Salehi, Jaime Teevan, Shamsi T Iqbal, and Ece Kamar. 2017. Communicating Context to the Crowd for Complex Writing Tasks.. In *CSCW*. 1890–1901.

[40] Thomas C Schelling. 1957. Bargaining, communication, and limited war. *Conflict Resolution* 1, 1 (1957), 19–36.

[41] Bran Selic. 2007. A systematic approach to domain-specific language design using UML. In *Object and Component-Oriented Real-Time Distributed Computing, 2007. ISORC'07. 10th IEEE International Symposium on*. IEEE, 2–9.

[42] Jaime Teevan, Shamsi T Iqbal, Carrie J Cai, Jeffrey P Bigham, Michael S Bernstein, and Elizabeth M Gerber. 2016. Productivity Decomposed: Getting Big Things Done with Little Microtasks. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 3500–3507.

[43] 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.

[44] Jaime Teevan, Daniel J Liebling, and Walter S Lasecki. 2014. Selfsourcing personal tasks. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2527–2532.

[45] Bill Tomlinson, Joel Ross, Paul Andre, Eric Baumer, Donald Patterson, Joseph Corneli, Martin Mahaux, Syavash Nobarany, Marco Lazzari, Birgit Penzenstadler, et al. 2012. Massively distributed authorship of academic papers. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, 11–20.

[46] Arie Van Deursen and Paul Klint. 2002. Domain-specific language design requires feature descriptions. *CIT. Journal of computing and information technology* 10, 1 (2002), 1–17.

[47] Ryan Vlastelica. 2017. Sex, pop culture and the other most popular Wikipedia topics by language. (Apr 2017). https://www.marketwatch.com/story/sex-pop-culture-and-the-other-most-popular-wikipedia-topics-by-language-2017-04-27

[48] Alexandre Yahi, Antoine Chassang, Louis Raynaud, Hugo Duthil, and Duen Horng Polo Chau. 2015. Aurigo: an interactive tour planner for personalized itineraries. In *Proceedings of the 20th international conference on intelligent user interfaces*. ACM, 275–285.

[49] Jeffrey M Zacks and Barbara Tversky. 2001. Event structure in perception and conception. *Psychological bulletin* 127, 1 (2001), 3.

[50] Jeffrey M Zacks, Barbara Tversky, and Gowri Iyer. 2001. Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General* 130, 1 (2001), 29.

[51] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. 2012. Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 217–226.

Received April 2018; revised July 2018; accepted September 2018