Usability Testing of an Annotation Tool in a Cultural Heritage Context

Karoline Hoff[1] and Michael Preminger[1]

Oslo and Akershus University College of Applied Science, Oslo, Norway
karoline.hoff@outlook.com, michaelp@hioa.no

Abstract. This paper presents the result of a usability test of an annotation tool. The annotation tool is implemented, used and tested in a cultural heritage context (CH), the TORCH project at the Oslo and Akershus University College of Applied Science. The experiments employed non-experts with the intention of facilitating for crowd-sourcing of annotations. Interesting problems and usability patterns from the literature manifest in our experiments. Despite some weaknesses in the interface of the tool version used for the experiments, the annotators show a reasonable rate of success.

1 Introduction

The proliferation of semantic web application in recent years, has been followed by development of and research into automatic conversion of legacy data into semantic-aware formats, identifying entities and roles in those data. In order to make advances here, ground truths are needed, and in order to establish those, we need to have large samples of those data annotated.

Such annotation is performed intellectually and is facilitated by annotation tools. The best annotation results (e.g. ground truth) would be expected if the annotators were both domain experts and technically proficient. In most cases, alas, such annotators are scarce and expensive. For very specialized knowledge domains, such as agriculture, spacecraft and the like, domain experts would be needed, although leniency could be shown with regards to the requirement of technical proficiency. In domains like cultural heritage, which is the domain of our project, leniency could be shown with regards to both requirement, and crowd-sourcing could be used, given that the annotation tool facilitates non-experts input, and the usability is good. Moreover, particularly for cases that need great amounts of data, crowd-sourcing would be the only viable alternative, which challenges the design of the annotation tool.

The TORCH project at our institute is a conversion endeavour, attempting to convert programme description from the archives of the Norwegian national broadcaster (NRK) into Semantic aware formats. As described in [1], we have developed our own annotation tool, which we, in the long run, would like to adapt to a wider context and make hospitable to crowd sourcing. Successful crowdsourcing demands that the tool has as few usability issues as possible, which is the main objective of this endeavour.

In this article we report a study into the process of manual annotation by end-users using our annotation tool. For this, both information expert and non-expert users were observed in a usability study. Our goal is to gain insight into the end-users understanding of the tool, and to develop an adequate design for the annotation tool, making it usable for non-experts. This paper reports on the results of this user study.

The remainder of this paper is structured as follows: in section 2 we explore related work in evaluating the usability of manual annotation tools. Section 3 introduces our annotation tool, while section 4 explains the methodology and setup of our study. Section 5 contains the results of the study, and discusses the implications of the results.

We start by presenting the current state of the art of user testing in annotation

2 Background

2.1 Usability Fundamentals

[2] claims that "usability is not some vague postulation, but actually a criterion that can be measured and systematically engineered".

Burghardt further explores this in his doctoral dissertation ([3]). Here Burghardt emphasizes the importance of an interface that makes the annotation process as convenient and efficient as possible, as manual annotation is typically a laborious task. To investigate the usability of the annotation tool, we conducted a study with participants who could be typical users of these systems.

Usability can be described as how well a system can be used, or the users' ability to carry out the task successfully. Nielsen ([4]) states that usability cannot be described as a one-dimensional criterion, but must be seen as a concept defined by five quality components: *learnability*, *efficiency*, *memorability*, *error rate* and *satisfaction.*

Each of these usability components can be measured individually. Learnability tells us something about how easy it is for users to accomplish basic tasks when they encounter the design. Efficiency is how quickly a user familiar with the design can perform tasks. Memorability is about how easily a user returning to the product after a period of not using it can re-establish proficiency. The error rate tells us how many errors the users make, how severe these errors are, and how easily a user can recover from these errors. Satisfaction is the users own experience and perception of the design.

A more formal definition, like the ISO 9241-11 standard from The International Organization for Standardization ([5]), states that the usability of a product is "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use".

2.2 Usability Testing

There are several ways to evaluate the usability of a product, depending on factors like available resources, evaluator experience, ability and preference, and the stage of development of the product under review. Scholtz ([6]) states that the three most discussed evaluation methods are user-based, expert-based and model-based. In user-based evaluation methods individuals from a sample of the intended users try to use the application. Expert-based methods makes a usability expert do an assessment of the application. In model-based methods an expert employs formal methods to predict one or more criteria of user performance. In their user study, Hinze et.al. ([7]) state that usability typically is a factor in the interface development of manual annotation tools, and that end-user evaluations of interface and user interaction are very rare. Hinze et. al. also state that typical evaluation strategies of Semantic Web technologies seldom contain complex user aspect, despite aspects of Humancomputer interaction and user involvement being identified as important aspects.

> Few studies involving end-users have been executed in the context of semantic annotations. In particular, manual annotation tools have so far not been systematically evaluated for appropriate interaction design and semantic understanding. System evaluations that incorporated human participants did not seek their feedback on interaction issues nor did they evaluate the participants' mental models of the system interaction. So far, issues of understanding of semantic annotations by (non-expert) users have not been studied in a systematic manner. [7]

Like Barnum ([8]), when we talk about usability testing we are referring to the activity that focuses on observing users working with a product, performing tasks that are real and meaningful to them.

For usability testing to be an effective tool for understanding user interface design strengths and weaknesses, it needs to engage actual users in performing real work. Conducting formative usability testing as part of an iterative design strategy is the most reliable way to develop a truly usable product ([9], s. 7).

2.3 Usability Testing of Annotation Tools

If crowdsourcing is to become the reality, one of our goals has to be to create an environment where non-expert users are able to create meaningful and consistent annotations. Hinze et.al. identified following key requirements for non-expert users of manual annotation tools: *established interaction patterns*, *simple vocabularies*, *contextual semantic identity* and *focus on the user's task* ([7]).

"Established interaction patterns" entails making semantic web applications look like traditional applications and to use familiar interaction paradigms. Simple vocabularies is pretty self-explanatory, it is important because research indicates that complex category structures are disadvantageous for quality annotations. Contextual semantic identity means bridging the gap between objective knowledge (as encoded in the RDF data model) and subjective knowledge of human cognition, e.g. computers identifying resources by URIs in RDF, versus humans identifying entities with labels and disambiguate meaning using context. ”Focus on the user’s task” entails that semantic authoring and semantic annotations is integrated in a good way. This is not applicable to us, as the TORCH annotation tool is a pure annotation tool ([7]).

Burghardt describes 17 unique requirements for annotation tools, whereas two requirements are categorized as pure usability requirements, and eight requirements are categorized as mixed functional and usability requirements. Five of these requirements are relevant to us, as they describe usability aspects, and are categorized as relevant for the core group Burghardt calls ”Annotators” i.e. our tool’s user group. We have also left out the mixed requirements for functions our tool does not possess ([3, p. 77-81]). The two pure usability requirements are *documentation*, i.e. the availability of a user manual, which is important for the learnability of a system, and the general purpose requirement of an *easy-to-use interface*. The three mixed requirements relevant to us are *visualization of primary data*, meaning the original text should be displayed correctly in the annotation tool, and the tool should clearly differentiate between original and annotated text. *Visualization of annotation*, closely tied with the previous requirement. *Marking of anchors*, i.e. providing an interface making for an intuitive and effective selection of different anchor scopes ([3, p. 77-81]). The documentation and the three relevant mixed requirements are fulfilled as described in 3, and while our study is designed with these requirements in mind, the requirement of an easy-to-use interface is our main focus ([3, p. 77-81]). Within the mentioned usability requirements, Burghardt also identified twenty-six usability patterns for the domain of manual annotation tools. These are divided into the categories *general UI*, *installation*, *primary data*, *annotation scheme*, *annotation process* and *annotation visualization* ([3, p. 143]). We will use these patterns to identify potential problems with our tool later.

3 The Annotation Tool

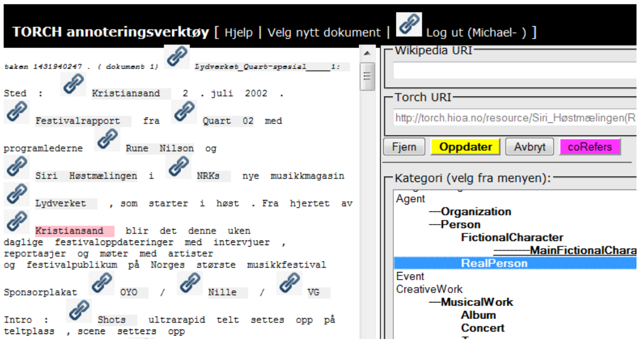
Our annotation tool (first described in [1]) is developed in order to provide ground truth-data for the TORCH project[1]. In order to allow the collection of enough data, we believe that crowd-sourcing with the emphasis of non-experts, is mandatory. Our tool is developed for use in the context of Cultural Heritage (broadcasted material), where the term non-experts refers to quite a wide public. We believe nonetheless, that also other contexts to which annotating textual materials is relevant could gain a larger number of annotators, and more annotated material of high quality, if the annotation process is made more accessible.

To this end, we seek to provide simplicity in use. Moreover, our assumption is that annotators, be it domain experts or non-experts, do not read guidelines very thoroughly, which means that heavily basing correct annotations on

[1] The TORCH project is an activity of the research group Information systems based on metadata: `http://tinyurl.com/k8gf7dr`

detailed guidelines is hazardous. Therefore the usage of the tool should be as self-explanatory as possible, the approach to the annotated material should be gradual in the process, and learning while annotating possible without compromising the results.

The thought behind the design is that named entities are annotated first, with as few as possible classes to choose from. The classes are ordered in hierarchies, the annotation done with as specific as possible classes (closest to the leaves), with the provision to fall back to more general (closer to the root) classes. This means that RealPerson is available as a class, but not Actor (role). Here the user is prompted to choose a Wikipedia URI (from a list constructed on the fly for each mention), to provide the mention with a unique, global URI. The second phase in the annotation is assignment of relations between already annotated entities. The current experiments do not test the assignment of other than co-references, which have a specially designed short-cut.



```
taken 1431940247 . ( dokument 1)  Lydverket_Quart-spesial____1:

Sted  :   Kristiansand   2  . juli  2002  .
  Festivalrapport   fra   Quart  02  med
programlederne   Rune  Nilson  og
  Siri  Høstmælingen  i   NRKs   nye  musikkmagasin
  Lydverket   , som  starter  i  høst  . Fra  hjertet  av
  Kristiansand   blir  det  denne  uken
daglige  festivaloppdateringer  med  intervjuer  ,
reportasjer  og  møter  med  artister
og  festivalpublikum  på  Norges  største  musikkfestival

Sponsorplakat   OYO   /   Nille   /   VG

Intro  :   Shots   ultrarapid  telt  settes  opp  på
teltplass  , scene  setters  opp
```

http://torch.hioa.no/resource/Siri_Høstmælingen(R

Fig. 1. The annotation tool screen with the class selector activated.

4 Usability Testing of the TORCH Annotation Tool

Our goal is to develop an adequate design for the annotation tool, making it usable for non-experts, and to gain insight into the end-users understanding of the tool. We wanted to know what usability issues could be identified, what problems users experienced during the annotation process, how satisfied the users were with the tool, and how the tool could be redesigned to minimize above issues.

We studied the usability of the annotation tool in an end-user study. We performed the usability tests and interviews with real users. Even though the annotation tool is fully operational, this user study made only use of the functions relevant for phase one of the annotation process, and the making of co-references. This was decided at an early stage, to make the design manageable in terms of variables and causality. Moreover, this is the phase that will most probably be exposed to crowd-sourcing, as we are still considering annotating relations by the TORCH staff.

Data was collected in two ways: through observation and through questionnaires. The facilitator sat with each participant throughout the session, recording observations, noting any difficulties, any comments made by the participant, and whether they successfully completed the task. Participants were asked at the beginning of the session to "think-aloud" ([4, p. 195-198]).

Besides the standard instructions given to all participants, no further explanations or assistance were given. In cases where the participant forgot to think-aloud, the facilitator would ask "what did you expect to happen there?". If the participant asked for more instructions, the facilitator would remind the participant that we were testing the usability of the tool and needed to see if people could use it without further explanations.

The participants were given realistic tasks to perform by interacting directly with the tool. The tool does not explain how the interface is supposed to work, making it possible to identify which parts of the interface are self-explanatory and which parts are confusing.

Before starting the actual interaction with the tool, each participant was asked to read the guidelines. The facilitator clarified the nature of the task, and explained how the initial interface worked. Each task began with a brief scenario explaining the goal of the task from a user's perspective, followed by task instructions, and instructions on how to report the task as completed. A short summary of the user tasks is presented below.

- Task 1 Annotate two personal names of your choice
- Task 2 Annotate two creative works of your choice
- Task 3 Annotate a geographical place of your choice
- Task 4 Create an equivalence-relation between two annotations of your choice (in our own terminology called co-reference).
- Task 5 Correct the errors in the pre-existing annotation
- Task 6 Delete an annotation of your choice

We used a text that contained only general knowledge concepts for the annotation. With the exception of one incorrectly annotated entity, the experimental text was plain and free of annotations. The error was annotated by the facilitator beforehand, making it possible to ask the participants to find and correct it during the test.

The participants had a pen and a printout of the tasks available during the whole session. During the sessions, the participants were asked to evaluate every feature of the interface design that they thought should be changed in looks or functionality based on their expectations of the tool. They were not obliged to follow the guidelines, but they were asked to think aloud and justify the decisions they made.

The participants were requested to notify the facilitator when they assumed they had finished the task, or gave up trying.

5 Results

Here we report the results of our study, and describe our observations of how participants interacted with the annotation tool. We differentiate between observations about the participants' interaction with the tool and the creation of annotations, quality of annotations and participant feedback. We have categorized our findings based on [3].

5.1 The Data Collected

We collected data as follows

- Participant demographics (collected pr. questionnaire before the test itself)
- Think aloud recording summaries
- Pre-task expectations per task (collected pr. questionnaire at the beginning of each task, not reported here)
- Feedback: post-task experience with reference to pre-task expectations (collected pr. questionnaire at the end of each task)
- System usability scale (not reported here)
- The annotation results

5.2 Participant Demographics

As we aspire to make the tool usable for non-experts, we selected seven participants with varied backgrounds and varied levels of technical knowledge. The participants were asked to rate their knowledge on a 5-point scale, with 1 being "no knowledge" and 5 being "very knowledgeable". We surveyed their familiarity with word processing as a measure of computer literacy, tagging as an annotation task, familiarity with usability and semantic web as technical expertise, respectively. Fig. 2 shows the distribution of expertise for the 7 participants. All participants are computer literate. Four participants were familiar with tagging and semantic web, while three participants knew little of it. Based on their self-assessment, we identified participants U1-U4 as technical experts, and U5-U7 as non-experts.

5.3 Observed Usability Problems

Here we report the usability problems described in [3] that we could identify by observing the participants' interactions with the tool.

Within the mentioned usability requirements, Burghardt also identified twenty-six usability patterns for the domain of manual annotation tools. These are divided into the categories *general UI*, *installation*, *primary data*, *annotation scheme*, *annotation process* and *annotation visualization*. We will use these patterns to identify potential problems with our tool.

The main bulk of the problems found belong in the "General UI" part of Burghardts listing of usability patterns.

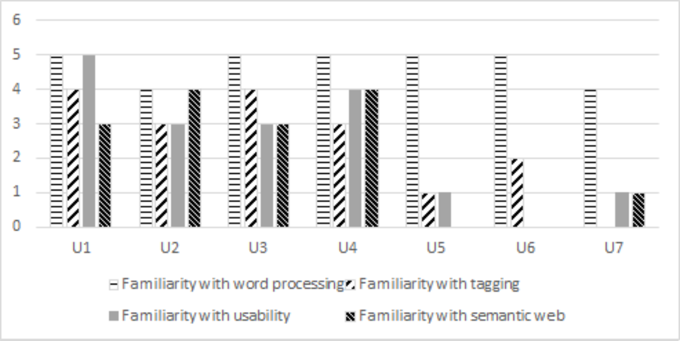


Fig. 2. The distribution of expertise for the 7 participants.

In ”General IU”, Burghardt has identified some problems that influence the overall user experience with the tool, with no explicit reference to any particular stage in the annotation work-flow. There are no generic solutions for these issues, but of the patterns Burghardt mentions, we could identify the following in our own study: Safe exploration, some users reported feeling afraid they would ”break something”. One user said ”I feel like the developer is trusting me as a userperhaps a bit too much”. We obviously have something to gain by making the tool seem less menacing, and ensure the users know that no mistake are unfixable.

We also identified the pattern Burghardt calls ”Help for domain-specific functions” (”domain” here, unlike elsewhere in the paper, referring to semantic annotation as an activity.). Several users reported being unsure of exactly what a co-reference was, and thus being unsure if they had annotated it correctly. As seen in Figure 3, creating co-references was the task with the greatest number of semantically meaningless results, despite the participants having read the section on co-references in the guidelines before starting the task. Burghardt proposes several solutions, one of which being a brief explanation inside the tool. Providing help for the unintuitive, domain-specific functions, increases the learnability of the tool. This is especially important for non-expert users, and a thing we will have to consider within the TORCH project if we are to crowd-source.

Burghardt lists ”Redundant controls” as a problem, and two participants were unsure of what ”set” of buttons they were to use. Burghardt also mentions ”no explicit save action” as a problem, and one participant in our study (an expert as it happens) exclaimed ”I do not want to *update* anything, I want to *save*”. However, he was the only one commenting on the wording of the button. In line with Burghardts ”General UI” item, one other participant commented on the colour, saying yellow made him think of danger. Three participants commented on the amount of primary data, feeling a bit overwhelmed. Burghardt describes ”tailored display of data”, and recommends the possibility of customizing things like font-size, font-family and line spacing. This is to give the user control and freedom, and ultimately increase the annotation speed, or *efficiency.* In the category *annotation process*, Burghardt identifies the steps in the annotation process,

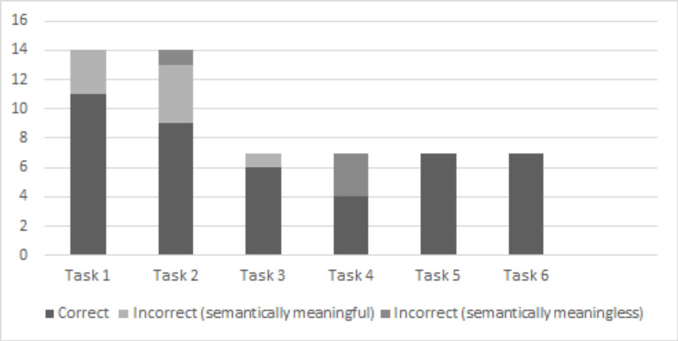


Fig. 3. Success rate by task.

and to us these are the important ones: *create anchor*[2], *select annotation*, *apply annotation*, *edit annotation* and *delete annotation*. All participants had some trouble with highlighting the anchor. Having to start the marking in the middle of the word they wished to select seemed foreign to them, but having done it once, all but one found it relatively easy. One participant wanted to find and select all instances of an entity with the ctrl+f-function. Another participant wanted to mark several instances of the same entity and annotate them at the same time to gain efficiency. Selecting annotations did not emerge as a problem, and as seen in Figure 3, each participant was able to successfully edit and delete annotations.

5.4 Quality of Annotations

Figure 2 shows a summary of the quality of the annotations created by the participants. The outcome of each task was recorded by the facilitator according to the following possible outcomes:

- an annotated entity which corresponds with the ground truth established by the facilitator is considered correct
- an annotated entity is considered incorrect, but semantically meaningful if it refers to a named entity (e.g., if participants annotated ”filmen Øyenstikkeren”[3])
- The annotations are considered semantically meaningless if they do not refer to a named entity, (e.g., ”innspillingen”[4])

All seven participants created at least one correct annotation, and three participants created nothing but correct annotations. There were two kinds of incorrect, but semantically meaningful annotations: Some were assigned the wrong Wikipedia URI, e.g. assigning the *movie* Øyenstikkeren [5] the article about the

[2] anchor being the highlighted text to be annotated
[3] The movie ”Øyenstikkeren”
[4] The shooting (of a movie)
[5] dragonfly in Norwegian

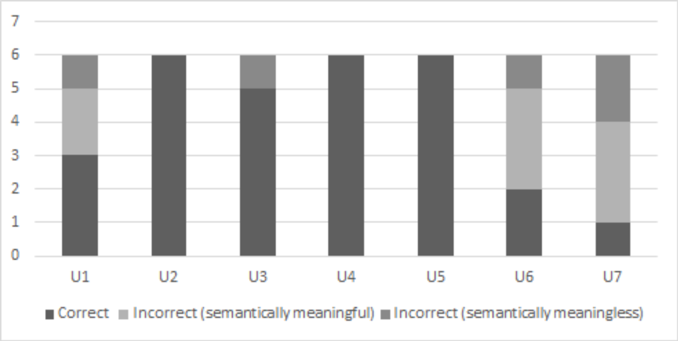


Fig. 4. Quality of annotations per participant

insect. The other kind had in common that participants included the determiner e.g. "the movie Øyenstikker" or "the host Mikal Olsen Lerøen". No one annotated wrong parts of the texts, however, this is most likely a consequence of the way we designed the tasks.

There were two kinds of incorrect annotations: some were faulty equivalence-relations between a referrer and the entity it was referring to, e.g. "the movie" referring to "Øyenstikker". The remaining were stand-alone nouns the participants wrongly interpreted as work-titles. It is clearly explained in the guidelines how to handle both of these occurrences, and none of them are to be annotated. Despite having access to the guidelines during the testing, and having to read through them in advance, only a single participant actually consulted them when in doubt. This indicates that our assumption about annotators not reading guidelines thoroughly is correct.

No participants gave up on a task without completing it, but some were unsure if they were done and if their annotation had been saved. Several participants reported that they had perceived little feedback from the tool, and were looking for some cue that they had been successful. They often commented "I *think* it did it" or "I guess I am done with that". Some participants perceived the tool as frozen when the tool was processing their request.

Like in [7] several participants switched from being an information provider to an information consumer in the course of the study. Four participants wanted to open the Wikipedia URI to read more about something, or making sure they had chosen the correct URI. Two participants showed interest in the ontology beyond what was described in the guidelines, e.g. "what if I found a pseudonym?".

5.5 Participant Feedback

Using a questionnaire as a guide while talking to the participants, we post-interviewed them about their experience in using the TORCH annotation tool. Figure 5 shows the participants' self-assessment regarding how difficult they found using the tool, with 1 being "very easy" and 5 being "very hard". We asked the participants for feedback on their experience of the different tasks:

create annotations (left), change annotation (second left), create co-references (second right), and their general experience of using the tool as a whole (right).

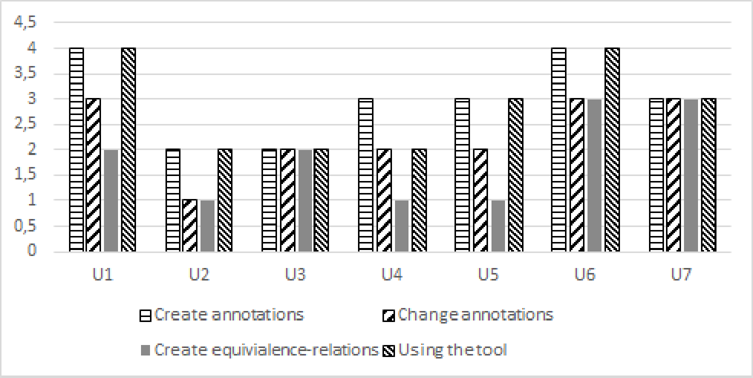


Fig. 5. How difficult the participants percieved the different tasks to be.

Five of the seven participants felt editing annotations and creating co-references was easier than creating new annotations and using the tool in general. The two others felt all activities were equally difficult. These results are also interesting in light of the quality of the created annotations (see Figure 2). U5 rated the tool fairly difficult to use (3), but all her annotations were correct. She was also the only non-expert participant with a 100 % success rate. Changing annotations was rated remarkably difficult, considering all participants successfully completed that task. We think this is because the participants were unsure whether they had indeed successfully completed the task, it being a tool-feedback issue. Figure 6 shows the participants' self-assessment regarding how interesting they found using the tool after completing the tasks, with 1 being ”not interested” and 5 being ”very interested”. Like in Figure 5 we asked for feedback on their experience of the different tasks: create annotations (left), change annotation (second left), create co-references (second right), and their general experience of using the tool as a whole (right).

6 Summary and Concluding Remarks

We have performed a usability test of the TORCH annotation tool, with emphasis on functions that we would like to expose to crowd-sourcing. The results show that annotators with reasonable computer literacy, not being domain experts, have a high success rate performing the tasks. In line with the existing literature we find that designing the interface as similar as possible to traditional applications increases annotators' confidence. We also found no clear connection between annotators' perceived difficulty and interest in participating in further activity of the kind, which is encouraging in the context of crowd-sourcing. As

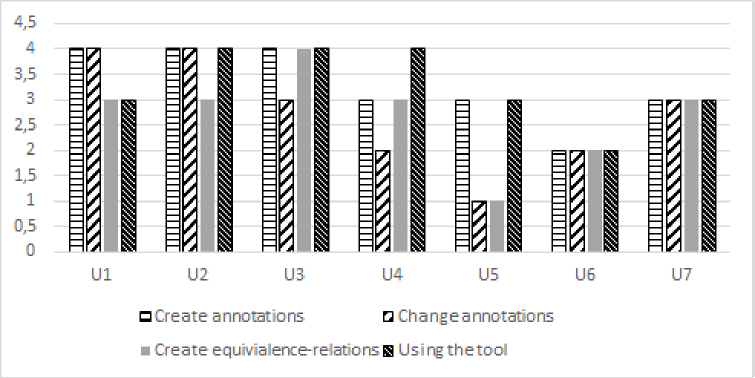


Fig. 6. The degree of interest participants were expressing.

further research we see the need to go deeper into users' attitudes towards annotating (performing in depth interviews with few participants) on the one hand, as well as exposing more of the functionality of our tool (and changes to the tool owing to the presented experiments) to further usability testing.

References

1. Tallerås, K., Massey, D., Husevåg, A.R., Preminger, M., Pharo, N.: Evaluating (linked) metadata transformations across cultural heritage domains. In: Metadata and Semantics Research - 8th Research Conference, MTSR 2014, Karlsruhe, Germany, November 27-29, 2014. Proceedings. (2014) 250–261
2. Burghardt, M.: Usability Recommendations for Annotation Tools. In: Proceedings of the Sixth Linguistic Annotation Workshop. LAW VI '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 104–112
3. Burghardt, M.: Engineering annotation usability - toward usability patterns for linguistic annotation tools (September 2014) Pattern wiki: http://www.annotation-usability.net.
4. Nielsen, J.: Usability Engineering. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
5. International Organization for Standardization: ISO 9241-210:2010 Ergonomics of human-system in- teraction Part 210: human-centred design process for interactive systems. (2010)
6. Scholtz, J.: Usability Evaluation (2004)
7. Hinze, A., Heese, R., Luczak-Rösch, M., Paschke, A.: Semantic Enrichment by Non-experts: Usability of Manual Annotation Tools. In Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E., eds.: The Semantic Web ISWC 2012. Volume 7649 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 165–181
8. Barnum, C.M.: Usability testing essentials: ready, set– test. Morgan Kaufmann, Burlington, MA (2011)
9. Norlin, E., Winters, C.: Usability testing for library websites: a hands-on guide. American Library Association, Chicago (2002)