Published in final edited form as:
J Biomed Inform. 2020 October ; 110: 103571. doi:10.1016/j.jbi.2020.103571.

User-Centered Design of a Web-Based Crowdsourcing-Integrated Semantic Text Annotation Tool for Building a Mental Health Knowledge Base

Xing He[1], **Hansi Zhang**[1], **Jiang Bian**[1,*]

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA

Abstract

Background: One in five U.S. adults lives with some kind of mental health condition and 4.6% of all U.S. adults have a serious mental illness. The Internet has become the first place for these people to seek online mental health information for help. However, online mental health information is not well-organized and often of low quality. There have been efforts in building evidence-based mental health knowledgebases curated with information manually extracted from the high-quality scientific literature. Manual extraction is inefficient. Crowdsourcing can potentially be a low-cost mechanism to collect labeled data from non-expert laypeople. However, there is not an existing annotation tool integrated with popular crowdsourcing platforms to perform the information extraction tasks. In our previous work, we prototyped a Semantic Text Annotation Tool (STAT) to address this gap.

Objective: We aimed to refine the STAT prototype (1) to improve its usability and (2) to enhance the crowdsourcing workflow efficiency to facilitate the construction of evidence-based mental health knowledgebase, following a user-centered design (UCD) approach.

Methods: Following UCD principles, we conducted four design iterations to improve the initial STAT prototype. In the first two iterations, usability testing focus groups were conducted internally with 8 participants recruited from a convenient sample, and the usability was evaluated with a modified System Usability Scale (SUS). In the following two iterations, usability testing was conducted externally using the Amazon Mechanical Turk (MTurk) platform. In each iteration,

[*]**Corresponding author**: Jiang Bian, bianjiang@ufl.edu. Affiliation: Health Outcomes & Biomedical Informatics, University of Florida Address: 2197 Mowry Road, Room 122 PO Box 100177 Gainesville, FL 32610-0177 Phone Number: (501)773-9074.

CONTRIBUTORS

JB and XH designed the initial concepts and framework for the proposed work; HZ and XH carried out the annotation work, and evaluation of the crowdsourcing results. XH designed and developed the protocol, conduct the user-centered design studies, and administrated the crowdsourcing experiments. XH and JB wrote the initial draft of the manuscript. All authors provided critical feedback and edited the manuscript.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflicts of Interest

None declared.

Declaration of interests

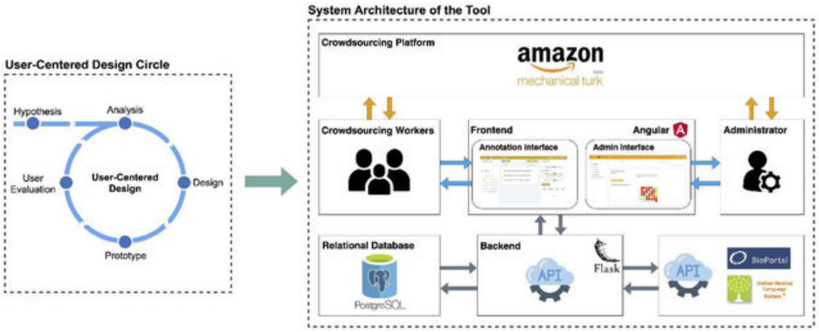
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

we summarized the usability testing results through thematic analysis, identified usability issues, and conducted a heuristic evaluation to map identified usability issues to Jakob Nielsen's usability heuristics. We collected suggested improvements in the usability testing sessions and enhanced STAT accordingly in the next UCD iteration. After four UCD iterations, we conducted a case study of the system on MTurk using mental health related scientific literature. We compared the performance of crowdsourcing workers with two expert annotators from two aspects: efficiency and quality.

Results: The SUS score increased from 70.3 ± 12.5 to 81.1 ± 9.8 after the two internal UCD iterations as we improved STAT's functionality based on the suggested improvements. We then evaluated STAT externally through MTurk in the following two iterations. The SUS score decreased to 55.7 ± 20.1 in the third iteration, probably because of the complexity of the tasks. After further simplification of STAT and the annotation tasks with an improved annotation guideline, the SUS score increased to 73.8 ± 13.8 in the fourth iteration of UCD. In the evaluation case study, on average, the workers spent 125.5 ± 69.2 seconds on the onboarding tutorial and the crowdsourcing workers spent significantly less time on the annotation tasks compared to the two experts. In terms of annotation quality, the workers' annotation results achieved average F1-scores ranged from 0.62 to 0.84 for the different sentences.

Conclusions: We successfully developed a web-based semantic text annotation tool, STAT, to facilitate the curation of semantic web knowledgebases through four UCD iterations. The lessons learned from the UCD process could serve as a guide to further enhance STAT and the development and design of other crowdsourcing-based semantic text annotation tasks. Our study also showed that a well-organized, informative annotation guideline is as important as the annotation tool itself. Further, we learned that a crowdsourcing task should consist of multiple simple microtasks rather than a complicated task.

Graphical abstract



Keywords

mental health; semantic annotation; user-centered design; crowdsourcing; semantic web knowledgebase

1. Introduction

In 2018, an estimated 47.6 million U.S. adults had any mental illness and an estimated 11.4 million adults had serious mental illness, corresponding to 4.6% of all U.S. adults. And the rate of youth experiencing mental health issues also continued to rise [1]. However, an astonishing number of Americans lacks access to mental health care [2]. Two out of 5 U.S. adults with any mental illness and more than half of those with serious mental illness cannot afford the cost of care; and similarly, more than 60% of youth with major depression did not receive any mental health treatment [2]. One well-documented reason for this situation is the severe mental health workforce shortage [2]. For example, to meet the need for mental health care, the providers in the state with the lowest-ranked mental health workforce availability must treat six times as many people as the providers in the highest-ranked state [2].

Besides, because many people are worried about being stigmatized if they admit that they have mental health issues, they often choose to keep their mental health issues private and are unwilling to ask for help [3–5]. With almost universal access to the Internet and the rising prevalence of smartphone use, the ways people learn and manage their mental health issues are changing. People often use the Internet to seek health information in general including mental health information [6,7]. However, the current online information related to mental health is poorly organized, without evidence to support it, and of poor quality [8–10]. Much of this online information consists of personal opinion, salesmanship, testimonials, and hypothesis-driven claims that are not evidence-based, mixed with high-quality information curated manually by experts such as those published on funding agencies (e.g., National Institute of Mental Health [NIMH]) or professional societies (e.g., Mental Health America).

A semantic web knowledgebase (KB) (also known as knowledge graph [KG]) using a formal knowledge representation (e.g., ontology) built with associated semantic web technologies can better organize and deliver quality mental health information to the public [11,12]. The World Wide Web Consortium (W3C) specifies the standards and protocols that define the semantic web, where Resource Description Framework (RDF) is the basis for semantic web data [13]. RDF can be used to structure knowledge statements in the format of semantic triples: subject-predicate-object [14]. For example, a mental health knowledge statement, "*antidepressant can treat depression*", can be encoded in a semantic triple as "*antidepressant (subject)-can treat (predicate)-depression (object)*". In a semantic web KB, entities (e.g., subjects, objects, and predicates) are standardized using ontologies that define the classes, relationships, concepts, and inference rules for the KB.

Existing KB or KG on mental health is sparse [15], although a number of well-known semantic web KBs and KGs exist such as Wikidata [16] and Google's Knowledge Graph [17] in the general domain. In the biomedical domain, there are approaches that extract knowledge statements from scientific literature such as those used in SemMed [18]. The SemMed platform uses SemRep [19]-a Unified Medical Language System (UMLS)-based natural language processing (NLP) tool-to extract triple statements from MEDLINE abstracts. Nevertheless, the performance of these automated methods is often suboptimal. To mitigate the performance issues, on the other hand, researchers have tried to manually build these semantic web KBs and KGs based on high quality, evidence-based resources, such as publications from high impact journals [20]. However, manual annotation and extraction of semantic triples in the form of subject-predicate-object expressions from publication abstracts are time-consuming, labor-intensive, and difficult.

Previously, Lossio-Ventura et al. demonstrated that crowdsourcing could potentially be a low-cost mechanism for collecting labeled data from non-expert laypeople. Even though individual layperson might not offer reliable answers, the collective wisdom of the crowd is comparable to expert opinions [21]. However, even for experienced annotators, the annotation tasks are laborious and difficult to perform without an easy-to-use annotation tool. In our previous work [22], we explored existing semantic annotation tools, but no one met our use case of constructing a mental health related semantic web KB through crowdsourcing. First, many of the existing tools are outdated, not well-maintained, and not web-based, thus are not suitable for crowdsourcing use as users need to download an desktop application and making coordination of the annotation tasks difficult. Second, most of these tools do not provide any annotation support such as suggesting candidate semantic classes of an entity. Further, these tools often do not provide a mechanism for monitoring the annotation quality (e.g., reporting inter-rater agreements) making it inconvenient for crowdsourcing tasks. Thus, we prototyped a web-based Semantic Text Annotation Tool (STAT) with an intended goal of being integrated with crowdsourcing platform [22].

A tailored user interface (UI) of a crowdsourcing system designed to reduce workers' cognitive load will help them focus on the task and perform better [23]. And using a user-centered design (UCD) process to develop such a system with a goal of minimizing users' cognitive load has been proven to free up mental resources and permit users to perform well on the crowdsourcing tasks [24]. UCD is an iterative design approach, in which targeted end-users influence how a design takes shape [25], which can potentially improve the chances of successful implementation of the technology in practice. Crowdsourcing systems and tools developed following a UCD process have reported improved user acceptance, user-friendliness, and ultimately better task performance [26,27].

In this study, we aimed to improve the user acceptance and user-friendliness of our STAT prototype iteratively following UCD principles and refine the annotation workflow to adapt to a crowdsourcing environment tailored for curating a high-quality mental health knowledge base.

2. Methods

2.1. Development of STAT following a User-Centered Design Process

The primary purpose of STAT is to assist laypeople in extracting semantic triples from scientific literature in a crowdsourcing environment. User experience (UX) is one of the most critical factors to be considered as crowdsourcing workers have a relative short attention span on the tasks. Based on the initial prototype [22], we followed UCD principles in the iterative development and refinement of STAT. As shown in Figure 1, the design and development of STAT consist of five steps: 1) summarizing existing research and semantic text annotation tools and making an initial hypothesis; 2) analyzing the needs of the intended end-users; 3) designing the prototype with the required functionalities and user interface (UI); 4) developing the working prototype; and 5) conducting usability testing and collecting user feedback. The last four steps (i.e., analysis, design, prototype, and evaluation) were conducted iteratively.

2.1.1. Hypothesis Making—In our previous work on building the STAT prototype [22], we hypothesized that a web-based annotation tool with real-time annotation recommendations and post-annotation quality analysis support for analyzing, monitoring, and managing the crowdsourcing annotation quality and results could help both experts and laypeople extract semantic triples from scientific literature effectively and accurately, and thus accelerate and facilitate the building of an evidence-based mental health KB.

2.1.2. Analysis of Needs and Requirements—In our previous prototyping work [22], we collected user needs and requirements for STAT by interviewing annotators who have extensive experience in semantic triple extraction tasks. We summarized their annotation workflows and designed the STAT prototype with a set of basic functionalities. In the subsequent UCD iterations, for each iteration, we analyzed the feedback collected from either usability testing focus groups or online surveys from the previous iteration and updated user needs and requirements accordingly.

2.1.3. Design—In the first UCD iteration, we reviewed existing scientific literature and online resources of semantic text annotation tools and found that web-based annotation tools are widely used and proven useful when annotating collaboratively. We also identified and examined the functionalities of some popular semantic text annotation tools (e.g., BRAT [28], GATE [29], WebAnno 3 [30], etc.). We then drafted the initial features desired by the intended end-users and developed the functional requirements based on user needs and requirements. In the subsequent design iterations, based on the user feedback from the usability testing focus group or online survey, we redesigned or refined existing features as well as added many new features. One fundamental design principle that we consistently followed throughout our UCD iterations is Occam’s razor principle: choosing the simplest solution whenever possible [31], where the UI is kept clean and simple yet provides all the necessary functionalities.

2.1.4. System Architecture—As depicted in Figure 2, STAT consists of three components: 1) a relational database PostgreSQL [32] to store literature text data and annotation data as well as additional management information (e.g., user settings, project information, etc.); 2) a Flask-based [33] Python backend with a Representational State Transfer (REST) [34] architecture to provide application programming interfaces (APIs) to the frontend; and 3) a web-based frontend UI built with the popular Angular [35] web application framework. The frontend UI consists of an annotation interface for the crowdsourcing workers and an administrator interface for managing annotation tasks. Workers recruited through crowdsourcing platforms (e.g., Amazon Mechanical Turk [36]) will carry out the annotation tasks via the annotation interface. The project owner or administrator will utilize the administrator interface to create new projects and then analyze, monitor, and manage the crowdsourcing annotation tasks.

2.1.5 User Acceptance and Usability Assessments—We engaged both internal (local) and external users to assess the user acceptance and usability of STAT.

2.1.5.1. Internal Usability Testing Focus Groups: During the initial stage of STAT, we focused on developing and completing the functionalities. A guideline on how to extract the correct information from the scientific literature was not developed initially.

In each initial stage of a UCD iteration, we examined the usability of STAT with a focus group of 8 participants recruited from a convenience sample (i.e., college students at the University of Florida). Each focus group lasted one hour with five sections: 1) we first provided an introduction of the study and the basic functionalities and interface of STAT (~10 minutes); 2) the participants were then asked to explore the UI and complete a list of pre-designed tasks (~15 minutes); 3) we assessed the usability using a modified System Usability Scale (SUS) [37] (~5 minutes); 4) the participants were asked to answer four open-ended questions to stimulate thinking (~10 minutes); and 5) the group carried out open discussions for us to gather user experience and feedback for improvements (~20 minutes).

2.1.5.2. External Usability Testing Over Amazon Mechanical Turk: After several UCD iterations, the internal usability testing result revealed a high acceptance and users expressed satisfaction of the STAT functionalities and UI. Thus, we conducted external usability testing using the popular crowdsourcing platform-Amazon Mechanical Turk (MTurk). We also designed and developed an annotation guideline to facilitate the annotation tasks.

In each of the following UCD iterations, we recruited 18 to 20 workers on MTurk. Participants were asked to finish three tasks: 1) using STAT to complete an assigned annotation task following an annotation guideline; 2) answering usability testing questions (i.e., the modified SUS); and 3) answering four open-ended questions to give feedback of using STAT. We also analyzed the annotation results from each participant to help identify possible usability issues.

2.2. Analysis of the Data Collected from the Usability Assessment Sessions

We analyzed the data gathered from the usability testing sessions both quantitatively and qualitatively.

2.2.1. Quantitative Analysis—During both internal and external usability testing, we used a modified SUS (see Appendix A) to evaluate the usability of STAT quantitatively. We modified the SUS because the original SUS questions were created to evaluate the usability of systems, such as "*I think that I would like to use this system frequently*." As STAT is a web application, we simply replaced the word "*system*" with "*website*," e.g., "*I think that I would like to use this website frequently*," to clarify the target of the evaluation. SUS is technology-independent and has been widely applied to evaluate the usability of hardware, general software, websites, and mobile apps. The 10-item SUS questionnaire is based on a 5-point Likert scale and scales to a maximum score of 100 on the users' impression of the usability of a product. A SUS score between 0 and 50 indicates that the usability is not acceptable, a score ranges from 50 to 70 means marginally acceptable, and a score higher than 70 deems acceptable [38].

In the external usability testing through MTurk, we collected MTurk workers' annotation results to assess annotation quality. Following best practice in designing crowdsourcing tasks, we divided the big and complex annotation task of extracting semantic triples into three microtasks. These microtasks are:

1. *Annotation of entities*, where the workers need to identify individual entities (i.e., subjects, objects, and relations) from the given sentences. For example, given a sentence "*depression may hinder cancer recovery*", the workers need to extract the subject "*depression*", the object "*cancer recovery*" and the relation "*may hinder*" from the sentence.

2. *Normalization of entities*, where the workers are expected to map the identified entities to the classes in existing ontologies or controlled vocabularies (e.g., UMLS). For example, the entity "*depression*" should be normalized to "*mental depression*" (CUI: C0011570) according to UMLS.

3. *Composition of semantic triples*, where the workers will compose the normalized entities into semantic triple statements. For example, the extracted and normalized entities in the above examples should be composed into a triple statement of "*mental depression [CUI: C0011570] (subject)-may hinder (predicate)-cancer recovery (object)"*.

We analyzed the annotation results and identified the tasks that were difficult for laypeople to complete. We calculated the rate of corrected completed assignments (i.e., number of participants who completed a microtask correctly divided by the total number of participants assigned to the task) of each microtask, suggesting a lower rate indicated that the task was more difficult for laypeople to complete. A lower correctness rate can potentially reveal difficult microtasks that might impair the usability of the system.

2.2.2. Qualitative Analysis

In each internal usability testing focus group, we posted four open-ended questions before each usability assessment session: 1) "*What is the major difficulty for you to use STAT?*"; 2) "*Do you have any ideas or advice for the improvement of STAT?*"; 3) "*List the most negative aspect(s)*"; and 4) "*List the most positive aspect(s)*." During the discussion session, participants were encouraged to think aloud and discuss these questions and any other related issues, such as their perceptions and attitudes about using a tool like STAT or suggesting missing or additional functionalities. With participants' consent, the focus group sessions were recorded and then transcribed.

In each of the external usability testing sessions, we included the same sets of open-ended questions on MTurk's task description page as thought-provoking questions. Participants were asked to answer these questions in a survey after they completed the annotation tasks.

We then identified and categorized the usability issues by themes and heuristics and analyzed the usability testing results qualitatively. We collected a set of usability issues from the answers to the open-ended questions and the transcripts of the recorded focus groups. All usability issues were encoded using themes derived from a thematic analysis and mapped to Nielsen's usability heuristics [39]. Through the thematic analysis, the usability issues reported in each design iteration were encoded by themes. We followed a well-established process for the thematic analysis commonly used in human-computer interaction projects consisting of five steps: 1) familiarizing with the data; 2) assigning initial annotation codes;

3) sorting and merging the coded data into broader themes; 4) reviewing and refining the themes identified before; and 5) naming and describing each of the themes. Similar usability issues were grouped as a unique issue type. We also extracted suggested improvements from the open-ended questions and the transcribed recordings and ranked the importance of the suggested improvements. Highly ranked suggestions were considered in the next UCD iteration to refine the design of STAT.

2.3. Evaluation of Annotation Tasks using STAT through a Case Study

After four rounds of UCD iterations (i.e., two internal rounds with local convenient sample and two external rounds with MTurk users), we conducted a case study on MTurk using mental health related scientific literature and evaluated the annotation results of the tasks from two aspects: annotation efficiency and quality. We identified five sentences related to "*mental health*" with different complexity levels, where the number of semantic triples in each sentence ranged from 1 to 8 (i.e., a sentence is more complex if it contains more semantic triples to be extracted). We recruited two experienced annotators (i.e., annotators with biomedical informatics training and had experience in extracting semantic triples from published literatures) to annotate the selected sentences manually following their existing workflow (i.e., the extracted triples were recorded in a Excel spreadsheet) without using STAT and recorded their annotation processes using a screen recorder. The consensus of their annotation results serves as the gold standard dataset. And we analyzed the recorded video to calculate the average time spent on each sentence.

We then released the annotation task on MTurk and recruited 5 workers. We tracked the amount of time they spent on various components of STAT such as the time spent on the onboarding tour and on the annotation of each sentence for each worker.

Furthermore, we evaluated the annotation quality on entity level (i.e., we only evaluated the quality of annotations for entities and relations and did not evaluate the quality of extracting the complete semantic triples). Following the evaluation metrics used in the SemEval-2013 challenge for Task 9.1 [40], for each sentence, we reported the average precision, recall and F1-score comparing each worker's annotation results with the gold standard created by the two expert annotators, where an annotation is considered as a true positive (TP) only if there is some overlap between a worker's annotation and a gold standard annotation of the same type, a false positive (FP) where (1) there is overlap but the annotated type is incorrect (e.g., a worker marked an entity as a relation), or (2) the worker's annotation does not appear in the gold standard, and a false negative (FN) if the worker missed an annotation in the gold standard. The formulas for precision, recall, and F1-score can then be calculated as follows:

$$Precision = \frac{\#\,of\,TP}{\#\ of\,TP + \#\ of\ FP}$$

$$Recall = \frac{\#\,of\,TP}{\#\ of\,TP + \#\ of\ FN}$$

$$F1\text{-}score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

3. Results

Table 1 summarizes the results of the four UCD iterations. And Table 2 shows the demographics of the participants for each UCD iteration. The first and the second iteration were mainly about UI and functionalities and conducted internally within the University of Florida. And the following three iterations focused on refining the annotation workflow and annotation guideline. And these iterations were conducted externally through MTurk. Both internal and external iterations were stopped when the SUS score reached acceptable, and there were no major usability issues identified.

3.1. Development of STAT following a User-Centered Design Process

We refined STAT through 4 rounds of UCD iterations. Figure 3 shows the main user interface of STAT after the two internal iterations. At that stage, the UI supports three operations: 1) annotating text, 2) normalizing selected text leveraging the terminology and ontology services from the National Center for Biomedical Ontology (NCBO) BioPortal and UMLS, and 3) composing semantic triples. As depicted in Figure 4, the normalization popup window supports various normalization operations on the selected text (e.g., querying BioPortal and UMLS for recommending candidate concept classes based on the selected phrases). The annotation task at the time was to extract semantic triples from the entire abstract (i.e., multiple sentences) of the scientific literature, as we thought that the annotators were able to not only extract triples within a single sentence but also relations between subjects and objects across different sentences (if exist). We also provided a very detailed usage guideline with a tutorial video and descriptions to instruct the users how to use STAT to finish the annotation tasks.

This UI worked well during our internal usability testing focus groups. Nevertheless, after a preliminary test (i.e., a quick testing to make sure the tool works fine on the crowdsourcing environment) on MTurk before the formal external usability testing, we collected valuable feedback and recognized that the overall task was too complex for crowdsourcing workers. We simplified the annotation workflow by splitting the abstract to sentences and retained only sentences from the conclusion section of the articles (i.e., as we aim to extract factual statements from that particular study, rather than statements cited by that study) to reduce the workload of each annotation task.

After the third UCD iteration, STAT was further simplified according to the usability testing results. We removed the normalization task and associated component because the normalization results were suboptimal. It is understandable as our previous experience also indicated that it was difficult even for an experienced expert to normalize a text phrase to an existing ontology class accurately. For example, in the sentence "*efforts to prevent cancer and promote health must attend to mental health disparities to meet the needs of young adults*", the entity "*mental health disparities*" does not match to any concept classes in the existing ontologies; and the recommended ontology classes are either "*Health Disparities*" or "*Mental Health*" and from multiple ontologies (e.g., the class "*Health Disparities*" can be found in the Psychology Ontology [http://ontology.apa.org/apaonto/termsonlyOUT%20(5).owl#Health_Disparities] and MedlinePlus Health Topics [http://purl.bioontology.org/ontology/MEDLINEPLUS/C1171307]). Neither of the two recommended ontology classes is appropriate to represent the entity "*mental health disparities*", making it difficult even for the experts to decide.

Further, the learning curve of using STAT is much higher with requiring the normalization operations. Considering the fact that crowdsourcing workers only spend a short amount of time on the tasks, we further simplified the detailed tool usage guide to a 3-step onboarding tour with a 40 seconds video. The further simplified UI of STAT is shown in Figure 5.

Figure 6 shows the simplified annotation workflow after the fourth UCD iteration. After an annotator accepts the task on the MTurk platform, she will be redirected to the STAT site. The onboarding tour will pop up automatically. As shown in Figure 7.A the first step of the onboarding tour introducing the purpose of this tour. Figure 7.B shows the annotation guideline that consists of textural descriptions of the tasks (i.e., telling workers what to extract from the sentences) and a demonstration video tutorial (i.e., how to use the STAT tool).

After the onboarding tour, the annotator will be asked to annotate 5 to 10 sentences. For each sentence, the annotator needs to read the sentence carefully, and then identify and extract all entities (e.g., noun phrases) or relations (e.g., verb phrases) from the sentence. After the annotator selected a candidate text, the annotator can either use the menu on the top of the sentence panel (Figure 8.A) or use the context menu by right-clicking on the selected text (Figure 8.B) to mark the selected text as either an entity of relation.

After the entities and relations are extracted, the annotator needs to drag these extracted entities and relations into the corresponding areas of the semantic triple panel as shown in Figure 8.C to construct semantic triples in the form of "*subject - predicate - object*". When the annotator completes the annotation task for all sentences, the annotator can click the "Finish Annotation" button to generate a reward code for them to redeem the incentives on the MTurk platform.

3.2. Analysis of the Data Collected from the Usability Assessment Sessions

In the first UCD iteration, we conducted an internal usability testing with 8 participants recruited from a convenient sample (i.e., staff or graduate students from the University of Florida). The average SUS score of the initial STAT prototype was acceptable (70.3 ± 12.5). Overall, seven distinct usability issues were identified from the focus group. And we grouped the usability issues into three themes as shown in Table 3. Most of these usability issues were related to the lack of certain functionalities. We also mapped these distinct usability issues to the ten usability heuristics described in Nielsen's book [39]. The most common usability heuristic is the "*flexibility and efficiency of use*"-urging the system to be flexible catering to both inexperienced and experienced users.

We then composed a list of usability improvements suggested by the participants, which were aligned with the identified usability issues. Broadly, these proposed improvements included: 1) adding new functions; 2) improving existing functions; 3) improving UI and UX; and 4) improving documentation of the system and the annotation guideline. Table 4 lists selected suggested improvements and the corresponding actions we had taken.

In the second UCD iteration, we recruited another 8 participants from a convenient sample, and four of them have attended the usability testing focus group in the first iteration. The average SUS score was increased to 81.1 ± 9.8. The total number of distinct usability issues (Table 3) identified was 7. The category of the suggested usability improvements (Table 4) remained the same. There were no major improvements suggested.

Since the SUS score in the second UCD iteration was acceptable and no major improvement was needed, we then conducted a pilot testing through MTurk to make sure STAT would work fine in a crowdsourcing environment before the formal usability testing and the third UCD iteration. We released a testing task (i.e., extracting all semantic triples from a complete abstract of a scientific paper) on MTurk without any restriction and collected 18 responses. However, after reviewing the usability testing survey results and the annotation data, only one worker finished the annotation work following the instruction, and filled the usability testing survey carefully. This worker expressed that the annotation work was too difficult, and it took a long time to complete all the tasks. The worker preferred to annotate a single sentence at a time rather than a full abstract. Considering the pilot testing results, we changed the annotation task from abstract based to sentence based.

After the pilot testing, we conducted the third UCD iteration externally. We released the usability testing tasks on MTurk and required all the workers to be MTurk master workers. Master workers are those who have consistently demonstrated a high degree of success across a large number of tasks on MTurk. The Amazon MTurk platform monitors and analyzes workers’ performance and certify these top performance workers as MTurk master worker. These master workers may not have experience with the annotation tasks; however, using master workers will eliminate either bots or cheaters who merely clicking through a task for the reward. At the end, we received responses from 20 master workers. After our further examination of the annotation results and the quality of their survey responses, only 14 of them met our requirements. The average SUS score decreased to 55.7 ± 20.1 in this iteration. Analysis of the usability testing results revealed that the annotation guideline was a critical reason for the significant decrease of the SUS score. In the internal usability testing, the participants were only required to test the functionalities and UI of STAT. However, in the external usability testing via MTurk, the workers were required to complete the actual annotation tasks, where the deficiency in the annotation guideline significantly affected the UX and annotation quality. As shown in Table 3, the majority of the usability issues in the third iteration were related to the annotation guideline.

We further analyzed the annotation results. The annotation task was divided into three microtasks: 1) annotation of terms; 2) normalization of the annotated terms; and 3) composition of triples. 10 of 14 workers could annotate the terms from the text correctly, and 9 of 14 workers could compose semantic triples correctly. However, only 4 of 14 workers normalized some of the annotated text correctly. The other ten workers did not conduct normalization or fail to normalize the annotated text correctly. Based on the analysis result and the usability issues, we think it is more appropriate to let the experts conduct the normalization operation in future studies. Combining the improvement proposed by workers and the analysis result, the improvements were listed in Table 4.

In the fourth UCD iteration, we focused on refining the annotation task workflow and the annotation guideline. We collected 20 usability testing responses through MTurk and 17 of them were valid. The SUS score increased from marginally acceptable to acceptable (73.8 ± 13.8). As shown in Table 3, the majority of these usability issues in the fourth iteration was about the annotation guideline. Even though we modified the annotation guideline multiple times, MTurk workers still felt that some terms were hard to understand in the guideline. Compared with previous UCD iterations, there were no new features required. Table 4 lists the suggested usability improvements and corresponding improvement actions of the fourth iteration.

3.3. Evaluation of Annotation Tasks using STAT through a Case Study

In the crowdsourcing evaluation case study, we recruited 5 workers from the MTurk platform and workers were asked to annotate 5 sentences (see Appendix B) of different complexity. In order to create a gold-standard dataset to assess workers' annotation quality, two experienced annotators (HZ and XH) manually annotated the same 5 sentences in the same order as the crowdsourcing workers. The inter-annotator agreement between the two experienced annotators was 0.83, where conflicts between the two annotators were resolved with the entire study team. The average time the workers spent on the onboarding tour was 125.5 ± 69.2 seconds. Table 5 shows crowdsourcing workers' performance of the annotation tasks by sentence. In terms of time spent on the annotation tasks, on average, the workers spent less time on each sentence comparing to experts. And it is clear that both crowdsourcing workers and experts spent more time on more complex sentences with more triples.

In terms of annotation quality, the average F1 scores of the sentences ranged from 0.62 to 0.84. We also noticed that the lowest average F1 score (0.62) was for the 5th sentence, which only contains 3 entities or relations. Comparing to the 1st sentence that contains 10 entities or relations, a lower F1 score of the 5th sentence indicated that the difficulty of annotating a sentence could not be simply measured by the number of entities, relations, and tipples it contains. Further, each individual crowdsourcing worker might produce suboptimal annotation results; nevertheless, when considering a majority rule, the performance of an aggregated result is comparable to expert annotations. Aggregated results based on majority vote achieved F1 scores ranged from 0.83 to 1.

4. Discussion

4.1. Principal Results

Semantic text annotation to curate semantic web KBs or KGs is a difficult and time-consuming task even for experienced experts. It could be time-saving and low-cost to utilize the power of crowdsourcing to facilitate the annotation task. In this study, we conducted four design and development iterations to improve a semantic text annotation tool, STAT, and achieved an acceptable SUS score (73.8 ± 13.8) in the last iteration. We followed user-centered design principles during these iterations, leading to a more user-friendly and easier to use annotation tool and a more reasonable annotation workflow. Our annotation case study demonstrated that the use of STAT could reduce the time spent on annotating the sentences compared with manual extraction without using a tool. This may be beneficial even for expert annotators. Combined with existing crowdsourcing platforms, we could use STAT to collect a large number of semantic text annotations from crowdsourcing workers within a short amount of time and at a low cost. We evaluated the quality of crowdsourced annotations using STAT and the aggregated results based on majority vote achieved an F1 score ranging from 0.83 to 1, outperforming individual worker. The evaluation results showed that the quality of the annotations on most of the sentences by the workers was acceptable.

By following the UCD principles to refine STAT, we had a better understanding of the iterative nature of design and development when incorporating user feedback from different sources. In the UCD iterations of refining STAT, we involved both experienced expert annotators and crowdsourcing workers in different stages of design and development. Leveraging both internal and external participants in the UCD process helped us achieve rapid prototyping in the early stage while ensuring the tool was acceptable and usable to the crowdsourcing workers in the production stage. When conducting the usability testing externally through Amazon MTurk, we managed to collect valuable feedback from the crowdsourcing workers and tailored STAT accordingly. Our success in designing and refining STAT through MTurk with remote participants demonstrated that we do not always need the participants to attend the UCD interviews or focus groups physically. A multi-approach mixed UCD process could be used to promote the usability and acceptance of future crowdsourcing tools and improve workers' performance on the crowdsourcing tasks.

We also gained significant insights during the design and development of STAT. First, we should always keep in mind that crowdsourcing tasks-in terms of both the crowdsourcing tool and the task workflow-should be as simple as possible. Most crowdsourcing workers will only perform the annotation task once with very short attention span, considering the small amount of incentives (i.e., $0.25 per task in our case) they get. Thus, the annotation tool should be easy to learn and use. There would be tradeoffs between simplicity and comprehensiveness when designing and developing the crowdsourcing annotation tool. Another lesson that we learned is we should consider the target end-users as early as possible. During the first two internal UCD iterations, we collected user feedback about the STAT prototype from convenient samples rather than the actual target users-crowdsourcing workers. Although with local users we could achieve rapid prototyping, we wasted development time on building some functionalities (e.g., normalization tasks) that were too complicated and not suitable for crowdsourcing users. Our original annotation workflow was too difficult for crowdsourcing workers. Thus, we divided the workflow into smaller microtasks (i.e., first annotation of entities and then construction of triples using annotated entities) and removed difficult steps (i.e., entity normalization). Nevertheless, the task of constructing semantic triples was still problematic leading to suboptimal annotation quality.

A more sensible approach may be mixing crowdsourcing workers and expert annotators into the workflow, where the crowdsourcing workers are only used to perform simple tasks (i.e., identification of entities and relations) on a larger scale while the experts perform more complex tasks (e.g., normalization and construction of triples). Such a hybrid approach would work as follows: 1) the administrators select publications and extract interested sentences from the publications, and then publish the annotation tasks on crowdsourcing platforms; 2) the crowdsourcing workers complete the entity annotation tasks on these extracted sentences; 3) the experts map the annotated entities to existing (or creating new if necessary) ontology classes; and 4) the experts constructed semantic triples using the normalized entities. During the ontology mapping operations, the expert annotators could utilize the normalization functionalities depicted in Figure 4. Moreover, in our evaluation experiments, we only reported the performance based on lenient matching (i.e., partial matching of the annotated entities to the gold-standard), as during the normalization process, it is sufficient to present the experts the consensus entities (extracted by the crowdsourcing workers) based on lenient matching results.

More lessons were learned when we tested STAT in a real crowdsourcing environment. The first is the need for mechanisms to control annotation quality. Not all crowdsourcing workers will treat each task carefully, and there should be a mechanism to filter these low-quality annotators. For example, we included an easy annotation task (e.g., extracting the entities from a simple sentence "*I like math, physics and geography"*) in every annotation task as a control task (i.e., similar to CAPTCHA)-a type of challenge-response test to determine whether or not the user is human or pays attention to the task. Another lesson worth mentioning is that we needed to create more assignments (i.e., 6.6 ± 1.1 assignments per task) than the number of responses we needed (i.e., 5 valid responses in our case) for each task because it is very common for crowdsourcing workers not being able to finish the task correctly.

After four UCD iterations, STAT is not yet perfect. Some workers still complained about the annotation guideline being too difficult to follow. Besides, how to utilize voting or other mechanisms to facilitate the generation of more accurate and better annotations from collected crowdsourcing data still needs further exploration.

4.2. Conclusions

User-centered design enabled us to create a user-friendly web-based crowdsourcing-integrated semantic text annotation tool, STAT, to facilitate and speed up the construction of a mental health KB, through an iterative development process. Our study explored the conversion process from an expert-based task to multiple laypeople-based microtasks through the UCD iterations. Our efforts and failures made in the UCD iterations indicate the importance to consider the needs from the target end-users in the design process as early as possible. Further, annotation guidelines for annotation tasks are very crucial in successfully carrying out the crowdsourcing tasks. More efforts are still needed to improve our annotation guideline. The lessons learned from the design and development of STAT could serve as a guide to further enhancement of STAT and the development of other crowdsourcing tools.

In future studies, we can potentially integrate machine learning or deep learning algorithms to reduce the complexity of the annotation task further. For example, we could leverage the advancements in natural language processing (NLP), especially named-entity recognition (NER) models, to highlight the potential concepts (i.e., entities and relations in our tasks) and the crowdsourcing workers will only need to determine whether the recognized concepts are correct or not. Our ultimate goal is to create an efficient pipeline that consumes high-quality scientific literature to curate and enrich a mental health KB with knowledge extracted from these publications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported in part by grants National Institute of Health (NIH) UL1TR001427, R21AG061431, R01CA246418 and the Cancer Informatics and eHealth Core program at the University of Florida Health Cancer Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This article did not receive sponsorship for publication costs.

Abbreviations

STAT	semantic text annotation tool
UCD	user-centered design
MTurk	Mechanical Turk
W3C	World Wide Web Consortium
RDF	Resource Description Framework
KB	knowledge base
KG	knowledge graph
UI	user interface
REST	representational state transfer
API	application programming interface
SUS	System Usability Scale
NER	named-entity recognition

References

1. Substance Abuse and Mental Health Services Administration. Key substance use and mental health indicators in the United States: Results from the 2018 National Survey on Drug Use and Health [Internet]. Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration; 2019 8 Report No.: HHS Publication No. PEP19–5068, NSDUH Series H-54. Available from: https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHNationalFindingsReport2018/NSDUHNationalFindingsReport2018.pdf

2. Hellebuyck Michele, Halpern Madeline, Nguyen Theresa, Fritze Danielle. The State of Mental Health in America 2019 [Internet]. Mental Health America, Inc.; Available from: https://mhanational.org/sites/default/files/2019%20MH%20in%20America%20Final_0.pdf
3. Barney LJ, Griffiths KM, Jorm AF, Christensen H. Stigma about depression and its impact on help-seeking intentions. Aust N Z J Psychiatry 2006 1;40(1):51–54. [PubMed: 16403038]
4. Eisenberg D, Downs MF, Golberstein E, Zivin K. Stigma and help seeking for mental health among college students. Med Care Res Rev MCRR 2009 10;66(5):522–541. [PubMed: 19454625]
5. Greene-Shortridge TM, Britt TW, Castro CA. The stigma of mental health problems in the military. Mil Med 2007 2;172(2):157–161. [PubMed: 17357770]
6. Horgan A, Sweeney J. Young students' use of the Internet for mental health information and support. J Psychiatr Ment Health Nurs 2010 3;17(2):117–123. [PubMed: 20465756]
7. Burns JM, Durkin LA, Nicholas J. Mental health of young people in the United States: what role can the internet play in reducing stigma and promoting help seeking? J Adolesc Health Off Publ Soc Adolesc Med 2009 7;45(1):95–97.
8. Nemoto K, Tachikawa H, Sodeyama N, Endo G, Hashimoto K, Mizukami K, Asada T. Quality of Internet information referring to mental health and mental disorders in Japan. Psychiatry Clin Neurosci 2007 6;61(3):243–248. [PubMed: 17472591]
9. Reavley NJ, Jorm AF. The quality of mental disorder information websites: a review. Patient Educ Couns 2011 11;85(2):e16–25. [PubMed: 21087837]
10. Kirby PL, Reynolds KA, Walker JR, Furer P, Pryor TAM. Evaluating the quality of perinatal anxiety information available online. Arch Womens Ment Health 2018;21(6):813–820. [PubMed: 29931445]
11. Hadzic M, Chen M, Dillon TS. Towards the Mental Health Ontology. 2008 IEEE Int Conf Bioinforma Biomed 2008 284–288. doi: 10.1109/BIBM.2008.59
12. Yamada DB, Yoshiura VT, Brandão Miyoshi NS, de Lima IB, Usumoto Shinoda GY, Lopes Rijo RPC, de Azevedo Marques JM, Cruz-Cunha MM, Alves D. Proposal of an ontology for Mental Health Management in Brazil. Procedia Comput Sci 2018 1 1;138:137–142. doi: 10.1016/j.procs.2018.10.020
13. RDF Working Group. RDF - Semantic Web Standards [Internet]. [cited 2020 Aug 13] Available from: https://www.w3.org/RDF/
14. Cyganiak R, Wood D, Lanthaler M. RDF 1.1 Concepts and Abstract Syntax [Internet]. [cited 2020 Aug 13] Available from: https://www.w3.org/TR/rdf11-concepts/
15. Huang Z, Yang J, van Harmelen F, Hu Q. Constructing Knowledge Graphs of Depression In: Siuly S, Huang Z, Aickelin U, Zhou R, Wang H, Zhang Y, Klimenko S, editors. Health Inf Sci [Internet] Cham: Springer International Publishing; 2017 [cited 2020 Apr 11] 149–161. doi: 10.1007/978-3-319-69182-4_16
16. Turki H, Shafee T, Hadj Taieb MA, Ben Aouicha M, Vrandečić D, Das D, Hamdi H. Wikidata: A large-scale collaborative ontological medical database. J Biomed Inform 2019 11;99:103292. [PubMed: 31557529]
17. Singhal Amit. Introducing the Knowledge Graph: things, not strings [Internet]. 2012 [cited 2020 Feb 13] Available from: https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html
18. Rindflesch TC, Kilicoglu H, Fiszman M, Rosemblat G, Shin D. Semantic MEDLINE: An advanced information management application for biomedicine. Inf Serv Use 2011 9 6;31(1–2):15–21. doi: 10.3233/ISU-2011-0627
19. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003 12;36(6):462–477. [PubMed: 14759819]
20. Zhang H, He X, Harrison T, Bian J. Aero: An Evidence-based Semantic Web Knowledge Base of Cancer Behavioral Risk Factors. SEPDA@ISWC 2019.
21. Lossio-Ventura JA, Hogan W, Modave F, Guo Y, He Z, Yang X, Zhang H, Bian J. OC-2-KB: integrating crowdsourcing into an obesity and cancer knowledge base curation system. BMC Med Inform Decis Mak 2018 23;18(Suppl 2):55. [PubMed: 30066655]

22. He X, Zhang H, Yang X, Guo Y, Bian J. STAT: A Web-based Semantic Text Annotation Tool to Assist Building Mental Health Knowledge Base. IEEE Int Conf Healthc Inform IEEE Int Conf Healthc Inform 2019 Jun;2019.

23. Rahmanian B, Davis JG. User Interface Design for Crowdsourcing Systems Proc 2014 Int Work Conf Adv Vis Interfaces [Internet] New York, NY, USA: Association for Computing Machinery; 2014 405–408. doi: 10.1145/2598153.2602248

24. Oviatt S Human-Centered Design Meets Cognitive Load Theory: Designing Interfaces That Help People Think Proc 14th ACM Int Conf Multimed [Internet] New York, NY, USA: Association for Computing Machinery; 2006 871–880. doi: 10.1145/1180639.1180831

25. Abras C, Maloney-krichmar D, Preece J. User-Centered Design Bainbridge W Encycl Hum-Comput Interact Thousand Oaks Sage Publ Publications; 2004.

26. Burghardt M, Spanner S. Allegro: User-Centered Design of a Tool for the Crowdsourced Transcription of Handwritten Music Scores Proc 2nd Int Conf Digit Access Textual Cult Herit [Internet] New York, NY, USA: Association for Computing Machinery; 2017 15–20. doi: 10.1145/3078081.3078101

27. Stromer-Galley J, Rossini PGC, Kenski K, Folkestad J, McKernan B, Martey RM, Clegg B, Osterlund C, Schooler L. User-Centered Design and Experimentation to Develop Effective Software for Evidence-Based Reasoning in the Intelligence Community: The TRACE Project. Comput Sci Engg 2018 11;20(6):35–42. doi: 10.1109/MCSE.2018.2873859

28. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a Web-based Tool for NLP-Assisted Text Annotation Proc Demonstr 13th Conf Eur Chapter Assoc Comput Linguist [Internet] Avignon, France: Association for Computational Linguistics; 2012 102–107. Available from: https://www.aclweb.org/anthology/E12-2021

29. Bontcheva K, Cunningham H, Roberts I, Roberts A, Tablan V, Aswani N, Gorrell G. GATE Teamware: a web-based, collaborative text annotation framework. Lang Resour Eval 2013 12;47(4):1007–1029. doi: 10.1007/s10579-013-9215-6

30. Eckart de Castilho R, Mújdricza-Maydt É, Yimam SM, Hartmann S, Gurevych I, Frank A, Biemann C. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures Proc Workshop Lang Technol Resour Tools Digit Humanit LT4DH [Internet] Osaka, Japan: The COLING 2016 Organizing Committee; 2016 76–84. Available from: https://www.aclweb.org/anthology/W16-4011

31. Knight W UX for developers: how to integrate user-centered design principles into your day-to-day development work. Apress; 2019 ISBN:978-1-4842-4227-8

32. The PostgreSQL Global Development Group. PostgreSQL: The world's most advanced open source database [Internet]. [cited 2020 Feb 7] Available from: https://www.postgresql.org/

33. Grinberg M Flask Web Development: Developing Web Applications with Python. O'Reilly Media, Inc.; 2018 ISBN:978-1-4919-9169-5

34. Richardson L, Ruby S. RESTful Web Services. O'Reilly Media, Inc.; 2008 ISBN:978-0-596-55460-6

35. Google LLC. AngularJS [Internet]. 2020 [cited 2020 Feb 7] Available from: https://angularjs.org/

36. Amazon Mechanical Turk, Inc. Amazon Mechanical Turk [Internet]. [cited 2020 Feb 7] Available from: https://www.mturk.com/

37. Brooke J SUS: a retrospective. J Usability Stud 2013 Feb 1;8(2):29–40.

38. Bangor A, Kortum PT, Miller JT. An Empirical Evaluation of the System Usability Scale. Int J Human-Computer Interact 2008 7 29;24(6):574–594. doi: 10.1080/10447310802205776

39. Nielsen J Usability Engineering. Elsevier; 1994 ISBN:978-0-08-052029-2

40. Segura-Bedmar I, Martínez P, Herrero-Zazo M. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). Second Jt Conf Lex Comput Semant SEM Vol 2 Proc Seventh Int Workshop Semantic Eval SemEval 2013 2013 341–350.

Highlights

- User-centered deisgn of a web-based semantic text annotation tool
- Annotation tool to facilitate the curation of semantic web knowledge bases
- Well-organized, informative annotation guideline is as crucial as the tool
- Crowdsourcing task should consist of multiple simple microtasks

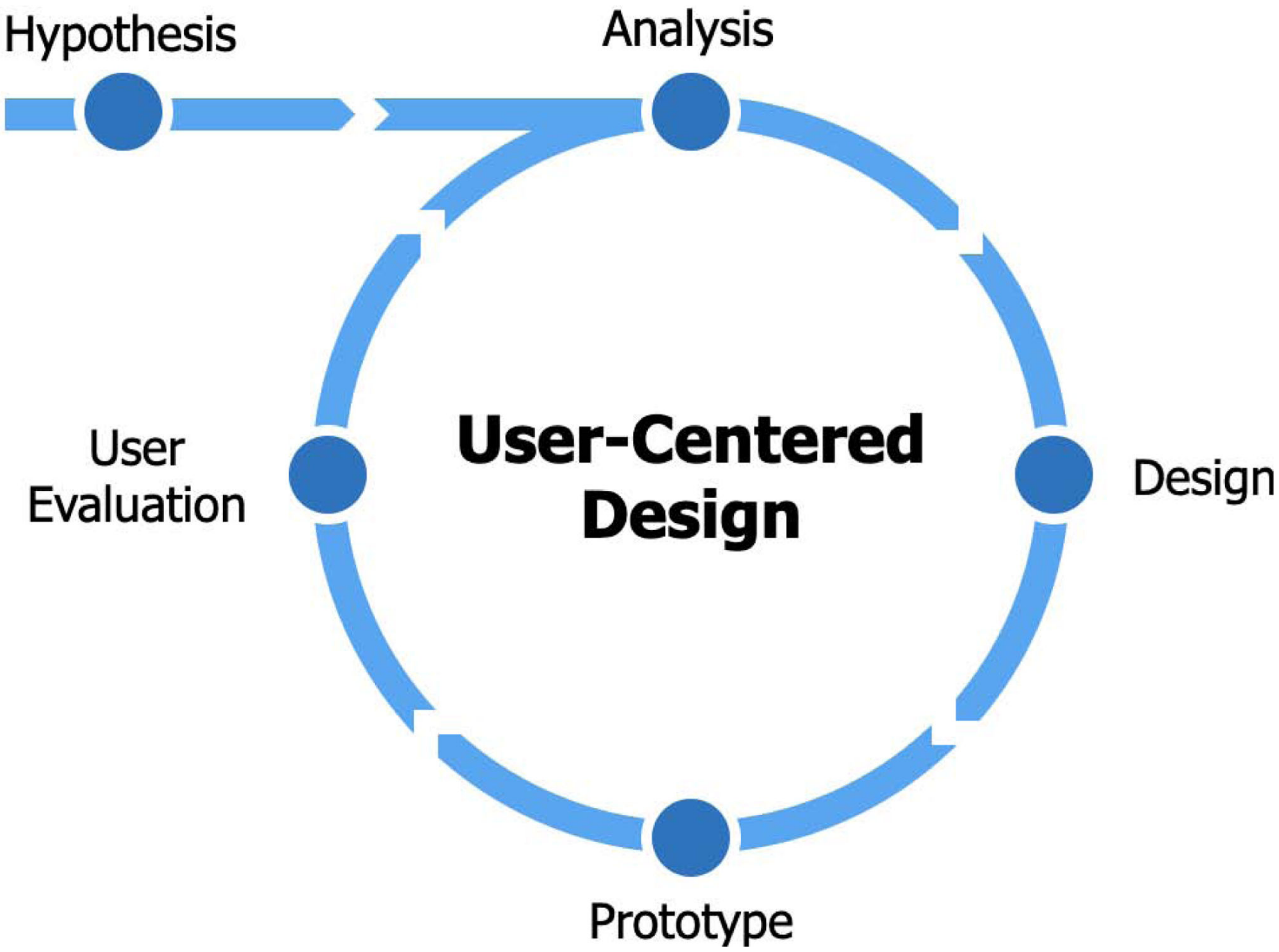


Figure 1.

An iterative user-centered design process for the development of STAT.

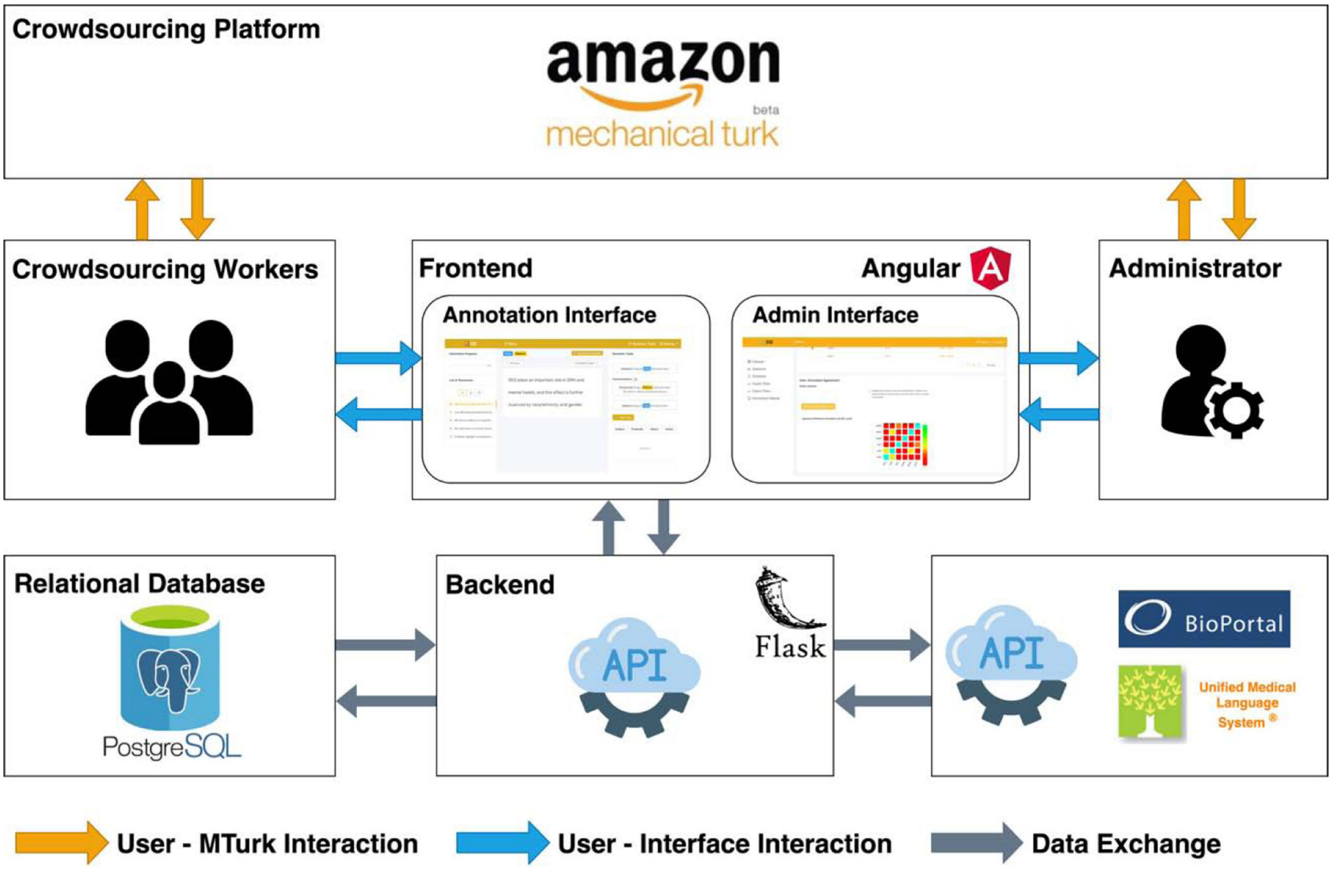
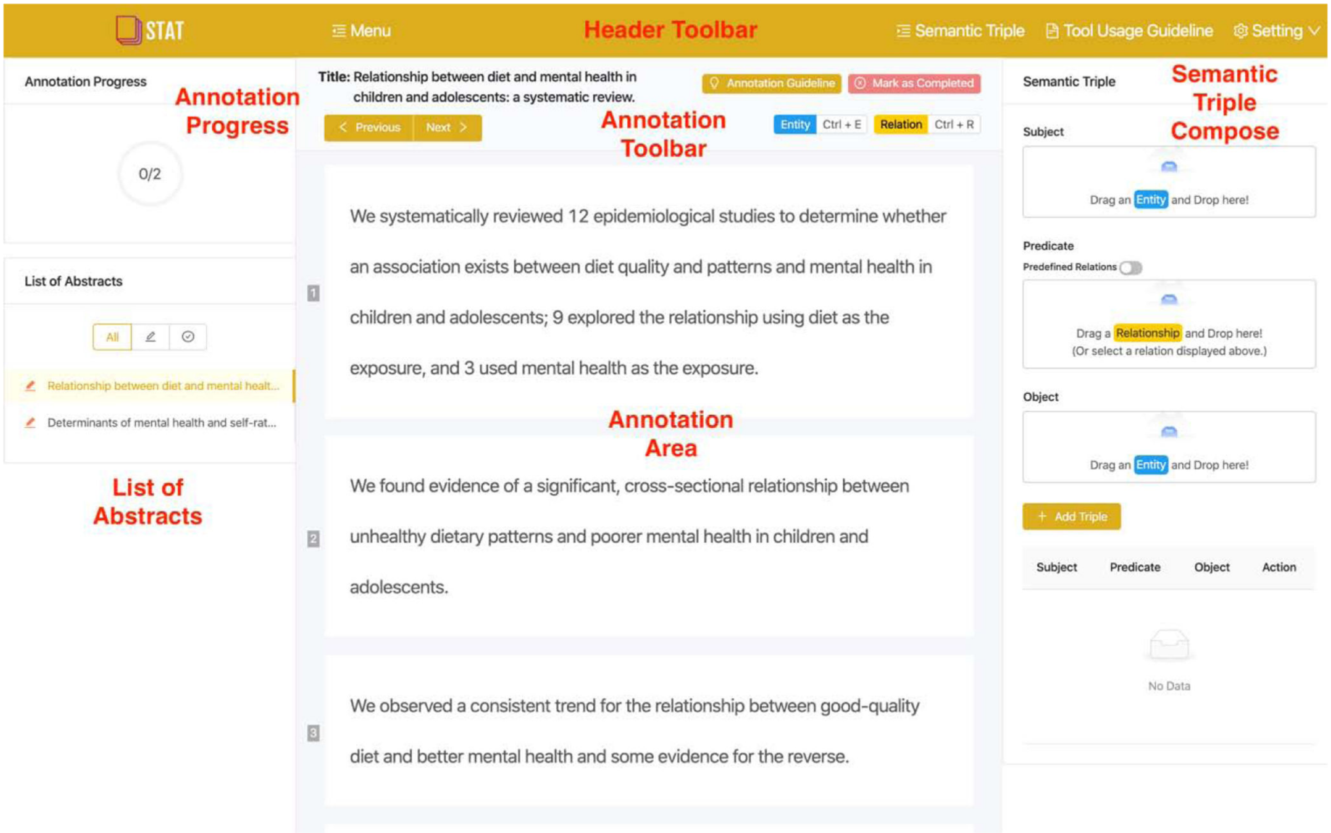


Figure 2.

The system architecture of STAT.



Title: Relationship between diet and mental health in children and adolescents: a systematic review.

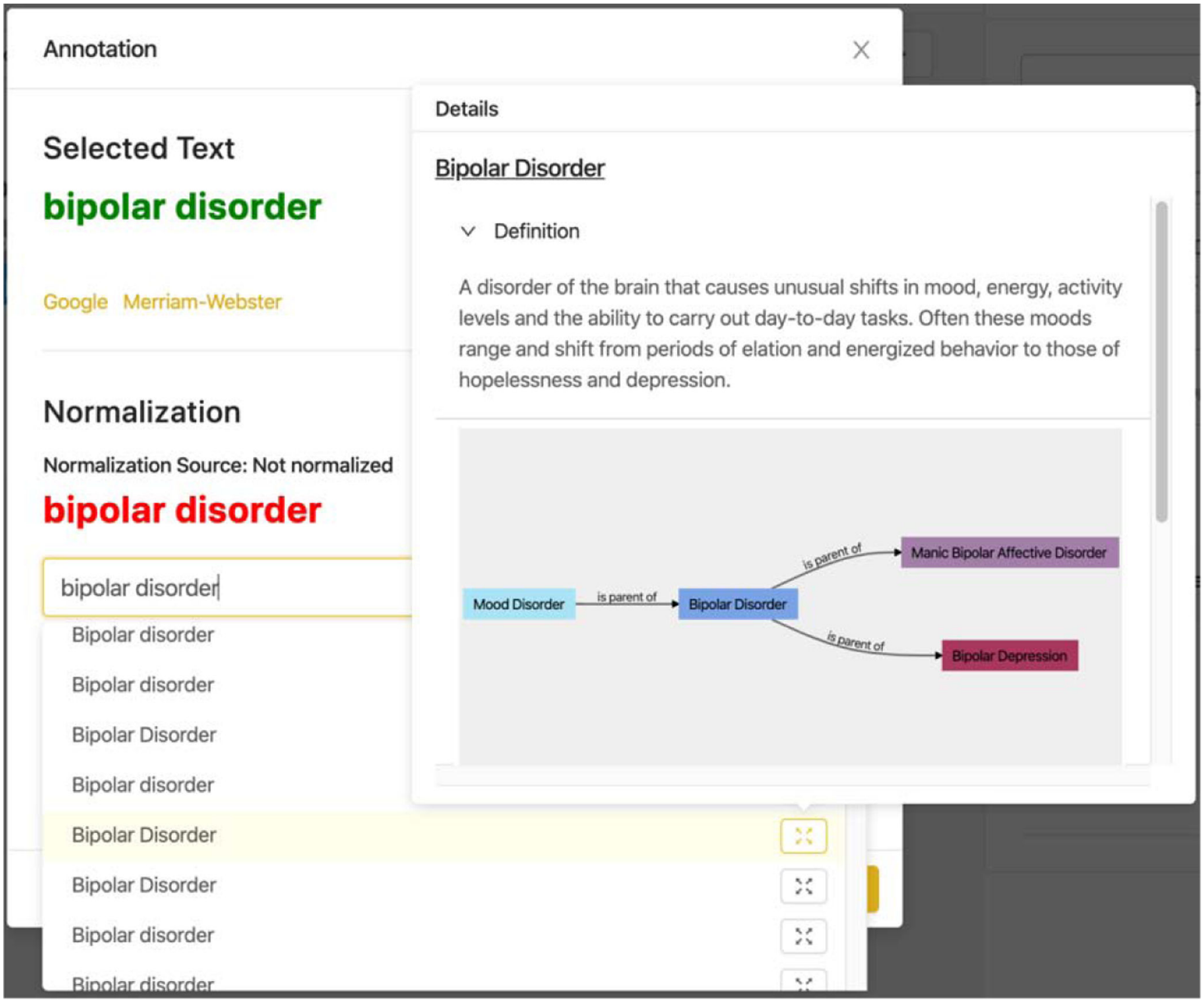
We systematically reviewed 12 epidemiological studies to determine whether an association exists between diet quality and patterns and mental health in children and adolescents; 9 explored the relationship using diet as the exposure, and 3 used mental health as the exposure.

We found evidence of a significant, cross-sectional relationship between unhealthy dietary patterns and poorer mental health in children and adolescents.

We observed a consistent trend for the relationship between good-quality diet and better mental health and some evidence for the reverse.

Figure 3.

The main user interface of STAT after the internal user-centered design iterations



Normalization Source: Not normalized

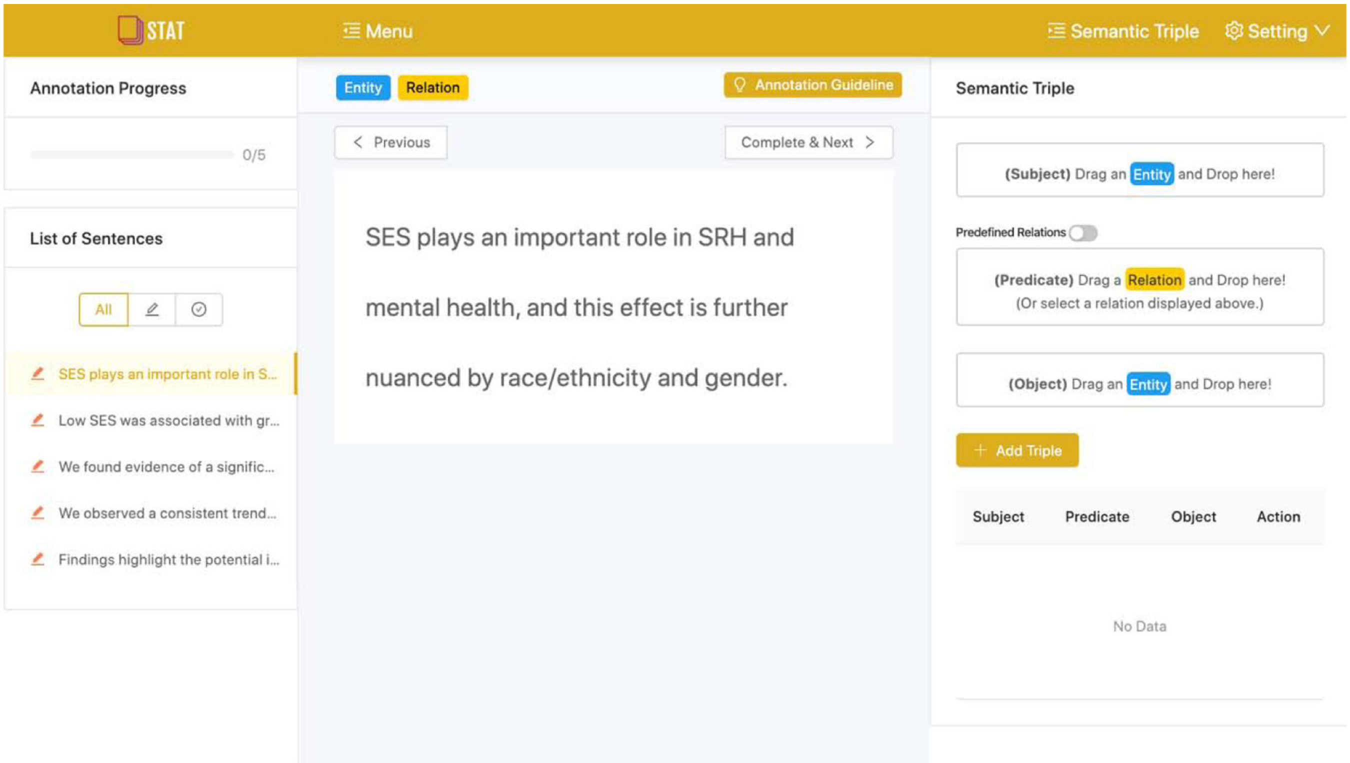
bipolar disorder

bipolar disorder

A disorder of the brain that causes unusual shifts in mood, energy, activity levels and the ability to carry out day-to-day tasks. Often these moods range and shift from periods of elation and energized behavior to those of hopelessness and depression.

Figure 4.

The normalization pop up window with a visualization of the ontology used to normalize extracted terms.



- SES plays an important role in S...
- Low SES was associated with gr...
- We found evidence of a signific...
- We observed a consistent trend...
- Findings highlight the potential i...

SES plays an important role in SRH and mental health, and this effect is further nuanced by race/ethnicity and gender.

Figure 5.

The simplified user interface of STAT after the fourth UCD iteration.

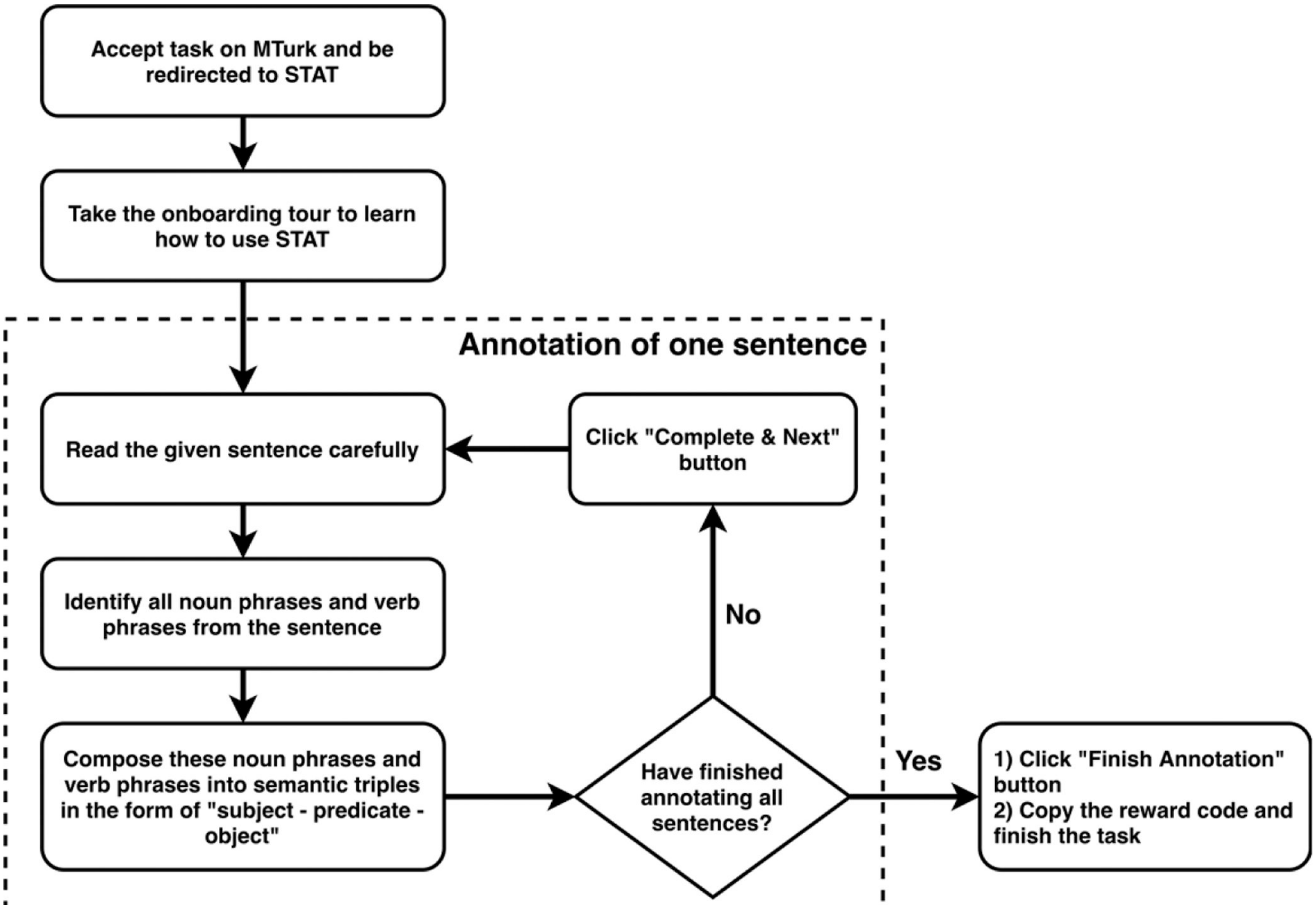
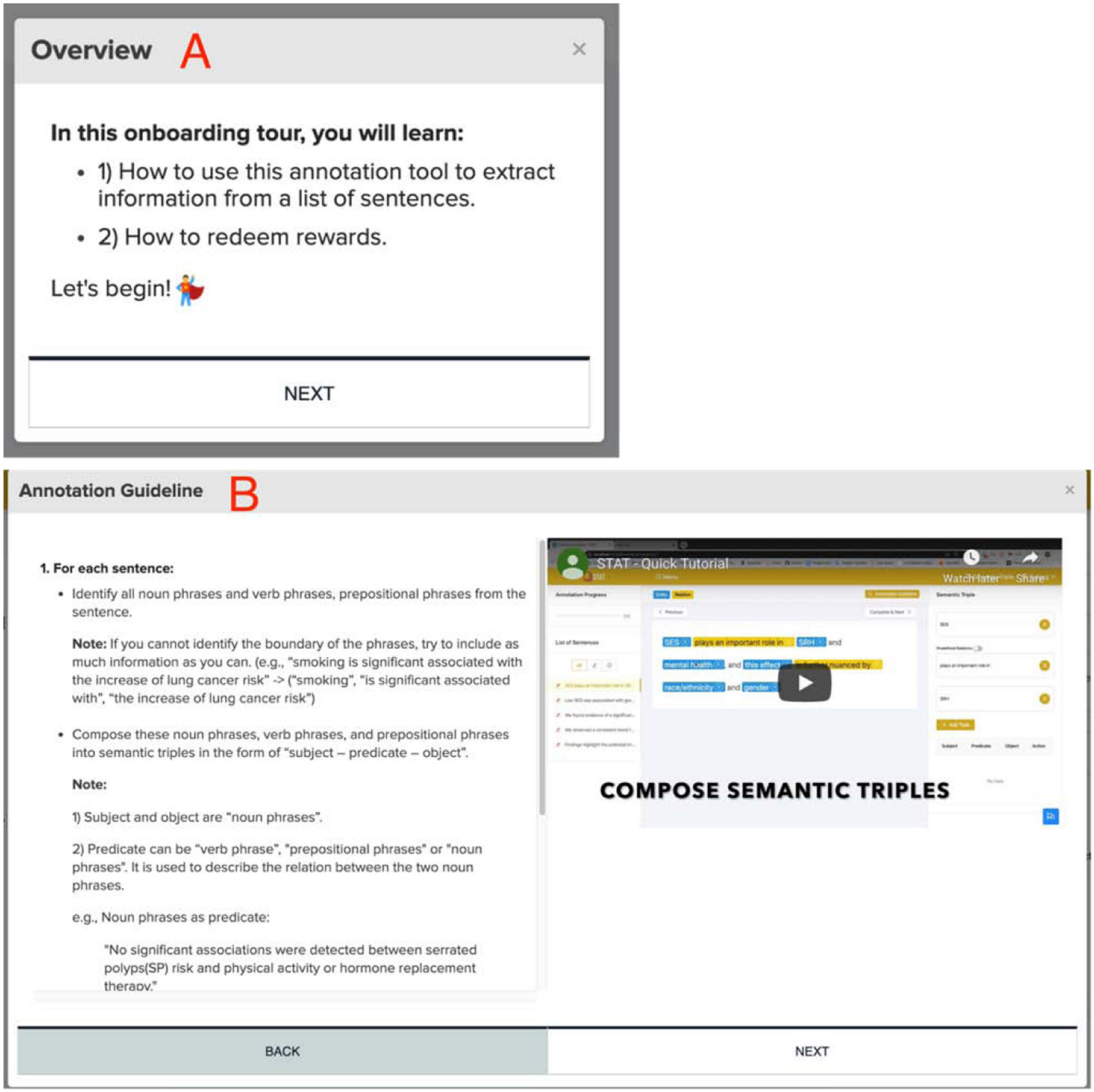

Accept task on MTurk and be redirected to STAT
Take the onboarding tour to learn how to use STAT
Annotation of one sentence
Read the given sentence carefully
Click "Complete & Next" button
Identify all noun phrases and verb phrases from the sentence
No
Compose these noun phrases and verb phrases into semantic triples in the form of "subject - predicate - object"
Have finished annotating all sentences?
Yes
1) Click "Finish Annotation" button
2) Copy the reward code and finish the task

Figure 6.

The simplified annotation workflow with STAT.



Overview

In this onboarding tour, you will learn:

- 1) How to use this annotation tool to extract information from a list of sentences.
- 2) How to redeem rewards.

Let's begin! 🦸

NEXT

Annotation Guideline

1. For each sentence:

- Identify all noun phrases and verb phrases, prepositional phrases from the sentence.

  Note: If you cannot identify the boundary of the phrases, try to include as much information as you can. (e.g., "smoking is significant associated with the increase of lung cancer risk" -> ("smoking", "is significant associated with", "the increase of lung cancer risk")

- Compose these noun phrases, verb phrases, and prepositional phrases into semantic triples in the form of "subject – predicate – object".

  Note:

  1) Subject and object are "noun phrases".

  2) Predicate can be "verb phrase", "prepositional phrases" or "noun phrases". It is used to describe the relation between the two noun phrases.

  e.g., Noun phrases as predicate:

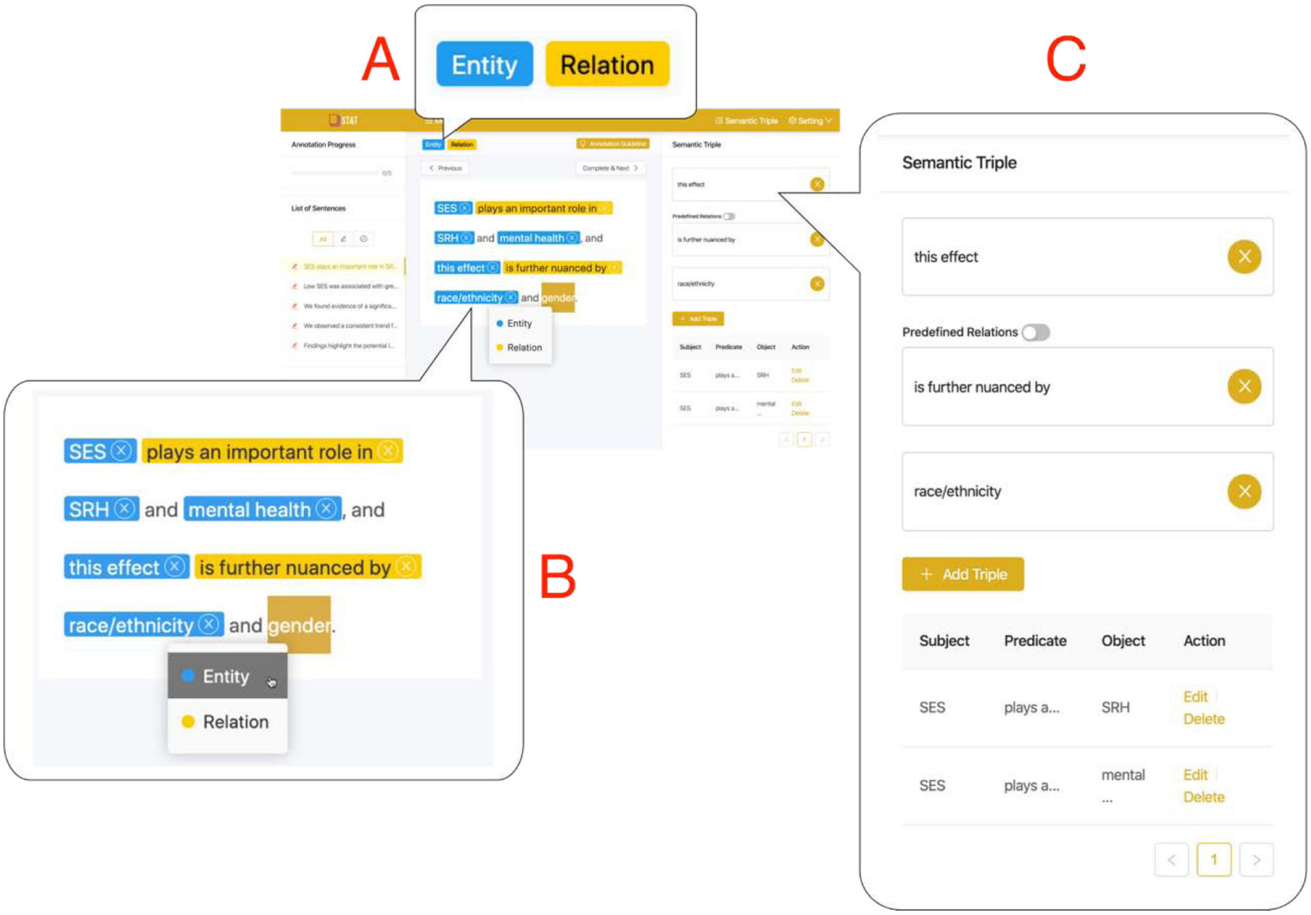
  "No significant associations were detected between serrated polyps(SP) risk and physical activity or hormone replacement therapy."

BACK

NEXT

Figure 7.

A) The "Overview" step of the onboarding tour; **B)** the "Annotation Guideline" step of the onboarding tour.



SES plays an important role in SR...
Low SES was associated with gre...
We found evidence of a significa...
We observed a consistent trend f...
Findings highlight the potential i...

SES plays an important role in SRH and mental health, and this effect is further nuanced by race/ethnicity and gender.

Semantic Triple

this effect

Predefined Relations

is further nuanced by

race/ethnicity

Subject	Predicate	Object	Action
SES	plays a...	SRH	Edit Delete
SES	plays a...	mental ...	Edit Delete

Figure 8.

A) Buttons that support the annotation operation; **B)** a context menu after a mouse right-click that supports the annotation operation; and **C)** a "Semantic Triple" module that supports the triple composition operations.

Table 1.

Results of each UCD iteration.

Iteration	Internal/External	# of Participants	SUS Score	Main Issues
1	Internal	8	70.3 ± 12.5	UI, Function
2	Internal	8	81.1 ± 9.8	UI, Function
3	External	14	55.7 ± 20.1	Annotation Guideline; Workflow
4	External	17	73.8 ± 13.8	Annotation Guideline

Table 2.

Demographic information of the participants in each UCD iteration.

	Iteration 1	Iteration 2	Iteration 3	Iteration 4
	N = 8 (100%)	N = 8 (100%)	N = 14 (100%)	N = 17 (100%)
Age				
18–24	2 (25.0%)	3 (37.5%)	1 (7.1%)	0 (0.0%)
25–34	6 (75.0%)	5 (62.5%)	5 (35.7%)	7 (41.2%)
35–44	0 (0.0%)	0 (0.0%)	4 (28.6%)	6 (35.3%)
45–54	0 (0.0%)	0 (0.0%)	1 (7.1%)	3 (17.6%)
55–64	0 (0.0%)	0 (0.0%)	3 (21.4%)	1 (5.9%)
Gender				
Male	5 (62.5%)	4 (50%)	—	—
Female	3 (37.5%)	4 (50%)	—	—
Race				
White	0 (0.0%)	2 (25.0%)	8 (57.1%)	9 (52.9%)
Black or African American	0 (0.0%)	0 (0.0%)	1 (7.1%)	1 (5.9%)
Native American or American Indian	0 (0.0%)	0 (0.0%)	1 (7.1%)	0 (0.0%)
Asian / Pacific Islander	7 (87.5%)	6 (75.0%)	4 (28.6%)	7 (41.2%)
Other	1 (12.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Education				
Doctorate degree	0 (0.0%)	1 (12.5%)	0 (0.0%)	0 (0.0%)
Master’s degree	7 (87.5%)	0 (0.0%)	2 (14.3%)	4 (23.5%)
Bachelor’s degree	0 (0.0%)	0 (0.0%)	8 (57.1%)	10 (58.8%)
Associate degree	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (5.9%)
Professional degree	0 (0.0%)	6 (75.0%)	1 (7.1%)	0 (0.0%)
Some college credit, no degree	0 (0.0%)	1 (12.5%)	2 (14.3%)	1 (5.9%)
High school graduate, diploma or the equivalent	0 (0.0%)	0 (0.0%)	1 (7.1%)	1 (5.9%)
Other	1 (12.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Table 3.

Usability issues identified in each UCD iteration's usability testing.

Iteration	Theme	Usability issues	Nielsen's 10 usability heuristics
1	Unclear information presentation	Hard to find the guideline button	Recognition rather than recall
	Incomprehensible concepts	Hard to understand some concepts used in STAT (e.g., entity, relation, and semantic triple)	Help and documentation
	Lack of functionality	Could not create a new relationship if no relation has been identified	Flexibility and efficiency of use
		Unsure how to create complex triplets when there are multiple associations presented	Flexibility and efficiency of use
		The annotation box appears every time by default; and there is no way to disable it	Flexibility and efficiency of use
		Cannot change the size and format of the text	Flexibility and efficiency of use
2	Unclear information presentation	Hard to find the menu button, because it was placed close to the STAT logo	Recognition rather than recall
		Hard to see existing (defined) semantic triples	Recognition rather than recall
		Tagging "*Previous Abstract*" and "*Next Abstract*" is redundant and somewhat bothering the participants to focus on the annotation work	Aesthetic and minimalist design
		Hard for a user to find the setting option for the "*normalization popup window*"	Recognition rather than recall
	Lack of functionality	Could not change the "*entity*" or "*relation*" type in the popped window	Flexibility and efficiency of use
		Difficult to find relations and put them into the semantic triple form when there were more than two entities being related	Flexibility and efficiency of use
		The font size of the entities that have been annotated will not change after adjusting the font size setting	Consistency and standards
3	Unclear documentation and annotation guideline	Confusion between entity and relation	Help and documentation
		Lack of an overall cohesive explanation of the point of the task explained in very simple terms	Help and documentation
		The normalization step was unclear	Help and documentation
		Learning of the functionalities was a little hard	Help and documentation
		Users were not certain whether performing the task correctly, despite reviewing the video and guidelines several times	Help and documentation
	Lack of functionality	Hard to remove a term normalized in error	Help users recognize, diagnose, and recover from errors
	Bugs	The definition box froze and needed to refresh the page	Error prevention
4	Incomplete functionality	Hard for users to highlight things precisely since they would include the leading or trailing space when highlighting things.	Flexibility and efficiency of use
	Unclear documentation / guidelines	Lack of examples	Help and documentation
		Hard to understand what terms are fitting for criteria	Help and documentation
		Hard to construct a relationship	Help and documentation
		Hard to understand the basic term (e.g., entity and relation)	Help and documentation

Iteration	Theme	Usability issues	Nielsen’s 10 usability heuristics
	Unclear information presentation	Finished triples were not visible properly	Flexibility and efficiency of use
	Bugs	The triples in the queue are not added or removed when moving to the next sentence	Error prevention

Table 4.

Selected important improvements suggested in each UCD iteration.

Iteration	Improvement category	Improvement action
1	Adding new functions	Provide a link to the MTurk website where workers can redeem their rewards
		Provide functions to support the change of font size
		Provide support for the creation of relations
		Provide a way in addition to the context menu to distinguish whether a selected text is an entity or a relation
	Improving existing functions	Support editing of created semantic triples
		Support "drag and drop" of entities or relations between different boxes within the "Semantic Triple" composition panel
		Allow users to create complex triples when there are multiple associations present in the sentence
		Give users a warning when deleting a semantic triple (e.g., "*delete*", "*confirm delete*")
	Improving UI and UX	Add a label to distinguish each panel (e.g., "*Annotation Progress*", "*List of Abstracts*")
		Clarify the meaning of the different coloring for each sentence
		Remove the hover effects on the annotation buttons
		Change the word "*documentation*" to "*guidance*"
	Improving documentation and annotation guideline	More detailed documentation with explanations of the different concepts and provide more sensible examples and demonstrations
2	Adding new functions	Include keyboard short cuts (i.e., ⌘ + E/R) for mac users to annotate entities and relations
		Support the fast composition of multiple entities relating to multiple entities
	Improving existing functions	Toggle between "*Complete*" and "*Next*" button when any changes happened
		Make the annotation guideline pop up automatically
	Improving UI and UX	Enlarge the menu buttons
	Improving documentation and annotation guideline	Add further description on what entities and relations are
3	Improving UI and UX	Add tool tip with hover effect
	Improving documentation and annotation guideline	Add more examples in the annotation guideline
		Add simple examples to the demo video to help end-users understand
		Make the instructions more explicit
		The video should have a narrative explaining what is happening as opposed to a music track
		Convert the usage guideline to an onboarding video tour
	Improving existing functions	Add more normalization terms for predicates, like "*cause*", "*associated with positive direction*", and "*associated with negative direction*"
	Simplifying workflow	Remove the "*normalization*" step from the crowdsourcing workflow
4	Improving documentation and annotation guideline	Add more examples and instructions
		Simplify the instruction to make it easier to understand.
	Improving UI and UX	Display triples clearly
		Display all triples of a sentence in one table

Table 5.

Worker performance of the annotation tasks by sentence.

			Individual Worker's Average Performance[a]			Majority Vote Based Performanc[b]				
Sentence	**# of Entities and Relations**	**# of Triples**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Worker Average Time (seconds)**	**Expert Average Time (seconds)**
1	10	8	0.84	0.61	0.69	1.00	1.00	1.00	209.6	348.5
2	7	5	0.87	0.66	0.74	1.00	1.00	1.00	199.5	249.5
3	3	1	0.85	0.84	0.84	0.86	0.86	0.86	66.5	111
4	3	1	0.78	0.75	0.76	0.83	0.83	0.83	62	238
5	3	1	0.64	0.62	0.62	0.75	1.00	0.86	61.8	293

[a] Individual Worker's Average Performance: The performance scores (i.e., precision, recall, F1-score) is the average score of the 5 crowdsourcing workers' annotation results compared to the gold standard.

[b] Majority Vote Based Performance: The 5 crowdsourcing workers' annotation results are aggregated based on majority vote (e.g., we consider an annotation as valid, if 3 out of the 5 workers agreed on the same annotation); and the aggregated result is compared to the gold standard to calculate the performance scores.