

Gamification of an annotation task

Konrad Ukens

s1617524

k.u.ukens@student.utwente.nl

Creative Technology Bachelor Thesis

Supervisors: Mariët Theune, Dolf Trieschnigg

University of Twente

19.02.2018

Abstract

In the following, a feedback interface for a computer based annotation task is developed with the goal of creating a better user experience around the task. By doing so, a consequence of the better experience should be more user engagement, resulting in higher productivity and more output from the workers assigned to the task.

Gamification is a modern design trend in user experience, defined in 2011 by Deterding as 'the use of game elements in non-game contexts'. Within this project, the annotation task and its context are examined, and an interface prototype was designed and realized using game elements intended on evoking a positive user experience - challenges, achievements, progress indication, a (work) data visualization and performance feedback. The original work process and software were not modified.

An experiment was conducted to evaluate the gamified feedback interface. Results indicate that the gamification was effective in improving user experience and motivation to some extent, as mean ratings for the gamified feedback were higher than those of the other two tested versions (non-gamified feedback and log file feedback). Test users appreciated the gamified feedback, noting that it was interesting to see and that they looked forward to their feedback during testing. Test users also commented on the motivational value of some of the implemented gamification elements, and that they felt more challenged when receiving feedback in the gamified format. However, statistically significant differences could only be distinguished in 5 out of 22 indicators for motivation, engagement and user satisfaction. This leads to the conclusion that future research can lead to further improvements for the gamified feedback system.

Keywords: Gamification, annotation, user experience, motivation, user interface

Abstract	1
Chapter 1 - Introduction	4
1.1 Motivation	4
1.1.1 Client MyDataFactory, problem statement	4
1.1.2 Gamification	4
1.2 Project outline, method of investigation	5
Chapter 2: Context analysis	6
2.1 The annotation task using brat annotation software	6
2.2 End user analysis	8
2.3 Context of annotation task	9
2.4 Problem description	9
2.5 Research question	9
2.6 MoSCoW requirements I: must have	10
Chapter 3: State of the art	11
3.1 Motivation within the Self Determination Theory	11
3.2 Requirements regarding motivation	13
3.3 Gamification	13
3.4 Gamification mechanics and dynamics	16
3.5 State of the art: related projects	17
3.6. MoSCoW requirements II: should have	21
Chapter 4: Ideation	22
4.1 Brief recap of essential aspects of the annotation task	22
4.2 Collection of possible gamification mechanics and dynamics	22
4.3. Selection of mechanics to be included in prototype	27
4.4 MoSCoW requirements: gamification M&D	28
4.5 System features	29
4.6 Reflection on fulfillment of gamification M&D MoSCoW requirements	30
4.7 Pen and paper prototyping	31
4.8 Results from pen-and-paper prototype testing	33
4.9 Adjustments for high-level development	34
Chapter 5: Specification	35
5.1. Visualization choice	35
5.2. Development of high-level design: usability testing	38
Chapter 6: Realization	41

Chapter 7: Evaluation experiment design	42
7.1 Test outline, variables and hypotheses	42
7.2 Questionnaire	46
7.3 Adjustments for test	49
Chapter 8: Evaluation results	51
8.1 Hypothesis on quantity of annotations made	51
8.2 Hypotheses on user engagement and motivation and usability	52
8.3 Other insights gained from test	56
8.4 Limitations of the test	58
8.5 Test conclusions	58
8.6 Reflection on MoSCoW requirements	59
Chapter 9: Discussion with client	61
9.1 Prototype discussion	61
9.2 Concept discussion	61
9.3 Recommendations for future work	66
Chapter 10: Conclusion	67
Chapter 11: Future work	68
References	70
Appendix	72
Pen and paper low-level questionnaire	73
Prototype test: all questions and answers	88
One-way ANOVA significance tests and follow-up Tukey tests	120

Chapter 1 - Introduction

1.1 Motivation

1.1.1 Client MyDataFactory, problem statement

MyDataFactory¹ (abbreviated MDF) is a small Dutch company which specialises in data cleansing and matching for clients with large databases. Their work consists of both providing intelligent software as a service (SAAS) as well as human-sourced data cleaning, correcting and matching activities.

Next to their client-related work activities, the company is creating their own dictionary of product descriptions in order to enhance their matching processes and cleansing tasks. The dictionary is created by performing an annotation task, which is done on a computer using an annotation software called brat² (brat rapid annotation tool). However, this dictionary is not directly linked to a particular client case and the work process involved in creating it is monotonous and repetitive. Thus, realization of it is slow and tedious, and employees struggle to maintain motivation. It is in the client's interest to propel the creation of this dictionary, ideally by motivating employees to work on it independently, without the current recurring requests of their superior for 'someone to work on it a bit'.

As stated by the client, the task is to (quote): "Entice the user to contribute many, and high-quality terms to the dictionary. How can a group of users be stimulated to contribute many terms to the dictionary (for instance by showing a dashboard with what colleagues contributed), and how can quality be managed (by cross-checking between users)". The focus of this work is on motivating users to voluntarily and regularly work on this dictionary.

1.1.2 Gamification

Gamification is an emerging trend in which game elements are applied to non-game contexts, such as the workplace, to increase employee job satisfaction as well as productivity. For obvious reasons, the gamification of repetitive and monotonous tasks is a popular choice both in and outside of the workplace, often occurring subconsciously. An example may be racing a coworker to see who can package more products for shipping in a certain amount of time.

Companies are becoming more receptive towards experimentation with gamification, enticed by the potential benefits such as increased employee satisfaction and productivity [3]. In one study [7] that collected employee's knowledge and opinions of gamification the majority of the asked

¹ www.mydatafactory.com

² <http://brat.nlplab.org/>

employees favoured the idea of gamifying certain tasks or processes, given that the context was appropriate and considerations were made. Many also described having used gamification in some aspect of life before, again frequently in the context of monotonous and repetitive tasks. Research such as that done by Jovanovic [8] shows that when properly implemented, gamification can hold true to its promises. However, as it is a fairly new field, the research on it is scattered and not clearly defined - previous experiments that use game mechanics are not labeled as gamification, and the contexts of application are many. The situation at hand offers itself to an implementation of gamification, as the task description (motivating workers to work, and work more, while increasing user satisfaction and productivity) fits the goals of gamification. Additionally, researching the topic will contribute to the knowledge base of this ambiguous young field, offering those who wish to implement gamification more scientific research on which they can justify design choices.

1.2 Project outline, method of investigation

The product development method used in this work is inspired by the Creative Technology design process, as described by Mader [9]. Through a context analysis, the task and its role in the workplace are examined to identify how the [activity of performing the] task can be improved. As a result, requirements are distinguished in the form of a MoSCoW prioritization that a solution must cater towards, and an appropriate research question is formulated. Following the context analysis, a state-of-the-art review sheds light on current thinking in the fields of motivation and gamification, offering applicable advice that can be used in the development of a prototype.

Based on the context analysis and recommendations from literature, gamification is applied to the annotation task, and a prototype is developed from a pen-and-paper prototype unto a functioning web application. This prototype is then tested with an experiment in its effectiveness in improving user experience and productivity around the annotation task. The resulting data is collected and evaluated in respect to the original research question, leading to a discussion on the strengths and weaknesses of the developed prototype. The success of the prototype in meeting the requirements is deemed, and a conclusion is met. Recommendations for future work are made in the end.

Chapter 2: Context analysis

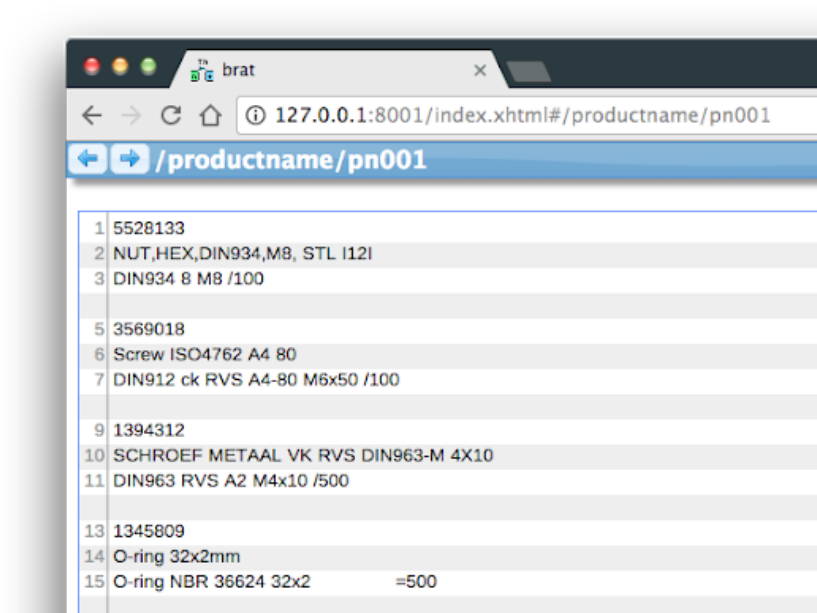
In the following chapter, the annotation task (with which the dictionary is created) in its current state is examined as well as the other relevant elements around it: the end users and their relation to the task, the tasks role in everyday company activities and the environment it is performed in. In understanding all these factors, the problem is defined more accurately and essential requirements can be outlined that any solution must aim at fulfilling.

2.1 The annotation task using brat annotation software

In the scope of this project, the activity of interest is the annotation of product names in a catalogue of unannotated product data. This catalogue consists of thousands of pages of product data, with each page having (on average) entries on 10 products, each entry consisting of a reference number and two individual product descriptions. A section of such a page can be seen in figure 1.

The annotation process is performed using a web application called brat³, in which three steps are taken to make an annotation:

1. Making a selection of entry content from the product descriptions by clicking and dragging the mouse over text,
2. choosing from a pre-made list of annotation labels (figure 2), and
3. annotating with the label (figure 3).



1	5528133
2	NUT,HEX,DIN934,M8, STL I12I
3	DIN934 8 M8 /100
4	
5	3569018
6	Screw ISO4762 A4 80
7	DIN912 ck RVS A4-80 M6x50 /100
8	
9	1394312
10	SCHROEF METAAL VK RVS DIN963-M 4X10
11	DIN963 RVS A2 M4x10 /500
12	
13	1345809
14	O-ring 32x2mm
15	O-ring NBR 36624 32x2 =500

Fig. 1. 4 entries from the example data set are displayed in the brat tool interface. Brat runs on internet technology, so here it is seen opened in Google's web browser Chrome. Enlarged image.

³Brat Rapid Annotation Tool, open source from <http://brat.nlplab.org/index.html>

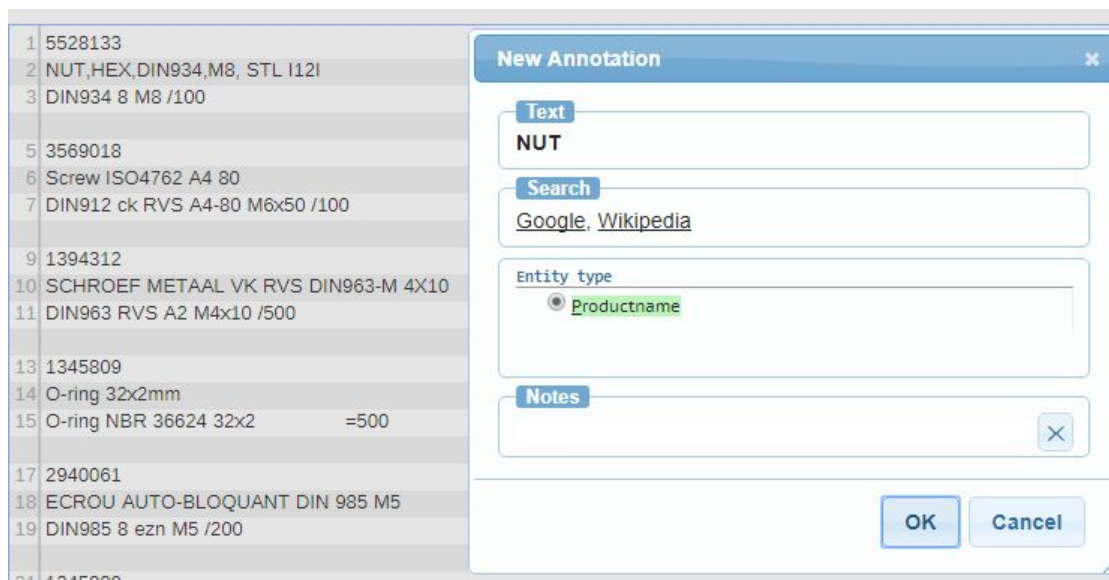


Fig. 2. A selection of characters from a data entry has been made (from the product description in line 2). When a section is highlighted, this window appears, allowing the user to choose from a list of predetermined labels. In the scope of this project, only one label (product name) is provided for annotation. Enlarged image.

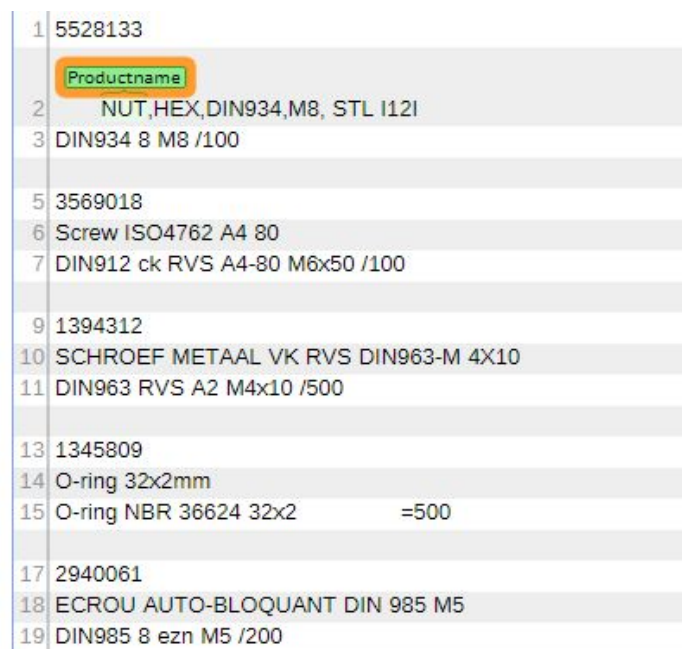


Fig. 3. Finished annotation. After the user clicks on 'OK' in the popup window (see figure 2), the label appears above the selected text. Enlarged image.

Ideally, on each page of 10 entries, a user will be able to identify and annotate at least 2 product names per entry, one in each product description (as each entry represents a single product,

the product names on each line should either be identical or interchangeable), resulting in 20 annotations of product names per page. As however some product names are indistinguishable or entries may be corrupted (e.g. have no product name or only one product description), this is not always possible. In some cases, multiple product names may be in a single product description (such as an industrial DIN standard as well as a text description).

When a user makes an annotation, the annotation is stored in a separate .ann file, along with all other annotations made on that page. An example of this text file can be seen in figure 4. This is the only way for anybody, the workers and their superiors, to see what work has been done (besides opening the page in the brat software).

During the annotation task, users will annotate new and recurring product names - some product names will be very frequent, often recurring multiple times on a single page and some may be very rare, appearing only once throughout a whole session or even in the whole database.



Fig. 4 Screenshot of the .ann file resulting from the annotation made in figures 1-3, opened with the Windows Notepad application. The annotation file contains the chronological listing of the annotations (T1 being the first, T2 the second etc.), the category annotated, the character span in the original file and the characters themselves.

2.2 End user analysis

The end users of the system are MyDataFactory employees (a group of domain specialists carrying out data cleansing activities), working in the office environment of the company. As part of a small company, the interpersonal relationships at the workplace are informal. The company tries to avoid bureaucratic structures and prefers common sense over imposed rules, wherever possible.

These workers have responsibilities and tasks of their own, often linked to client cases that are time sensitive. As the success of the company depends on meeting demands of their customers, the workers can only afford to contribute to this annotation task whenever there is nothing else more urgent, often only for an hour per day and in irregular intervals.

As the workers perform the task, they are not receiving any feedback or indication on the amount of work they have performed, how many (new) annotations they have made for the dictionary or which section of the product database they have annotated. The software itself does not offer any form of feedback besides the .ann files produced from annotating. This has been expressed by a user and the client as a demotivating factor - workers pour time and energy into the annotation work, and have no indication in how far they got or how valuable their

contributions are. Further, they have no way of showing other people (for example the person who asks them to do the work, their boss or visiting clients) the work they have done.

2.3 Context of annotation task

The annotation task is performed to create a database of annotated data that can be used for machine learning systems. As such, these systems are intended on gradually increasing the amount of products recognized and the accuracy with which the companies' software can identify products. However, the database of unannotated data is vast, consisting of thousands of pages, and there is no noticeable software improvement from one annotation session to the next.

This annotation task is an ongoing process with practically infinite amounts of data to be annotated. As mentioned, the annotation task is not linked to any immediate or pressing projects, such as client work with real deadlines and tangible deliverables. As such, it is an extracurricular task that workers are asked to participate in whenever they can afford to between their regular obligations. This frequently occurs due to reminders from their superior, upon which the task is performed for a couple days before being dropped again. Thus, the main motivator to work on the task is to comply with these requests, as the work itself is monotonous, repetitive and not inherently rewarding.

The annotation task, as all other work performed by the company is performed on a company PC in a quiet office space shared with other employees of Mydatafactory. Currently, it is only performed within normal working hours, and as mentioned, only when circumstances permit and when the workers either remember to do the task or are asked to.

2.4 Problem description

From the context analysis, the challenges around the annotation task can be seen: workers dedicate the little time they have inbetween obligations to a repetitive, monotonous task that offers no reward and no indication of progress. The work is extrinsically motivated by requests from above [in the company hierarchy], and performance or contribution can not be acknowledged. It is in the client's interest to increase motivation, and sequentially participation, in performing the annotation task, and offering the workers who perform it a better, more rewarding experience around the task.

2.5 Research question

The following research question was formulated to guide the research and method of investigation:

Can the annotation task be enhanced with gamification?

2.6 MoSCoW requirements I: must have

Based on the task at hand, namely addressing the lack of feedback and motivation around the annotation task, must-have requirements in the MoSCoW (Must have, Should have, Could have, Won't have) prioritization hierarchy are established :

Must have:

The product must motivate the workers to perform (and keep performing) the annotation task.

The product must increase user satisfaction around the annotation task.

The product must increase the amount of annotations workers contribute.

The product must increase (or maintain) a high quality in the annotations the workers contribute.

Chapter 3: State of the art

The task at hand centers around motivating people (workers) to perform activities they are not inherently motivated to do. Thus, gaining an understanding of motivation and various motivational factors is crucial to a successful implementation. Gamification, or 'the use of game elements in non-game contexts' [3], is an approach gaining momentum in the industry of user experience, user engagement and software/interface design. It is well suited for the task at hand, as it is in the client's interest to create a stimulating experience and entice users to contribute frequently to the to-be-created dictionary. In this chapter relevant research on motivation and gamification is presented which will help in making an effective product.

3.1 Motivation within the Self Determination Theory

Within the field of psychology, the Self Determination Theory (SDT) of Deci and Ryan is a generally accepted and leading framework of theories on human motivation. While there are other theories and frameworks of behaviour and motivation, Deci and Ryan's is considered appropriate for the situation at hand, explaining motivation and the factors necessary to evoke it in related contexts.

According to Deci and Ryan, people are not only motivated to different degrees, but also by different types of factors, which can be grouped into intrinsic and extrinsic motivators.

Furthermore, they distinguish three core needs that are at the center of self-motivation: the need for competence, relatedness and autonomy [1].

Ryan and Deci describe motivation on a spectrum, ranging from amotivation, the 'state of lacking the intention to act' resulting from 'not valuing an activity', 'not feeling competent' or 'not expecting it to yield a desired outcome' [1] to intrinsic motivation, the doing of an activity for inherent satisfactions. In between are different degrees of extrinsic motivation. This spectrum can be seen in figure 5.

The Self-Determination Continuum Showing Types of Motivation With Their Regulatory Styles, Loci of Causality, and Corresponding Processes

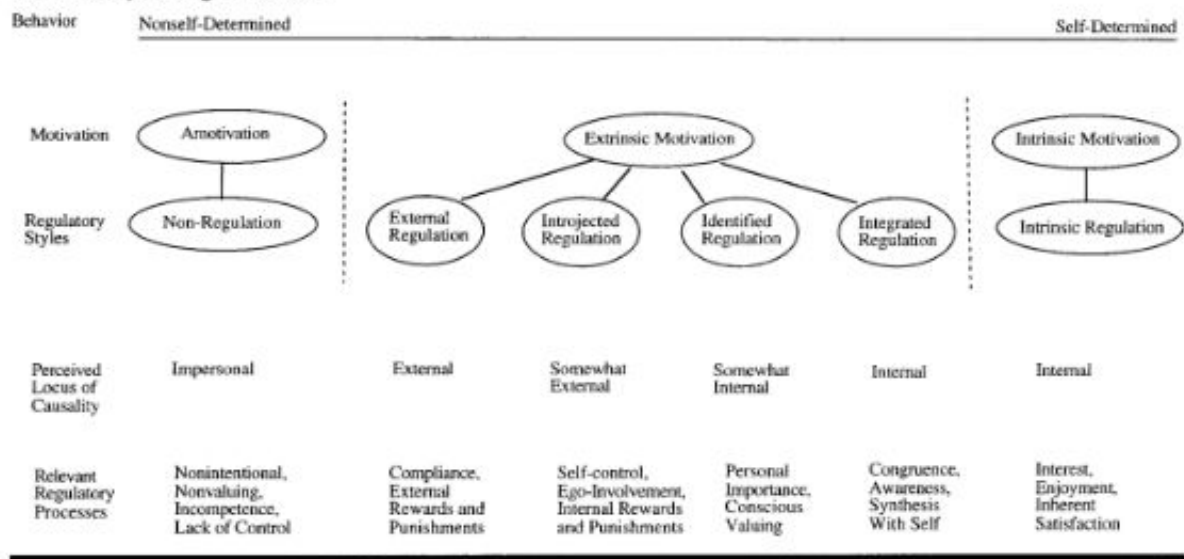


Figure 5. The self-determination continuum showing types of motivation with their regulatory styles, loci of causality and corresponding processes. Taken from Ryan & Deci [1].

Extrinsic motivation is shown to vary in terms of perceived locus of causality (the source of motivation as the person perceives it), between external and internal (to the person) and in terms of the accompanying behavioural processes that reflect the person's attitude towards the object of motivation (bottom row figure 5, relevant regulatory processes).

Intrinsic motivation refers to doing something because it is inherently interesting or enjoyable, whereas extrinsic motivation refers to doing something because it leads to a separable outcome. The closer to intrinsic motivation a form of motivation is on the scale, the greater the person's persistence, positive self-perceptions and quality of engagement with the activity [2]. As previously mentioned, competence, relatedness and autonomy are critical to self determination and intrinsic motivation, and can be seen in varying measures in the regulatory processes of extrinsic motivation. Ryan and Deci recommend creating contexts that support these three factors to increase internalisation and integration and evoke commitment, effort and high-quality performance in the activities and goals they pursue, and they warn against excessive control, nonoptimal challenges and a lack of connectedness [1].

3.2 Requirements regarding motivation

From this, we draw requirements that should be met to support motivation. Based on these requirements, considerations regarding how they can be supported are suggested:

- The product should aim to support competence, relatedness and autonomy.
 - Competence can be supported by keeping the product as close as possible to the original format in which it is performed, minimizing adaptation costs for the workers.
 - Relatedness can be supported by creating a product that offers feedback as close and relevant as possible to the real performance of the workers.
 - Autonomy can be supported by respecting the context in which the work is performed (a busy schedule and quiet workplace), and allowing workers to decide when it is most appropriate for them to work on the annotation task, essentially placing it in their hands. The product should not interfere with their usual obligations or anything else workplace related.
- The product should aim to shift motivation from externally regulated extrinsic motivation to intrinsically regulated intrinsic motivation.
 - This can be supported by creating a product that makes use of motivators based around the task, and creating challenges and rewards that the workers can relate to.

3.3 Gamification

While the term is popularly used in an ambiguous manner and is a topic of debate, researchers often reference Deterding et al. who defines it as ‘the use of game design elements in non-game contexts’ [3]. In detail, they write ‘Gamification refers to:

The use of (rather than the extension) **design** (rather than game-based technology or other game related practices) **elements** (rather than full-fledged games) **characteristic for games** (rather than play or playfulness) **in non-game contexts** (regardless of specific usage intentions, contexts, or media of implementation). Gamification in this form has been applied in various contexts and is, as Bunchball describes below, often applied to existing websites or applications. In the following, some examples of gamification are listed, and the elements, as defined by Thiebes et al. (see section 3.4) used in them are highlighted.

Examples of popular gamified services include KhanAcademy⁴, an online learning platform in which users can earn points and badges that are displayed on their public profile (see figure 6) and the smartphone application/game called Zombies, run!⁵ in which users who want to improve

⁴ www.khanacademy.org

⁵ <https://zombiesrungame.com/>

their running performances are offered a storyline experience. In *Zombies, run!*, users take the role of one of the last survivors on earth after a zombie epidemic and must save mankind by running better. As users run, the story unfolds, distracting from the otherwise (for some users) repetitive, monotonous or strenuous activity of running. In figure 7, one can see a screenshot from the interface.



Fig. 6, screenshot⁶ from Khanacademy user center. One can see earned achievements and progress indicators, both elements of gamification.

⁶ Taken from

<https://62e528761d0685343e1c-f3d1b99a743ffa4142d9d7f1978d9686.ssl.cf2.rackcdn.com/files/75220/area14mp/image-20150318-2490-3vpbnh.png>



Fig. 7 Image⁷ from *Zombies, run!* Game. Use of fantasy and feedback are also gamification elements.

By using game elements in these otherwise non-gamified contexts, designers aim to motivate desired behaviours and drive engagement [5].

Bunchball, a successful gamification company with over 300 clients, defines gamification as 'the process of taking something that already exists - a website, an enterprise application, an online community - and integrating game mechanics into it to motivate participation, engagement and loyalty' [4]. Stackoverflow is an example of a gamified online community platform: it is a forum/discussion site on which programmers and developers can ask and answer questions, and participate in discussions. In figure 8, one can see that reputation, another gamification element has been applied. By answering questions correctly and receiving according feedback from the community, users can earn themselves reputation in form of points, and acquire powerful abilities such as editing, deleting or moving posts from other users.

⁷ Taken from <http://www.thecollector.com/wp-content/uploads/2014/06/sites.jpg>

Type to find users:



Jon Skeet
Reading, United Kingdom
420k • 123 • 1818 • 3121
c#, java, .net



Darin Dimitrov
Rouen, France
299k • 18 • 333 • 543
c#, asp.net-mvc, asp.net-mvc-3



Marc G Forest
297k •
c#, .ne



Hans Passant
Madison, WI
237k • 16 • 102 • 270
c#, .net, winforms



SLaks
New Jersey
212k • 15 • 286 • 492
c#, .net, javascript



VonC
France
204k •
git, ecl

Fig. 8, top contributors on the stackoverflow discussion platform⁸. Top contributors have high point amounts, and thus gain social recognition. Reputation and points are gamification elements.

Both definitions of gamification seem appropriate given the situation at hand - namely, increasing user engagement with the to-be-processed database, and pushing the completion of the task forward while motivating the employees.

3.4 Gamification mechanics and dynamics

In 2014, Thiebes et al. produced a comprehensive overview of game mechanics and dynamics (M&D) described in recent studies, which they clustered into five categories which designers of gamified systems should all consider when gamifying an information system [6].

Game mechanics are described as “functional components of a gamified application and provide various actions, behaviours and control mechanisms to enable user inter-Action” [10]. Examples of these might be point systems or leaderboards. Dynamics, on the other hand, determine the individual’s reactions as a response to using the implemented mechanics [6]. Their summarised mechanics and dynamics (in future M&D) can be seen in table 1 (below). In their synthesis Thiebes et al. only selected studies that had focused on empirically investigating the effectiveness of each mechanic/dynamic, and where the workplace (in contrast to education or health) had been the study context. The game mechanics & dynamics were derived as isolated, individually investigated factors implemented in gamification experiments, with the intent of improving user motivation and productivity. This makes the synthesis a

⁸ Taken from <http://vonbismark.com/wp-content/uploads/2012/06/Stackoverflow2.png>

suitable catalogue of elements from which appropriate elements can be implemented in a prototype.

Cluster	Categorised game mechanics & dynamics
System design	Feedback; Audible feedback; Reminders; Meaning; Interaction concepts; Visual resemblance to existing games; Fantasy
Challenges	Goals; Time pressure; Progressive disclosure
Rewards	Ownership; Achievement; Point system; Badges; Bonus; Loss aversion
Social influences	Status; Collaboration; Reputation; Competition; Envy; Shadowing; Social facilitation; Conforming behaviour; Leaderboards; Altruism; Virtual goods
User specifics	User levels; Ideological incentives; Virtual characters; Self-expression

Table 1. Clusters of game mechanics & dynamics according to S. Thiebes et al [6].

The M&D are divided into clusters with regard to their meaning and method of motivating users/evoking certain behaviours. Each M&D represents a way of motivating users by using the named mechanic or dynamic in a gamification setting. Thiebes et al. recommend selecting M&D appropriate for the respective context of the system based on a context analysis that considers the task, the workers and goals of both. They advise against a ‘one solution fits all’ approach, noting that the wrong M&D in the wrong context can have detrimental effects.

3.5 State of the art: related projects

Upon researching the application of gamification to word sense labeling, annotation work and language notation, some examples were found that are worth examining. While none of the examples are specifically designed for the workplace or (industrial) product name annotation, they nonetheless exemplify successful implementations of game elements in the context of word annotation and language resource creation.

Crowdsourcing complex language resources: playing to annotate dependency syntax [11]:

In an attempt to harness the power of crowdsourcing to create databases of high-quality, manually annotated text bodies, Guillaume et al. created a Game with a Purpose called ZombiLingo [11]. In contrast to gamification, which makes use of game elements in non-game contexts but doesn’t necessarily change the way an existing task is performed, through a Game

with a Purpose users create the desired data (in this case the annotated texts) by playing a (newly created) game. A screenshot from the final game can be seen in figure 9.

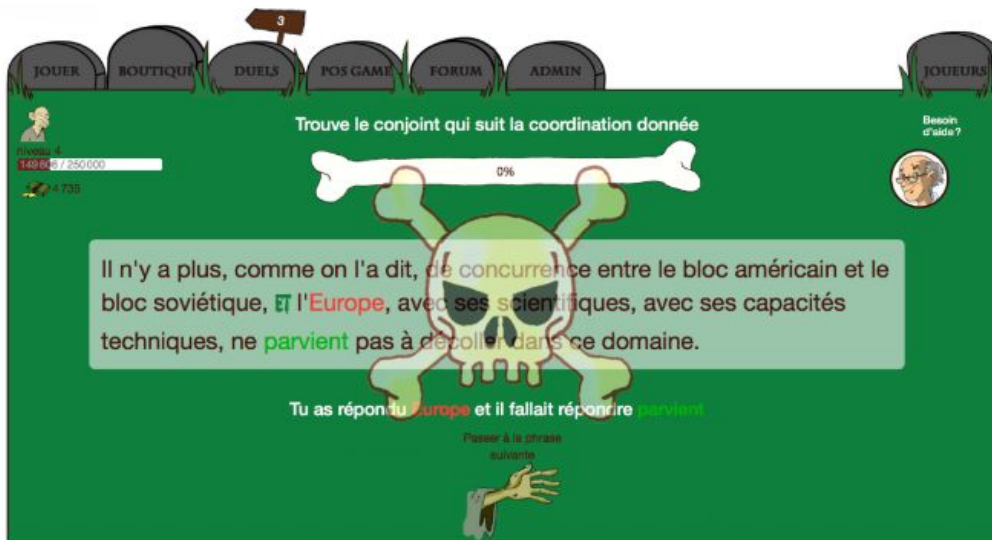


Fig. 9. Main interface of the game during the training phase [11].

The purpose of this game was to recruit as many participants as possible, and entice them to perform annotation work on (french) text bodies by framing the activity in an engaging way. The experiment proved successful both in training participants to annotate with high quality as well as creating a database of cross-corrected data, but, being a participant-driven task it was dependent on constant communication with the players and planned events to motivate and maintain participation.

While this work successfully enticed many users to participate and ultimately created valuable data, this concept (and theme) can not be applied to the situation at hand. The workplace is not an appropriate setting for a zombie-themed game to annotate product names for industrial clients.

Phrase Detectives: A Web-based Collaborative Annotation Game [12] :

In a similar effort to that of Guillaume et al. to create a database of annotated language data large enough to train and evaluate intelligent annotation software, Chamberlain et al. created an interface using game-styled elements with which non-expert users can participate in annotating and validating the work of other annotators. One can see a screenshot of the interface in figure 10.

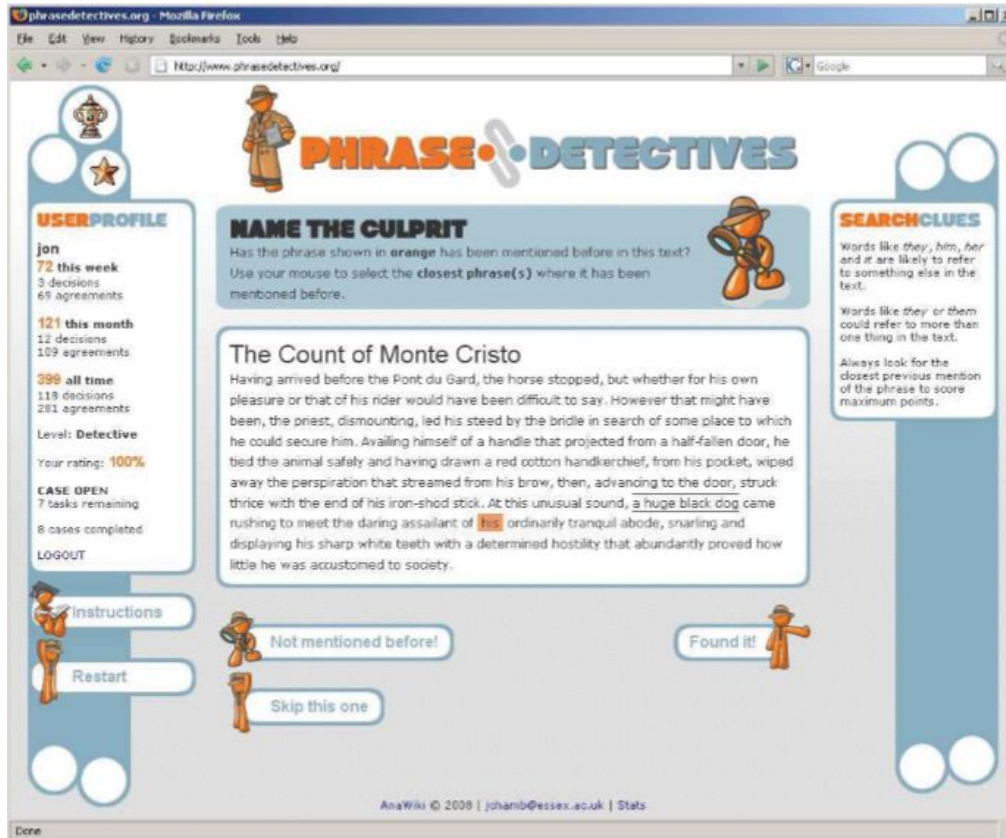


Fig. 10 Screenshot from the annotation mode in which a user is given a text in which they must make annotations. One can see a user profile on the left hand side, in which feedback and challenges are shown.

In this work, users are motivated using comparative and collaborative scoring, and leaderboards. Upon testing, all users who also use Facebook (a social media platform) said that they would be motivated to play if the game were integrated in their profile. This as well as the comparative and collaborative scoring indicate that amongst the test users (university staff and students), a high desire to integrate social elements was present.

The built in quality control used here is relevant to the task at hand - the client asked for a way to do peer-reviewing for quality control of the annotation work. The applied game elements may inspire a similar mechanism in this or future work on gamifying the product name annotation.

Gamification for Word Sense Labeling [13]:

Venhuizen et al. developed a collection of games with a purpose called Wordrobe with the goal of expanding a database of language annotations by enticing users with games using multiple-choice questions. An example can be seen in figure 11.

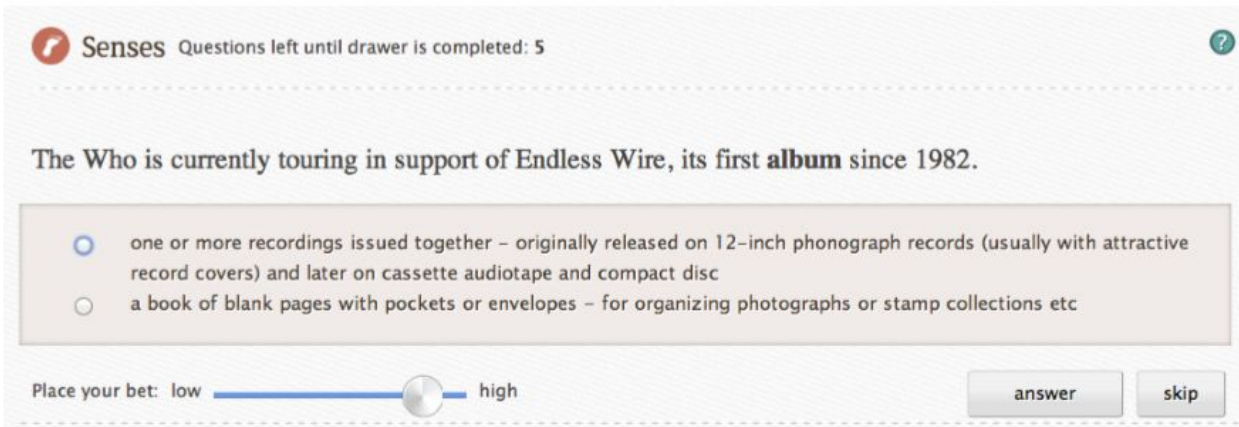


Fig. 11. Screenshot from the Wordrobe game.

Similarly to Phrase Detectives, in this game with a purpose, attaining high quality is a focus of the research. By placing bets on their correctness (which users make based on their confidence in being correct), users can, similarly to sports bets, score higher points when their answers are deemed correct. Answers are deemed correct depending on agreement amongst participants, as there is no gold standard to which the answers can be compared.

A high amount of precision was obtained in the scope of this work, but certain questions evoked unanimity amongst users that was different than what the test standards had been defined as. This lead to the conclusion that, at least in language annotation, more precise questions and available answers and a wider range of quality control are necessary to catch exceptions such as the above, specifically when collecting data from non-expert participants.

Concluding the research on the annotation of product names (so not of linguistic resources, such as book or website texts), no gamification of a similar annotation task has been done in the past, at least not in the scope of an academically written and peer reviewed study. Thus, the research into how gamification can serve to motivate participants in this context can be deemed original research.

3.6. MoSCoW requirements II: should have

Incorporating the recommendations on motivation and the recommendations by Thiebes et al. on proper gamification, should-have requirements are added:

- **Must have:**
 - The product must motivate the workers to perform (and continue performing) the annotation task.
 - The product must increase user satisfaction around the annotation task.
 - The product must increase the amount of annotations workers contribute.
- **Should have:**
 - The product should be integratable with the existing annotation software, as to reduce adaptation costs and thus maintain competence with the existing task.
 - The product should offer feedback as close and relevant as possible to the real performance of the workers, to keep any kind of intervention relatable to performance and the task.
 - The product should not be invasive or demanding, and allow workers to autonomously decide when and how they contribute to the dictionary with annotation work.
 - The product should use appropriate gamification elements based on the context analysis of the task, the workers, and the goals of both.
 - The product should not use gamification elements that are not relatable to the annotation task or inappropriate for the workplace or context in which the task is performed.

Chapter 4: Ideation

In the following chapter the theory on motivation and gamification is applied to the annotation task. Game M&D are chosen which are relevant to the annotation task, context analysis and established requirements. After taking consideration for which M&D can feasibly be implemented given the scope of the project, a low-level pen-and-paper interface prototype is developed which realizes the chosen gamification elements. This is then tested for essential usability aspects, in preparation for high-level development.

4.1 Brief recap of essential aspects of the annotation task

To guide the selection of gamification M&D potentially applicable to the annotation task, some essential aspects of and around the annotation task are recapped:

- In the annotation task, users will annotate pages upon pages of product names - some will be new, some will be recurring instances of existing product names.
- There is currently no form of feedback or visual representation of any of the work done (other than opening the annotation files), neither of the annotation work nor of the quantity of work done by the workers (in any given session or in total).
- There are no clear goals other than to 'do the work' - how much should be done is not defined.
- The current motivator to perform the annotation task is to comply with requests from the company/the workers' boss. Besides this, there are no challenges or rewards that would entice users to pick up the task on their own.
- The annotation work is performed inbetween other tasks, whenever the workload affords it.
- The annotation task is performed in a quiet, shared office environment, by domain specialists.

4.2 Collection of possible gamification mechanics and dynamics

Based on the context analysis and main aspects reiterated in section 4.1, the list of gamification M&D analysed by Thiebes is scanned for feasible elements. In the following table 2, the 31 elements are listed and described. If an element is deemed potentially applicable, an implementation is listed. If a M&D was not appropriate for the situation at hand, the exclusion criteria is given and it is excluded from further research.

Gamification M&D & description	Application to brat product name annotation work in Mydatafactory workplace	Exclusion criteria
Category: System design		
1. Feedback Give players awareness of their progress and/or failures in real time, e.g. with a progress bar and a color indication of right or wrong entries	Number of annotations done (in current session, in total) Number of sheets that are completely annotated (and which still have entries without an annotation in each description line, so that need further processing)	Live feedback is too obtrusive; end of session feedback should be sufficient
2. Audible feedback Sound effects and/or music		Workplace is (quiet) office environment, unnecessary if visual feedback is already present
3. Reminder Remind a user of their past performance	Progress bar showing how much a user has contributed, possibly over (how much) time	
4. Meaning Use the background of the user and the contextual placement of the task to give it meaning	Feedback on annotation content: How many different annotations has a user contributed to the system?	
5. Interaction concepts Attractive user interface, interaction and visually stimulating elements	Appealing visual design with mixture of text and graphic elements	
6. Visually resembling existing games Resemble existing games, e.g. Tetris, for familiarity		Inappropriate for workplace context
7. Fantasy Emotionally enhance the user experience with elements of fantasy		Inappropriate for workplace context
Category: Challenges		
8. Goals	Implementing daily, weekly or	

Create appropriate challenges and goals for users	<p>monthly goals of how many entries or pages should be annotated</p> <p>Communal goals that the work team can/should accomplish together</p>	
9. Time pressure Create time pressure with countdowns or similar time based mechanisms		Sessions can range from minutes to hours and are generally open ended - the longer a user works, the more they contribute. Further, difficult entries may take longer to resolve, and rushing workers may compromise annotation quality
10. Progressive disclosure Help players increase their skill by gradually disclosing knowledge and challenges	'Database' visualization of the different DIN and ISO standards that a user has contributed (and possibly which ones are missing)	The single annotation activity has the same degree of difficulty throughout the whole project
11. Ownership Users have a positive, sustained feeling of ownership towards their work	Feedback on collective contribution of a worker, e.g. how much of the total annotation work they have contributed, how many different product names they have annotated may evoke a sense of ownership	
12. Achievement Reward users for completing clear and desirable goals	<p>Additional feedback when milestone amounts of annotations, e.g. 100, 200 or 500 total, or different, have been achieved</p> <p>Additional feedback when a sheet needs no more work (so has all necessary annotations)</p>	
13. Point system Reward points for completing actions; points are cumulative and rewarding follows a system		Feedback on the annotation work is already numeric and incremental: points would be redundant

14. Badges Optional rewards and goals rewarded for participation outside of the main activities of the work process		The work consists of singular activity - there is no additional work in the scope of this that could be encouraged and rewarded (future expansions of work may have such however)
15. Bonus Extra reward for accomplishing a series of challenges or core functions	Additional feedback and reward, same as achievement	
16. Loss aversion Influence behaviour by making users lose something if they e.g. don't perform regularly or consistently, create something worth keeping by maintaining performance	Achievements or other elements that reward and encourage frequent and regularly occurring work, and that are removed/reset when not maintained	

Table (continued)

Gamification M&D & description	Application to brat product name annotation work in Mydatafactory workplace	Potential exclusion criteria
Category: Social influences	General: public visualization	
17. Status Status in a social environment, earned by working in isolation [in contrast to in a group task]	Public visualization of each or top users' contributions: who has done the most annotations? Who has done the most complete pages?	
18. Collaboration Create opportunities for colleagues to help each other on a set of tasks or large challenge	Public visualization of communal goals to annotate certain amount of entries in a week/month and how far the team is in accomplishing them	
19. Reputation The reputation of a user reflects what other users think about that person's performance and contribution	As status, a public indication of noteworthy contribution efforts (e.g. most annotations) can give a worker reputation	
20. Competition Users are given the chance to challenge each other		Participation in system is voluntary, and longevity of task doesn't lend itself to competition - also, users participate as their individual schedules

		allow, making the competition ground unbalanced
21. Envy Create elements a user can have/earn and that makes other users want to earn it too	Can arise from public visualization of top workers' contributions	
22. Shadowing Competition with one's own previous performances	Show users their past session achievements (e.g. 200 annotations, 16/20 pages completely annotated), and inform them when they have outdone themselves	
23. Social facilitation Create a social environment that makes performing (simple) tasks and/or collaboration easier for individuals		Straightforward, singular activity doesn't offer leeway for improvement in this manner
24. Conforming behaviour Also called peer pressure, users adapt to the behaviour of the majority of other users	Public visualization may have effect of conforming behaviour - the more some users work, and this is made public, the more other users might feel compelled to contribute more themselves	
25. Leaderboards Leaderboards rank (top) users according to predetermined criteria, indicating who is 'performing the best' and are intended on evoking productive competition in desired behaviours	As status, the implementation of a publicly visible list of the top contributors (in different categories, such as most annotations, most complete pages) can challenge lower performing users to improve their ranking and thus reputation	
26. Altruism Users can gift each other (virtual) gifts to strengthen relationships		Not suitable for this work
27. Virtual goods Non-physical, intangible goods that can be bought, traded or otherwise exchanged amongst users		Not suitable for this work
Category: User Specifics		
28. User levels Levels show a users general skill level and proficiency in the desired task	Users can increase their level with the amount of annotations they contribute	May be redundant to feedback

29. Ideological incentives Use attitudes and values to evoke motivation	By showing a user which annotations (so also DINs and ISOs) they have contributed that they are improving their expertise and knowledge	Not suitable for this work
30. Virtual character Use virtual/fictional characters to represent participants	See self expression	Not suitable for this work
31. Self expression Let users exhibit some degree of self-expression or personality while participating in the gamified task	User info page, which reflects a users' general use profile? Can show frequency of sessions, length of sessions, total annotations etc. (makes sense outside of a single user system, as it would otherwise be redundant)	Not suitable for this work

Table 2. Table of M&D as described by Thiebes et al. and application or exclusion criteria.

4.3. Selection of mechanics to be included in prototype

In the scope of this project, there are limiting factors that exclude the testing and evaluation of various M&D. These factors are:

- Unavailability of professional users for testing
 - Can not test in or simulate social environment of workplace → this is too subjective to try to draw conclusions from a test with pseudo-users
 - Test should thus focus on the single user experience and not hypothesize for a social context
 - Testing will be done with pseudo users, in a testing scenario → can not investigate long term effects, goals or game elements, and should choose M&D that can be evaluated in a single test run

Based on these, the following M&D were excluded from testing:

- All M&D of the social influences category
- Reminders, as it relies on past work
- Bonuses, as they imply completing a series of challenges (which is not meaningful within the scope of a single-time test)

While the excluded M&D's potential value should not be ignored, their proper implementation and adjustment through testing and feedback must wait until it can be executed in the real workplace, with the real users. Further, as some of the elements build on past performances and performance over time, these are equally unsuited for testing with non-users who are unlikely to voluntarily commit more time and effort to doing annotation work than required.

4.4 MoSCoW requirements: gamification M&D

Based on the aspects listed in section 4.1 listed aspects around the annotation task, derived from the context analysis, a MoSCoW prioritization is performed to prioritize the most potentially helpful features. The M&D which can not be tested (the social influences M&D, reminders and bonuses) as well as the M&D deemed inappropriate for the workplace and context of the annotation task (audible feedback, visual resemblance to existing games, fantasy, time pressure, point system, badges, virtual characters and self expression) are listed under the 'won't have' features.

Must have:

Feedback (was requested; whole system is a feedback system)

Goals (was requested, can serve to replace external motivator of compliance with requests)

Interaction concepts (offering a visualization of the annotation database was requested and may be more stimulating than a text list)

Should have:

Ownership (could have motivational benefits, if properly evoked)

Achievement (could have motivational benefits and increase user experience with rewards)

Loss aversion (may increase productivity around the annotation task, if properly evoked)

Could have:

Meaning (as a dynamic, meaning can be inferred by representing how much a user has contributed, how significant their contribution is)

Progressive disclosure (can be realized through a visual representation of the annotation work)

User levels (can be implemented in combination with goals or achievements)

Won't have:

Social influences M&D (can not be tested)

Reminders (rely on past use, which can not be)

Bonuses (relies on more experience/work than can be done within project scope)

Audible feedback (inappropriate for workplace)

Virtual resemblance to existing games (inappropriate for workplace)

Fantasy (inappropriate for workplace)

Time pressure (inappropriate for context in which work is performed)

Point system (redundant to counting mechanisms)

Badges (relies on more experience/work than can be done within project scope)

Virtual characters (inappropriate for workplace)
Self expression (inappropriate for workplace)

4.5 System features

After filtering through the limitations of the project scope and selecting elements that can be potentially tested, evaluated and incorporated in a final prototype, a wireframe mockup of an interface design was created that included all the potential mechanics & dynamics. This can be seen in figure 12.

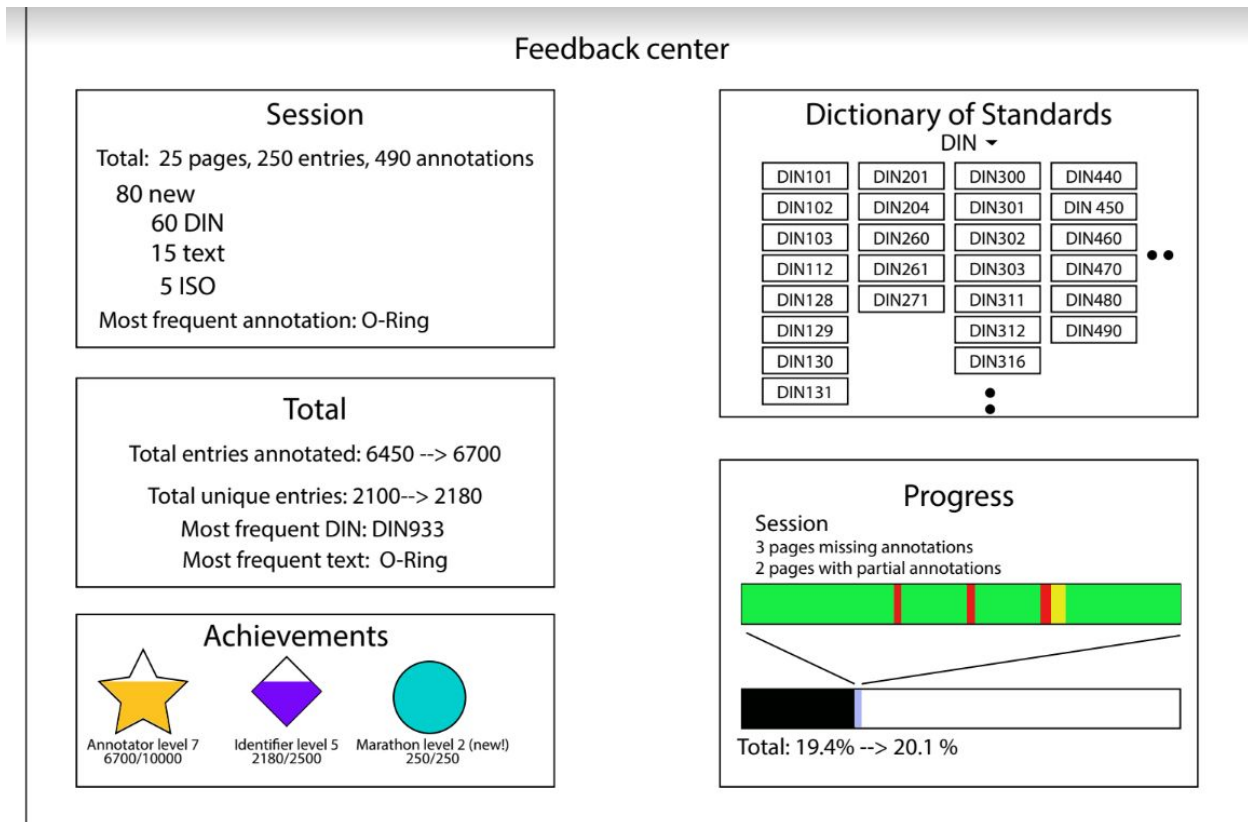


Figure 12. Wireframe mockup of gamified system

For a pen-and-paper evaluation, the features were grouped according to function and data that they give feedback on. These features are defined as such:

Session feedback

This section gives users feedback on their most recent work session, meaning the work they produced from opening the program to closing it. It informs on how many annotations they did (and of which type), and how many new (i.e. previously unannotated product names) annotations they made. The most frequent annotation is offered as a fun-fact.

Total feedback

This section gives users an overview in numbers on their total contributions to the annotation task, in terms of bulk (total annotations) and variety (number of different annotations). As DIN standards and text descriptions are much more frequently occurring than product names in ISO standard, these are offered as fun-facts additional to the numbers.

Achievements

Achievements can be attained with milestone amounts of work, and users can see how these visual symbols fill up in relation to the amount of work completed. When an achievement is reached, this can be added to for example a users profile or announcement board, which can function as a collection board for the various achievements a user can strive to acquire. Examples of these might also be: how often a user has worked on the project, how regular (without taking days off) their participation is, how many annotations they have done in a single session etc.

Dictionary of Standards

As product names are often given in their industrial standard, the whole database will have a vast amount of these that can be 'collected' and visualized as a form of dictionary of the standards that the user has 'discovered' in their work. Whenever a user has found new standards in their latest session, these can be visually highlighted, emphasizing their novelty to a users' growing collection.

Progress

In this section, the thoroughness of a users' annotation work of the last session is visualized in the top bar. This is then shown in relation to the total completion of the project (or a predetermined smaller milestone, if the project is too large to show how single sessions contribute).

4.6 Reflection on fulfillment of gamification M&D MoSCoW requirements

As for the prototype a prioritization of features (gamification M&D) was outlined. These are reviewed in their implementation here:

Feedback - the whole system is a feedback system, intended to give the user the opportunity to reflect on their progress and see their contributions.

Meaning - with the dictionary, as well as progress bar, users are supposed to be given a sense of value and significance that their contributions have: visually quantifying their work can show them what they are building; it functions as a form of visual feedback.

Interaction concepts - the system design should be aesthetically pleasing yet not distracting.

Progressive disclosure - the amount of work and dedication needed to earn achievements increases, and the dictionary grows with new entries.

Goals - challenging goals related to the annotation task can be seen in the 'achievements' section. The selected challenges/goals build on two content-related qualities of the work: the amount of work and the variety of product names. Further goals relating to additionally desired annotations, such as materials or manufacturers, can be implemented in later versions.

Ownership - with the personal dictionary of entries growing as well as achievements that reflect a users' personal milestones, these features aim to evoke a sense of ownership over the contributed body of work.

Achievement - Achievements can be earned and collected by working more, and aim to evoke satisfaction by informing a user (and potentially other users in future, socially dynamic systems, thus evoking social recognition) when they have contributed milestone amounts of work to the project.

Loss aversion - While not yet implemented in this wireframe mockup, loss aversion is in later versions linked to an achievement called 'streak', in which the number of sequential days doing annotation work are counted - if a user interrupts these, their progress on the achievement is reset. This may lead to more regular participation on the database, and thus result in more work done.

User Levels - Workers can increase their personal achievement levels with more work. The higher a level, the more work is required to advance. In future versions, users can have a user level which increases when e.g. all achievements have reached that level as well.

4.7 Pen and paper prototyping

In order to evaluate the functionality, insightfulness and initial reception of the various realizations of the game M&D, a pen-and-paper prototype was created and tested with three pseudo users. Pseudo users were male and female students of the University of Twente. The format of the test can be seen in figure 13. In figure 14, one can see the pen-and-paper prototype that was used to investigate the various elements.

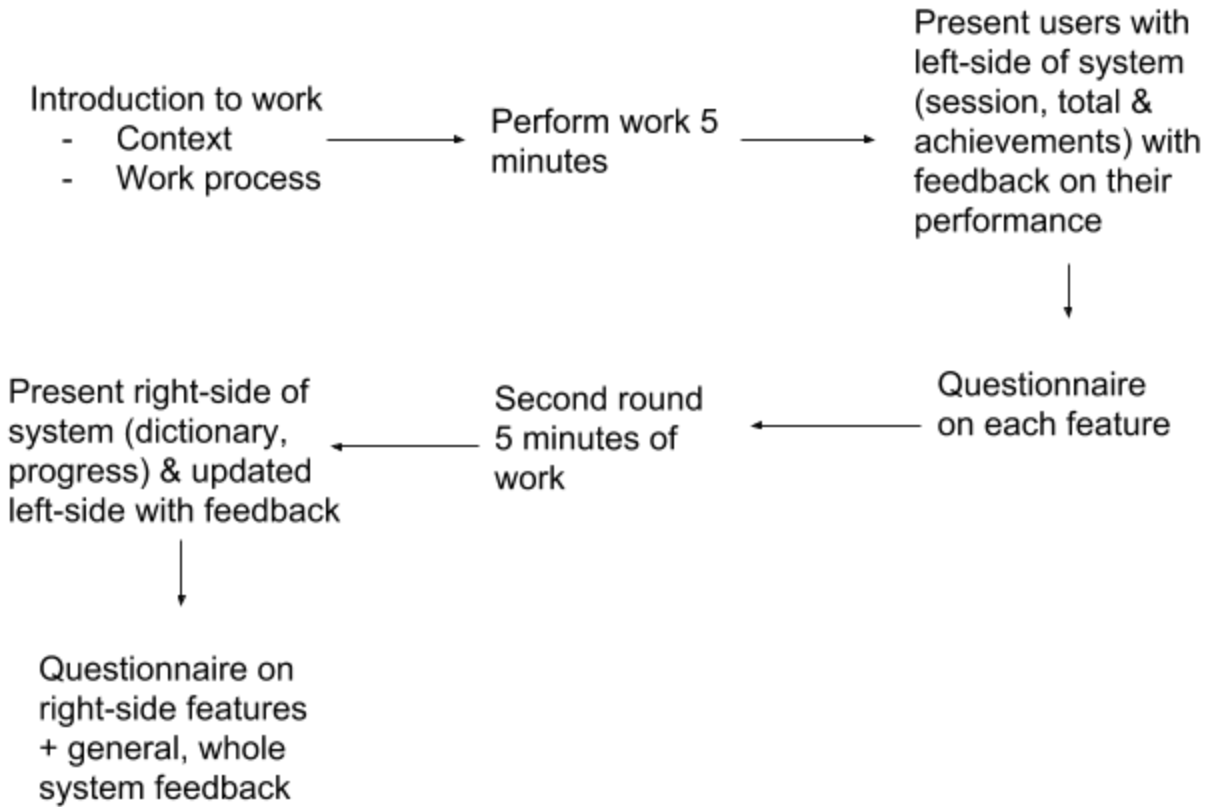


Figure 13, format of pen-and-paper testing & user feedback collection

The questionnaire and feedback can be found in the Appendix. Users were first given only half of the feedback to focus their answers on the left sides elements. Additionally, this allowed them to become somewhat more accustomed with the format of the system before they had to give whole-system feedback.

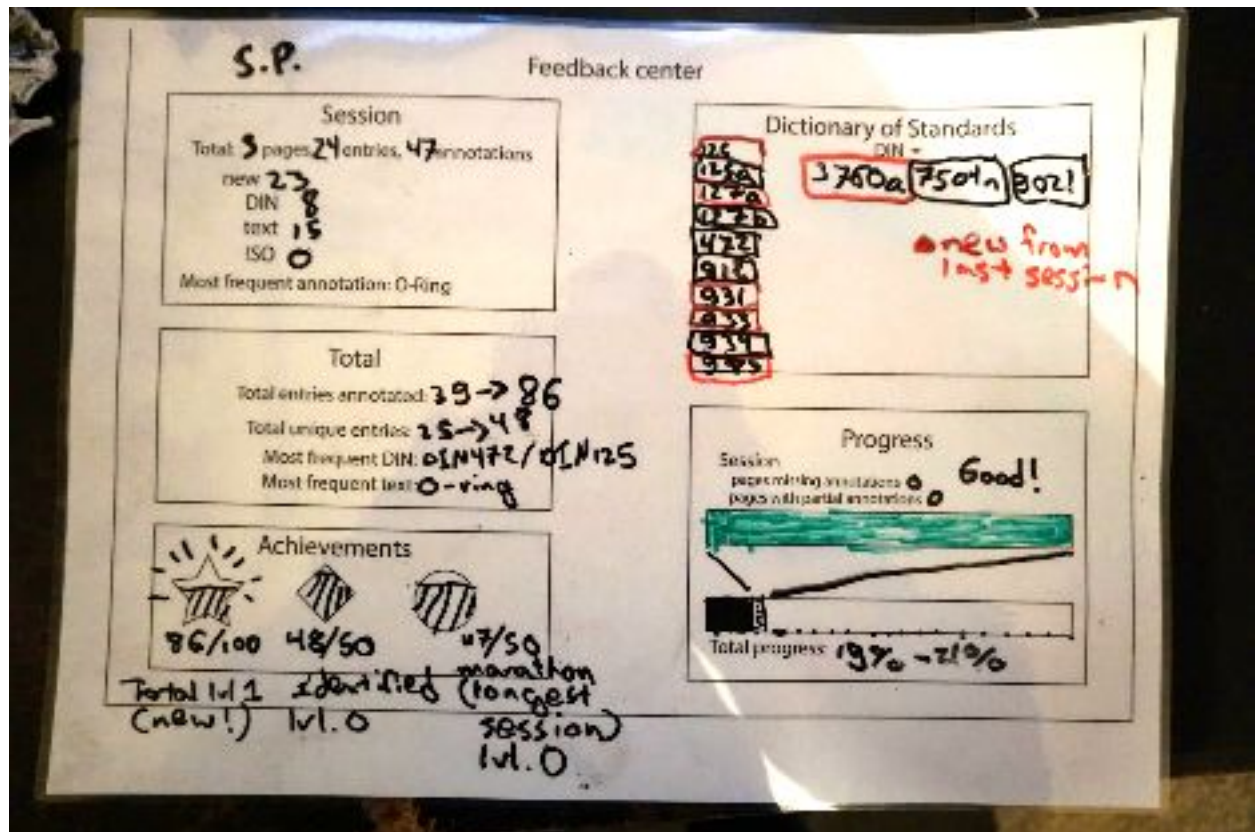


Figure 14. photo of pen-and-paper feedback center. After a user had performed 5 minutes of work, their progress was coded by hand and transferred to the laminated wireframe, and presented to the users, upon which they gave feedback in form of a questionnaire. This happened in two rounds, first revealing the left side (i.e. session, total and achievements) and then the whole interface, with an updated left side.

4.8 Results from pen-and-paper prototype testing

Except for the dictionary function, testers reported that all features were understandable, insightful and relatable to the work process. The achievements section and progress section were said to be motivating, as they offered visual feedback. Users commented that a time related feedback might be motivating, in the form of a time/productivity ratio, indicating how 'productive' a session might've been.

The 'most frequent' feedback elements in the session and 'total' features were considered superfluous by one.

The dictionary feature did not receive as good feedback as the other features. It was harder to understand, less insightful and was reported as less reflective of a user's work efforts. One user suggested that it be turned into a collection mechanic, so that users could collect and show rare annotations.

Generally, users reported that this kind of feedback regarding their work was appropriate, likely to be helpful and likely to be motivating. One user reported that, while the system was 'nice to have', the work process itself 'remains unmotivating'.

On the additional question if users would be comfortable showing their feedback center in a public office space, answers were mixed: one user said yes, one user said 'maybe when I feel comfortable with it' and one user said 'only in the features 'Dictionary and Progress''. As the testing and focus of this project is on the individual experience and conclusions from these users can not be used to hypothesize the real workplace and professional users, further investigation into public display of workers' results is omitted.

4.9 Adjustments for high-level development

The dictionary feature, intended to evoke the dynamic of 'Meaning', was not effective in the presented format. This is partially due to the wireframe nature of the test, but implies that it should be realized in a different format. Furthermore, the progress section and achievements section did not achieve 100% in the questionnaire question 'easy to understand', indicating that in a high-level version these features should be accompanied by explanations. In the design of high-level mockups for a final version, these were focus points, and would be criteria in deciding what kind of design, layout and visual elements would be chosen for implementation in a high-level prototype.

An additional observation made during testing was that the testers had different levels of confidence and understanding of the data, and that they were not proficient enough in recognizing what was or wasn't supposed to be annotated - this can be considered indicative of how other outsiders to the task may perform in the test. As there is no standard (no pre-annotated and quality checked set of the data) to hold testers against, giving users feedback on the amounts of missing or partial annotations, as in the upper progress bar in the progress section, isn't valuable in this stage of the product development. Thus, for the further scope of this project the investigation into how quality can be maintained or improved is omitted from the requirements. When the system is adapted for the real work place to be used by expert users, options of peer-review and quality control can be reintroduced.

Chapter 5: Specification

5.1. Visualization choice

With an outline of system features and improvements that must be applied, the system was redesigned to incorporate the well-understood features, and offer a better format of visualization than the previously tried dictionary of standards. In order to do so, three different designs were sketched, and can be seen in figures 15-17. In designing the interfaces, some features were kept in their original form (session and total feedback, achievements), whereas the other features (the dictionary of standards/data visualization and the progress indicator) were experimented with in each version.

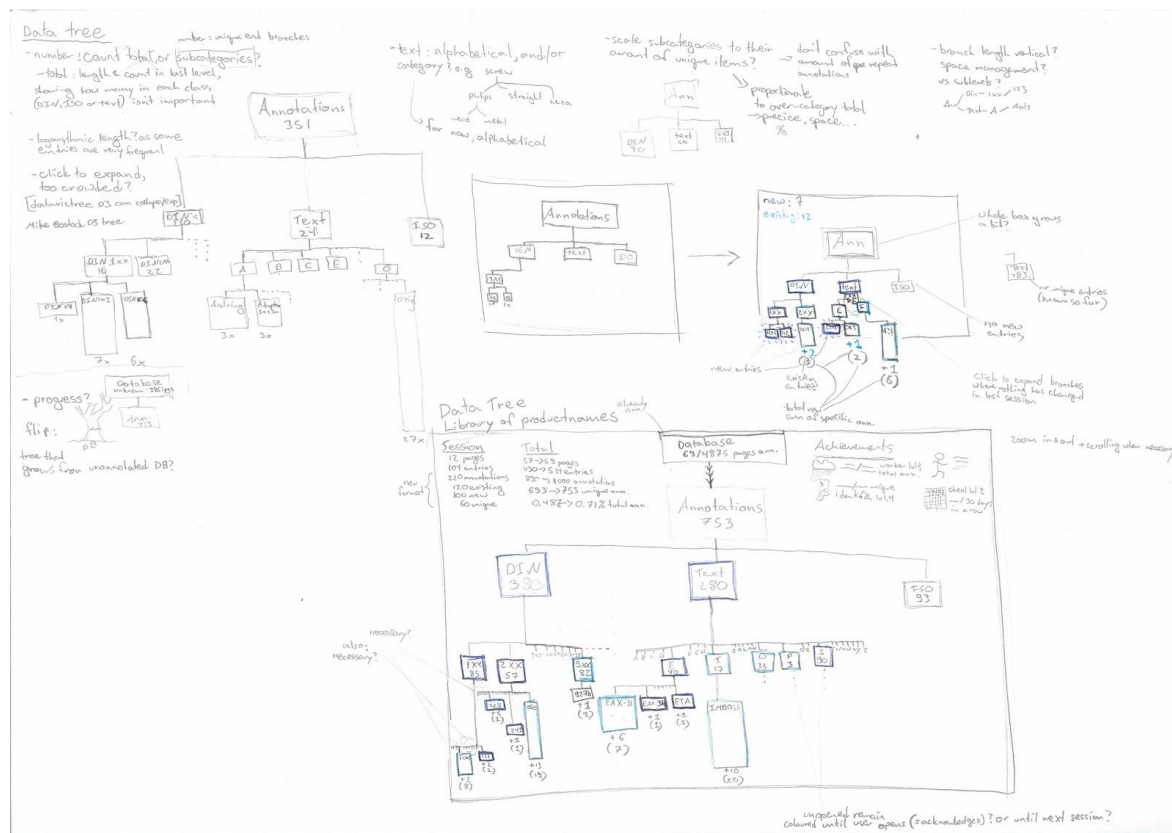


Fig. 15 interface sketch with DataTree visualization.

DataTree visualization

As users annotate, the tree is expanded, both in breadth (for the amount of different annotations) and depth/branch length (for the amount of recurring annotations). The section on top of the tree represents the unannotated database (similar to a progress bar), intended on showing it being 'processed' into the tree. This version emphasizes categorization between the

Annotation progress bar

Idea: have w/ring that, with work, goes over 'gray' blank sheets, revealing the unknown annotations.

Magnify current session as main part of visualization, showing which annotations are new.

- comparison to e.g. last 3 sessions to see if sessions are constant, shorter, longer, trend

- focus on linear progress

- tie graph above total bar?

Step by step... Incremental progress!

Session
9 pages
23 entities
150 annotations

Annotations
100 existing
50 new

Achievements
Identifier level 3!
500 unique annotations

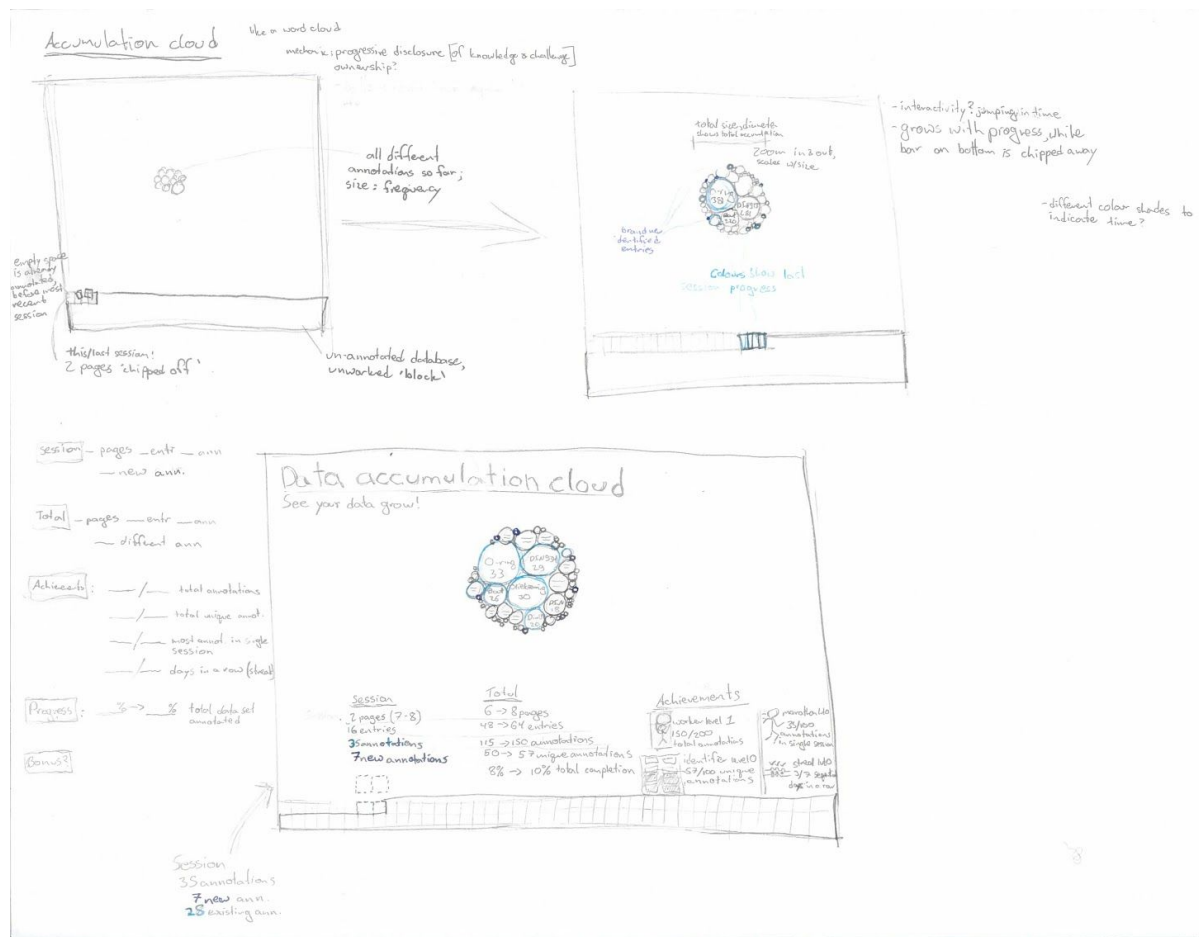
Total
03/06 page:
410-7487-21-105
200 → 550 annotations
8.0% → 8.3%

Total Annotations
503 unique annotations

Scale 10 pages?

Progress bar visualization

36



The bubble graph concept was chosen to further develop for a number of advantages that it has over the other visualization designs. Amongst the three versions, it is the most straightforward representation of a users accumulated annotation efforts, and how 'large' their body of work is so far - in one glance a user can see how many different annotations they have come across. This is in contrast to the tree concept, where a user would have to navigate through branches to see in each subdivision which new annotations they have made, and the progress bar concept, which focuses more on the progression through the whole database than showing the accumulated volume of annotations. This advantage is particularly relevant for testing with non-professionals, as it does not require e.g. potentially confusing subcategories, such as in the data tree model, and is assumed to offer a more gratifying change (growth) from session to session than the more linear progress bar concept does.

Another argument that spoke for the Datacloud concept is feasibility. In researching different available resources on data visualization that could be used in a prototype application, examples and usable code was found for the Datacloud concept, but not for the other concepts. Furthermore, the bubble cloud concept is the simplest in terms of visual elements, and is thus estimated to be the easiest in realization. While there are surely other visualization concepts that can be explored and applied to an annotation feedback system, the DataCloud concept was chosen for its simplicity and feasibility in the scope of this project. After all, it is only one of the many gamification elements that are part of the whole system which is being tested and evaluated. In future development, other models can be designed and developed with the expert users to meet their specific preferences.

5.2. Development of high-level design: usability testing

In order to develop the visual design of a high-level interface, the DataCloud design was redone in digital format and improved through multiple rounds of informal usability testing.

Starting with a first version (see figure 18), usability testing was conducted by asking previous testers (those who performed in the paper prototype testing, see section 4.5) to perform work, and then present them with the digitized design, upon which they were asked to give written, open feedback in terms of:

- Ease of understanding
- Any features or elements that they found confusing or misleading
- Aesthetic impression regarding the layout and information shown
- Missing or unnecessary information
- General impression of system and any other comments or improvements they may have

After conducting the usability test with three previous users, who all gave feedback, the interface design was accordingly adjusted, and the process was repeated three times, until all necessary changes had been made. The final interface design can be seen in figure 19.

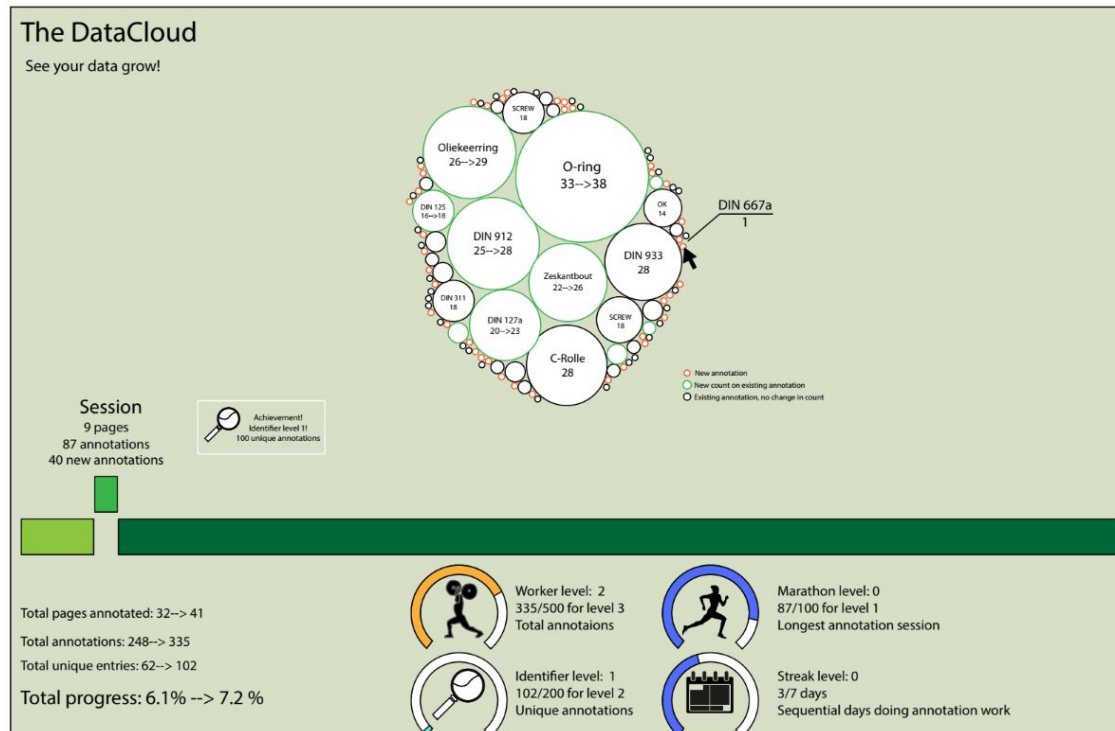


Fig. 18. First digital version of DataCloud interface design.

First digital version of the DataCloud interface design

All of the previously implemented features (progress bar, session and total feedback, achievements with user levels) are used, as well as an example of how the bubble chart visualization will look. In the first version, the layout and visual elements were commented on - colours served here primarily as cues to distinguish different elements from each other.

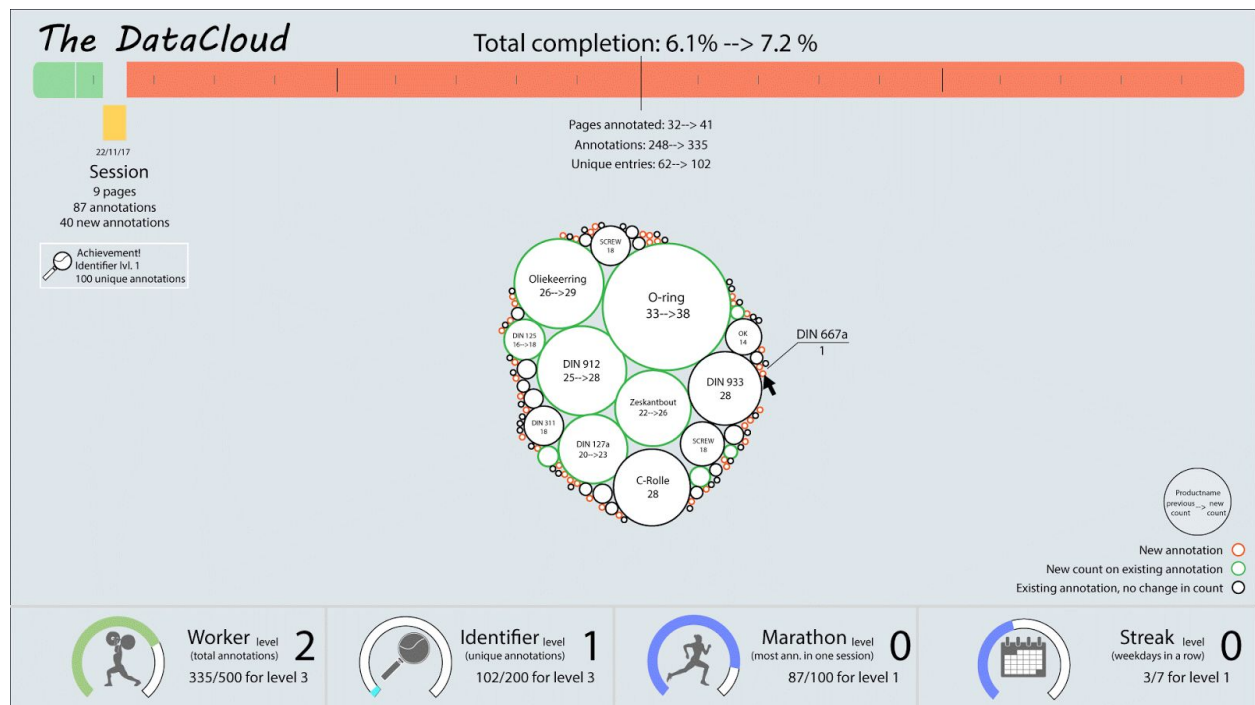


Fig. 19. Final layout and design for high-level prototype.

Final layout and design for the high-level prototype

The most significant changes are the placement of the progress bar, the placement of the total feedback, and the distribution of the achievements on the bottom of the interface. For ease of use, the numbers were made larger, as well as the bubble chart legend. During the prototype build, some final adjustments to the text sizes and annotation counts were applied.

Chapter 6: Realization

Based on the visual mockups and requirements, the prototype was realized as a web application using HTML, CSS and JavaScript. A screenshot of the final prototype can be seen in figure 20. The application runs on a web server, and can be used parallel to the brat annotation system and files that are being annotated.

For technical reasons, there are some (aesthetic) differences between the final mockups and the tested prototype. These are:

- Labels in the bubbles of the visualization appear through hovering with the mouse over them
- The progress bar has been color-adjusted to have coherence with the 'Total' and 'Session' titles, as well as the color of the bubbles in the visualization. The 'Session newly discovered' bubbles have the same color as the session progress in the progress bar, and the same color as the title on the left side; the 'Total' progress bar section has the same color as the 'Total' percentage indication in the top and the 'increased count' bubbles in the visualization.
- The progress bar does not have percent-markings, as this was not manageable within the development timeframe.
- The page count was taken out, as it was unreliable during pre-testing
- The achievement sections use full circles in used prototype, instead of the previously designed C-shapes, as this was more feasible within the development timeframe
- A button has been added to update the feedback (this was technically more feasible than a system that would update on its own)
- Some of the numbers were enlarged for legibility

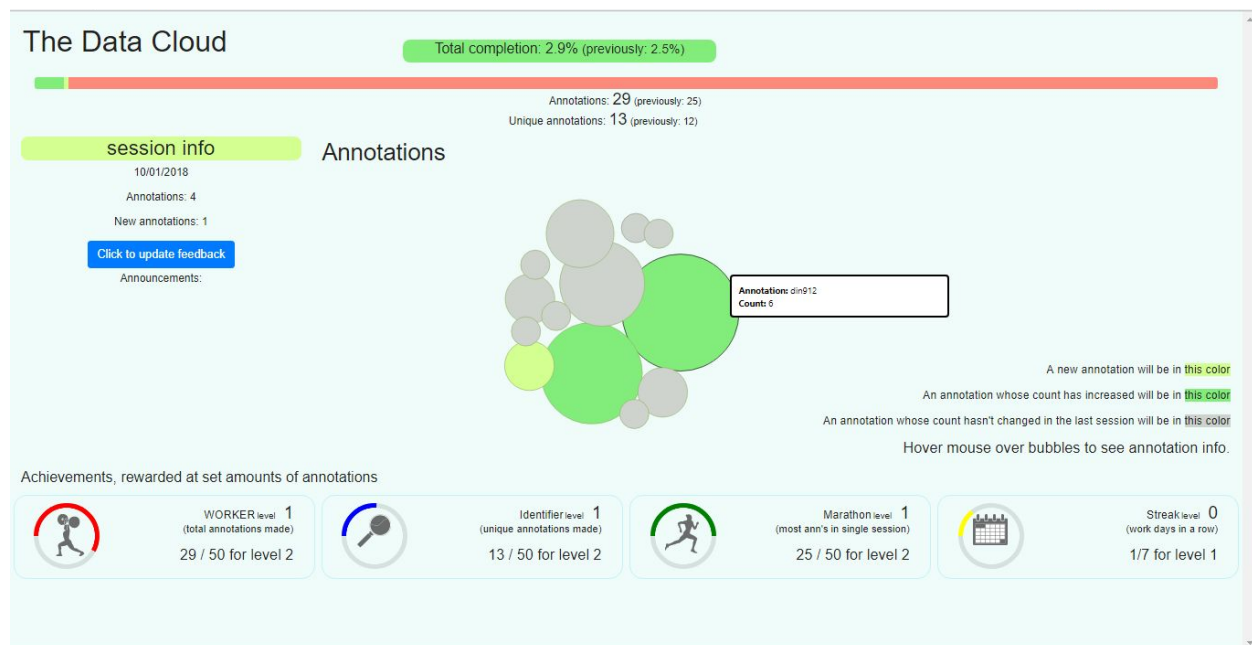


Figure 20. Screenshot of prototype developed and used in testing.

Chapter 7: Evaluation experiment design

7.1 Test outline, variables and hypotheses

In order to investigate the impact of gamifying the annotation work process, an experiment was designed. As a whole product concept, the goal is to improve users' engagement with the task, motivating them to contribute more time and effort to the project. The gamification methods, drawn from theory and applied to the annotation work process are intended to facilitate this, and are offered in the form of the previously described visual feedback system.

Ideally, the system would be presented and tested with real annotation users, who would use it in their workplace over a longer amount of time, thus indicating its usefulness in its proper context. As the limited availability of real users excluded this approach, the experiment was designed with pseudo users as testers to project if and how potentially effective this system could be in the workplace. To do so, the designed experiment investigates two things: how many annotations users make in their given time, and how users react to feedback given in different formats. Specifically, their (self-reported) levels of engagement, satisfaction and resulting motivation towards the annotation task and feedback are of interest. This is investigated by having pseudo users perform the annotation task, and be given one of three kinds of feedback (the original file-feedback, a non-gamified feedback and the developed gamified feedback prototype).

Independent variable: feedback users receive

In this experiment, the format of feedback is the independent variable imposed on test participants. In order to distinguish the effects of the gamification, three conditions are set up under which users would be tested:

1. receiving no feedback, but having to actively retrieve the log files (which is the current state of the work activity),
2. receiving non-gamified feedback in the form of a simple info page, and
3. receiving feedback with the gamified feedback system.

This way it should be possible to distinguish the difference between non-gamified feedback and the gamified feedback, and these two from the current state (in which, if users have interest, they must retrieve and interpret the brat system log files themselves). A screenshot of the developed non-gamified feedback can be seen in figure 21. The non-gamified feedback form offers feedback on 3 aspects: A users total work (total annotation and total unique annotation count), a users most recent session (total annotation and total unique annotations of the last session), and an overview of the different annotations a users has made and the number of occurrences of each. These feedback points are currently non-existent in the brat annotation software, which is given in condition 1.

By giving users a version of feedback that was not gamified, the goal was to differentiate the effects of the gamification from a feedback form that used essentially the same data but in a

non-gamified format. Thus, if the non-gamified feedback would score higher or equal to the gamified feedback in the test, it may indicate that the gamification offers no significant improvement in terms of engagement, satisfaction or motivation.

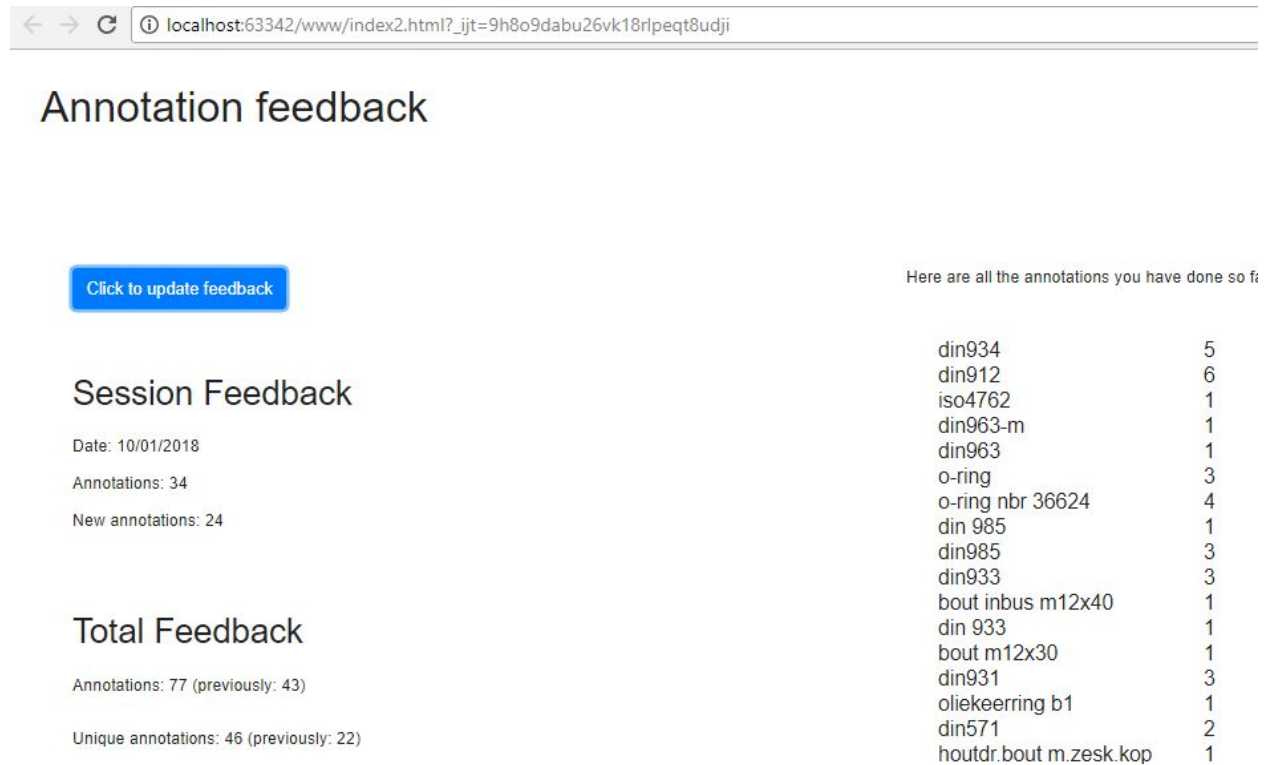


Fig. 21. Screenshot of non-gamified feedback developed for testing; this was given to users of condition C2 (non-gamified feedback).

Dependent variable: work and reception; Hypotheses

The two dependent variables of interest are

- the amount of work (numbers of annotations done) by users, and
- their reaction to the type of feedback they receive (in terms of perceived engagement, motivation and usability).

From these, 3 hypotheses were formed which were tested in the experiment:

1. H0: There is no difference in quantity of participants' annotations when getting feedback from a log file, a simple info page or a gamified feedback system.
Ha: The quantity of participants' work session is statistically higher when they receive feedback in the gamified format than when they receive feedback in either other condition.

To assess this, the amount of annotations created per user during their test were documented either in form of interface screenshots after completion of the 10 minutes of work (non-gamified feedback and gamified feedback) or manual count (file feedback).

2. H0: Users perceive no difference in motivation and engagement when getting feedback from a log file, a simple info page or a gamified feedback system.

Ha: When users receive feedback in different forms (original log files, non-gamified feedback and gamified feedback), they are more engaged with the task and feel more motivated and challenged when receiving feedback in the gamified format than when receiving feedback in the non-gamified format or original log file format .

3. H0: Users perceive no difference in usability when getting feedback from a log file, a simple info page or a gamified feedback system.

Ha: Users report improved usability when receiving feedback in the gamified format in comparison to the non-gamified format or original format.

To assess hypotheses 2 and 3, a questionnaire was developed with 22 statements and attributes which will be evaluated by users after their test participation. The questionnaire is detailed in section 7.2.

Study design

For the experiment, the between-group design was chosen. By letting participants only experience one form of feedback, a preference bias can be avoided, and participants can only decide whether they favor their assigned form of feedback over no feedback, and not in comparison to either of the other forms. Additionally, as the work is repetitive and not inherently interesting to outsider testers, fatigue and boredom from performing the task could affect their perception of second and third feedback forms.

Disadvantages of the between-group method are that users do not have any other formats of feedback to compare the version they received to - this is however reflective of the situation which future workers will also be in, as there is no existing feedback system currently in place. Furthermore, there is a risk that individual differences can bring in noise and strong variations in tester performance and subjective perception of each system. To counterbalance this, a group of 30 testers will be recruited (10 per condition, so 30 in total). These will then be randomly assigned to one of the three conditions before conducting the research experiment with them. The experiment can be conducted in any quiet, non distracting environment such as a library room or office space, as long as there are no other people present or environmental influences that may impact the experiment. Only a computer running the brat system and the feedback system are essential to run the experiment. Furthermore, as the product and investigation focus on the experience of the individual user, no other people (besides the experiment conductor) are required to be present.

The sequence of events in the experiment is planned as follows, according to the guidelines of Lazar, Feng & Hochheiser [14]:

1. Ensure the systems (the brat annotation software, the log system, the feedback systems) are ready for the experiment.
2. Greet the participant.
3. Introduce the purpose of the study and the procedures.
4. Get the written consent of the participants.
5. Randomly assign the participants to one of the three conditions.
6. Participants complete training task (1 page of brat annotations) to become familiar with the data and annotation process.
7. Participants complete actual tasks.
 - a. Brat annotation for 5 minutes
 - b. Receive feedback according to the condition they were assigned (if condition 1, presenting users with brat annotation files; if condition 2 or 3, presenting them with the corresponding screen)
 - c. Brat annotation for second round of 5 minutes
 - d. Receive updated feedback in the assigned form
8. Questionnaire regarding satisfaction and reaction to their assigned feedback form
9. Debriefing session.
10. Thanking for participation.

In order to mask the main research goals of the experiment (investigating the effect and reception of the feedback forms) and thus avoid biased answers, the experiment will be explained as the testing of an annotation software (as a whole, in contrast to bringing attention to the feedback aspect). Users will be told that they are simulating the use of such an annotation system in the office environment. After the second round of testing, the questionnaire will then specifically investigate if and how the feedback the users were given was stimulating, motivating and generally positively received.

As each round is timed, it will be possible to quantify the amount of work done in that time and compare users' amount of annotations against each other. As the task itself is fairly straightforward, a single-time test with users should suffice to collect an impression which is relatable to their efforts.

As part of the investigation, users will be asked via questionnaire about their past experience in doing work of similar nature: repetitive, linear tasks that do not vary in difficulty and for which they did (or did not) receive any form of feedback. Examples of these might be physical labour (such as packaging, sorting or assembling), coding data or correcting or manipulating text files. As follow-up questions, users will be asked if they did or didn't have any feedback, and if they can see how a system similar to the one tested could (or would) be motivating in performing such work. Testers being exposed to the three different forms of feedback should be able to relate the tested system to past experiences, if they have any to share.

7.2 Questionnaire

The questionnaire is a combination of questions from various established questionnaires [15], [16], [17], [18] and custom questions.

The questionnaire designed for the test consists of 6 sections:

1. Entering of tester number and introduction text
2. Likert scale of statements ranging from 1 (completely disagree) to 7 (completely agree); 15 statements in total, of which 3 are negative statements.
3. Rating of feedback on 7 attributes of the feedback users received, rating from 1 (negative attribute) to 7 (positive attribute).
4. Open questions 'What did you like about the feedback you received?' and 'What did you dislike about the feedback you received?'.
5. Yes/No/Other and open questions regarding past experience doing repetitive, monotonous work and if and how they might have received feedback that made the work more satisfying, easier or more fun.
6. Final comments on what they would change on the prototype if they would have to use it in the future, and basic demographic info (age, gender, level of education and employment status)

In the following, the questions of sections 2 and 3 are detailed, as they will be used to measure the success in creating an improved and motivating/enticing user experience:

Questionnaire section 2: Likert scale statements on engagement and motivation

The following statements were given to users, who rated their agreement with the statement on a scale from 1 (completely disagree) to 7 (completely agree).

1. If I could, I would have worked longer to complete a set amount of annotations (e.g. 50, 100, 200, 500 etc.).

This statement aims to investigate a (new) willingness to go further than they could in the given time, thus indicating whether the feedback motivated them to work more. As the gamified version uses achievements that are rewarded for set amounts of annotations, these elements ideally trigger the ambition to reach them, thus doing more work than when no such achievements are present.

2. Having done a set amount of annotations (e.g. 50, 100, 200, 500) would give me satisfaction. Again, aimed at the effect of the achievements linked to increased annotation work. Ideally, achievements work as a challenge, and then give the user a feeling of satisfaction when they are accomplished.

3. I feel like I accomplished something.

This statement is used to investigate the feeling of accomplishment users experience - if agreement is higher in the gamified version, it was successful in framing work progress in a rewarding way, better than in the other two versions.

4. To see my progress grow, I would often pick up the annotation task inbetween other tasks.

This statement aims at investigating if the gamified feedback, with a progress bar showing completion of the project, a visualization showing the body of work and the various achievements honoring milestones of contributions, evokes an ambition to perform the work more often than when they are not present. If users feel a drive/motivation/ambition to perform the work more often, this may be indicative to a higher amount of output than when workers are given one of the other two feedback versions.

5. I felt challenged to do more annotation work.

If users feel challenged to do more annotation work when given the gamified feedback, the gamification elements were successful in that respect. If users given the gamified feedback feel more challenged, this may be indicative to a higher amount of output than when workers are given one of the other two feedback versions.

6. In the second round, I looked forward to seeing the results of my work.

This statement investigates users' eagerness to engage with the (feedback part of the) annotation system, and a curiosity towards the results of their work. If users given the gamified feedback have higher agreement with this statement, it can indicate that the already existing work has been reframed in a more stimulating and engaging format.

7. I would have updated my feedback more often, if I could have.

This statement, similar to statement 6, investigates if users had a curiosity to see the feedback, updated whenever new work was performed. This may be indicative that, with the gamified feedback, work will result in visual feedback that is interesting to see, and thus more stimulating and engaging than when feedback is given in the two other formats.

8. The feedback I got was childish.

This statement investigates the appropriateness of each feedback format. If a feedback is childish, it is workplace inappropriate.

9. The feedback I got was inappropriate for a workplace, such as an office.

This statement is similar to statement 8, and is intended on investigating appropriateness for a workplace.

10. I feel a sense of ownership over the work I've done.

If users report a larger sense of ownership over their work when given feedback in the gamified version, the gamification dynamic of ownership was successfully realized. This may be

indicative to a heightened attention to the annotation task, and feelings of responsibility around it.

11. I feel like my work was not significant.

This negative statement is intended on investigating if users feel that their work was meaningless or insignificant, which may have demotivating effects. If users given the gamified feedback have high agreement with this statement, it may be indicative that the gamification was not successful in framing the work and challenges in meaningful and rewarding ways.

12. Doing the annotation work was satisfying.

This statement investigates whether users given the gamified feedback felt a greater sense of satisfaction knowing that they would be given feedback in that format - if users feel a greater sense of satisfaction doing the work, it may be indicative that the user experience of doing the annotation work has been improved.

13. I would share my feedback with others who were tasked to do the same work.

This statement investigates whether users are open to showing other users their feedback, which may be evoked by a sense of pride and ownership over the results, which may be indicative of a greater engagement and better user experience around the annotation task.

14. I would be interested in seeing the feedback of others such as colleagues, who also did this annotation work.

This statements investigates test users' curiosity to see how other users performed the annotation task. If users given the gamified feedback rate higher agreement with this statement, it may be indicative of a sense of competition and curiosity, and may indicate that social elements may be successfully implemented in future versions.

15. I was interested in the content of the database.

This statements investigates users' curiosity and interest in the actual data of the database. If users given the gamified feedback have higher agreeability with this statement, it may be indicative that the gamified feedback resulted in a more engaging and interest-awakening experience than the other two versions.

Questionnaire section 3, feedback evaluations/rating

The following attribute pairs were given to users to rate on a scale from 1 to 7.

16. Ease of understanding (1: very difficult to understand, 7: very easy to understand)

Ease of understanding is essential to good user experience and a grasping of the mechanics being used. While this attribute must not necessarily be higher in the gamified feedback relative to the other versions to show effects of gamification, it should be high, which may be indicative of understandable design and good user experience.

17. Frustration vs. satisfaction (1: very frustrating to see, 7: very satisfying to see)

This attribute is directly indicative of user satisfaction. If users of the gamified version rate this attribute higher than users of the other two versions, it is indicative of higher user satisfaction around the annotation task.

18. Boring vs. interesting (1: very boring/dull, 7: very interesting)

This attribute aims at distinguishing whether receiving feedback in the gamified format is more interesting than the other two formats. It may be indicative of an improved user experience.

19. Relevance to the task (1: completely irrelevant to the annotation task, 7: highly relevant to the annotation task)

This attribute aims at investigating the relevance of the gamified feedback to the actual annotation task. If users given the gamified feedback rate a high relevance to the task, this may be indicative of a higher user engagement - what users see is meaningful and not superficial.

20. Motivational value (1: very demotivating, 7: very motivating)

This attribute directly aims at investigating the motivational value of the gamified feedback. If users given the gamified feedback rate this higher, it may be indicative of a better, more engaging and motivating user experience around the annotation task.

21. Usefulness (1: completely useless, 7: very useful)

Similar to 'relevance to the task', this attribute investigates the functional usefulness and relevance to performing the annotation work. If users given the gamified feedback rate this high, it may be indicative that the gamified feedback is meaningful and has weight, and is not useless, which may be indicative of a lesser user experience.

22. Visual value (1: ugly, aesthetically displeasing, 7: good looking, aesthetically pleasing)

This attribute investigates whether the gamified feedback offers a visually improved and superior experience in comparison to the other two versions. If users given the gamified feedback rate this higher than users given the other two versions, it may indicate a better user experience.

7.3 Adjustments for test

In order to test the effects of the various gamification elements within the scope of the 2x5 minutes time frame that users have to work and experience feedback, some adjustments were made to the gamified feedback prototypes elements to give the test users an amplified impression of their meaning and intended effect. These adjustments are:

- Achievements (Worker, Identifier and Marathon): the amount of annotations needed in order to reach levels 1, 2 and 3 were set to 10, 50 and 100 annotations respectively. These numbers were chosen as in pen and paper testing levels 1 and 2 were reached by all users, and that the amounts necessary seem within reasonable reach.

- The streak achievement was set to a default value of 1, as users were one-time participants, participating on a single day.
- The progress bar, indicating completion of the whole annotation goal, was set to 100 pages; this comes on the one hand from having 100 pages of sample data, as well as that it affords a simple percentage calculation (1 page is 1 percent progress) and that it would be a relatable number calculation for test users - if users see they have done 5 pages, it shows progress of 5%. Due to technical challenges on the page counter, the completion value was adjusted to be calculated based on the amount of annotations made vs. the estimated amount of annotations that could be done on 100 pages (each page should have 10 entries, with each entry having at least 2 possible product names: with x being the completion percentage, the value was calculated $x = (\text{Nr. annotations} / 2000) * 100$). After pilot testing, the total annotation task scope was reduced from 100 pages to 50 pages, meaning one annotation = 0.1%, which was intended on giving users in the short testing time a greater sense of achievement and progress.

Chapter 8: Evaluation results

In the following the test results are summarized, and a conclusion in regard to the hypotheses testing is made based on the results. The complete test results, comments and demographic information on testers can be found in the appendix.

Significance was measured as follows:

1. All values for a particular statement were collected and run through a one-way ANOVA test (Analysis of Variance)⁹ to investigate whether a statistical difference between conditions may be present.
 - a. If statistical significance can be inferred, a post-hoc Tukey HSD (Honest Significance Difference) test¹⁰ is performed to investigate between which conditions a statistical significance may be present.
 - b. If no statistical significance can be inferred, further investigation is omitted.
2. Results inferring statistical significance are run through a post-hoc Tukey test to determine between which conditions a significant difference may be present.
 - a. If a statistical difference can be inferred, a significant difference between the tested conditions in respect to the statement or attribute can be claimed.
 - b. If no statistical difference can be inferred, no difference between the tested conditions in respect to the statement or attribute is claimed.

8.1 Hypothesis on quantity of annotations made

As one of the objectives in gamifying the annotation task is increasing the amount of work done by the workers, the first hypothesis was defined as follows:

H0: There is no difference in quantity of participants' annotations when getting feedback from a log file, a simple info page or a gamified feedback system.

Ha: The quantity of participants' work session is statistically different when they receive feedback in different forms.

The following counts regarding the amount of annotations done by test users were obtained:

Condition	Average annotations done in 10 minutes (std. dev.)
C1 (file feedback)	118.7 (32)
C2 (non-gamified feedback)	115.7 (39.4)

⁹ All results calculated with the online one-way ANOVA test tool at:
<http://www.socscistatistics.com/tests/anova/default2.aspx>

¹⁰ All results calculated with the online Tukey HSD test tool at:
http://astatsa.com/OneWay_Anova_with_TukeyHSD/

C3 (gamified feedback)	125.9 (43.7)
------------------------	--------------

Table 3. Average annotation count across conditions 1-3

The following statement from ANOVA testing at $p = 0.05$ was made:

At $F(2,27) = 0.184$, $p = 0.833$, the differences are not statistically significant.

There is no statistically significant difference in the amount of annotations made when users are given different formats of feedback.

Thus, the Null-Hypothesis can not be rejected - none of the versions were proven to be significantly more effective than the others in making users do more annotations in the given time.

8.2 Hypotheses on user engagement and motivation and usability

In the following table 4, the average ratings per condition per statement and attribute from questionnaire sections 2 and 3 are listed, as well as if there is a statistically significant difference between two or more conditions. The list of statements and attributes can be found in section 7.2. The average values for agreement with the statements per condition can be seen in figure 22, the average values for attribute rating can be seen in figure 23.

Statements

State ment	C1 avg. (std. dev.)	C2 avg. (std. dev.)	C3 avg. (std. dev.)	Anova Significa nce?	Tukey significance?
1	4.2 (1.8)	3.4 (2)	4.7 (1.2)	No	No
2	4.7 (1.5)	3.1 (2.2)	4.9 (1.1)	Yes	No
3	3.8 (1.2)	3.8 (2.3)	4.2 (1)	No	No
4	3.3 (1.6)	3.7 (2.1)	4.4 (1.5)	No	No
5	3.2 (1.7)	2.6 (1.8)	4.9 (1.5)	Yes	Yes (C2 < C3)
6	3 (1.6)	3.3 (2.5)	6 (0.7)	Yes	Yes (C1 < C3 and C2 < C3)
7	2.8 (1.6)	2 (1.8)	3.8 (1.8)	No	No
8	2.7 (1.5)	1.3 (0.7)	2.2 (0.6)	Yes	Yes (C2 < C1)
9	3.2 (1.6)	2.1 (2)	2.8 (1.5)	No	No

10	3.9 (1.5)	3.2 (2)	4.2 (1.4)	No	No
11	3.3 (1.6)	3.5 (2.5)	3.9 (1.2)	No	No
12	3.6 (1.5)	3.3 (1.5)	3.8 (1.1)	No	No
13	4.4 (2.1)	3.8 (2.5)	4.9 (2)	No	No
14	4.3 (2.2)	3.8 (2.4)	4.9 (2)	No	No
15	3.6 (1.8)	2.8 (2.1)	4 (2.1)	No	No

Table 4. Average user agreement ratings with statements 1-16, standard deviations and statistical significance

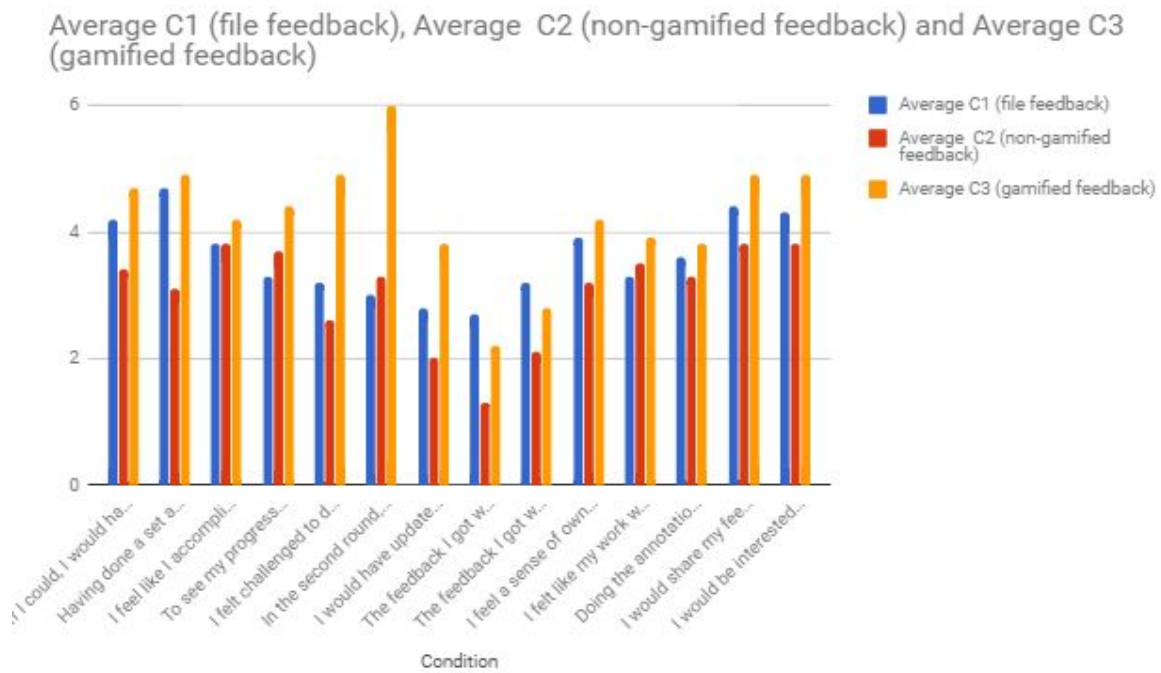


Fig. 22 average user ratings on the statements 1-15. The complete statements can be found in section 7.2. With exception of statements 8, 9 and 11, a higher score is better (indicates higher agreement with the statement).

The following claims can be made about the different feedbacks in which the Tukey post-hoc test proved statistical significance:

- **Statement 5: I felt challenged to do more annotation work.**

In condition comparison, there is a significant difference in the agreement on the statement between conditions C2 and C3. Users given gamified feedback had significantly higher agreement with the statement than users given the non-gamified feedback.

- **Statement 6: In the second round, I looked forward to seeing the results of my work.**

In condition comparison, there is a significant difference in the agreement on the statement between conditions C3 and the other two conditions. This means that users given the gamified feedback had significantly higher agreement with the statement than users given either other format of feedback.

- **Statement 8: The feedback I got was childish.**

In condition comparison, there is a significant difference in the agreement on the statement between conditions C1 and C2. This means that users given the non-gamified feedback had significantly lower agreement with the statement than users given the file feedback.

While statement 2 is inferred to have statistical significance in the ANOVA test, the Tukey test could not distinguish any statistical difference between any of the conditions.

Attributes

Attribute	C1 mean (std. dev.)	C2 mean (std. dev.)	C3 mean (std. dev.)	Anova Significance?	Tukey significance?
16	5.3 (1.6)	5.2 (2.3)	5.7 (1.1)	No	No
17	4.2 (1.1)	4.1 (1.4)	5.4 (0.7)	Yes	Yes (C2 < C3)
18	2.9 (1.4)	2.8 (1.6)	4.3 (1.1)	Yes	No
19	4.9 (1.4)	4.3 (1.8)	4.5 (1.3)	No	No
20	4.3 (1.1)	3.8 (1.4)	5 (1.3)	No	No
21	4.4 (1.8)	3.8 (1.5)	4.7 (0.7)	No	No
22	3.8 (1.7)	2.8 (1.9)	4.8 (1.3)	Yes	Yes (C2 < C3)

Table 5. Average ratings on attributes 16-22, standard deviations and statistical significance.

Average C1 (file feedback), Average C2 (non-gamified feedback) and Average C3 (gamified feedback)

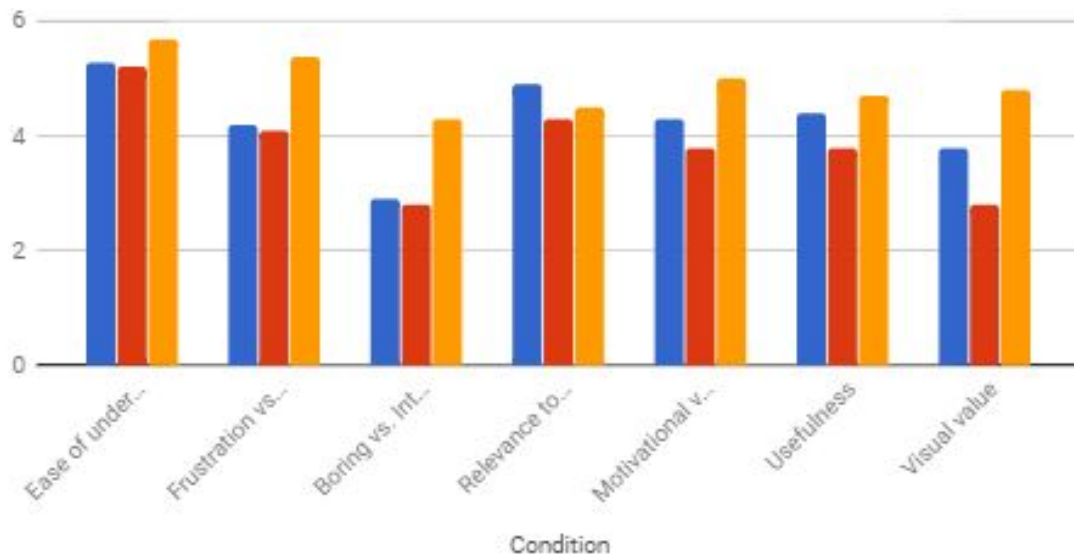


Fig. 23. Average ratings of attributes. The complete attributes can be seen in section 7.2. Higher ratings are better (closer to the positive attribute).

The following claims can be made about the different feedbacks in which the Tukey post-hoc test proved statistical significance:

- **Attribute 17: Frustration vs. Satisfaction**

In condition comparison, there is a significant difference in evaluation of the attribute between conditions C2 and C3. This means that users given the non-gamified feedback rated the attribute significantly lower than users given the gamified feedback.

- **Attribute 22: Visual value**

In condition comparison, there is a significant difference in evaluation of the attribute between conditions C2 and C3. This means that users given the non-gamified feedback rated the attribute significantly lower than users given the gamified feedback.

While attribute 18 is inferred to have statistical significance in the ANOVA test, the Tukey test could not distinguish any statistical difference between any of the conditions.

8.3 Other insights gained from test

In the following, answers from the tests are listed. A complete list of all the answers and data collected in the test can be found in the appendix.

Section 4, open questions on like and dislike:

Condition C2 (Non-gamified feedback):

2 users liked seeing how many annotations they had done in the given time.

2 users liked seeing the count on the different annotations they had done.

1 user liked being able to compare the second session with the first session.

2 users did not like that there was no graph in the feedback, one asked specifically for a graph to visualize progress.

2 users complained that the feedback was not appealing or challenging, and that it was uninteresting and not interactive.

1 user complained that there was no comparative feedback, and no information on how they did in comparison to others.

Condition C3 (gamified feedback):

5 users made positive comments about the achievements and level system.

1 user mentioned that the bubbles were not particularly relevant, as the achievements 'were all about the total number of tags'.

1 user mentioned that 'it could come across as checking up on employees by the company, it could also make the workplace quite competitive'.

1 user mentioned that 'it might be more motivating if one knew how these terms would be organised in a later database step helping the user make connections between products, names and their functions...'.

Section 5, previous work experience:

27.6% of all users who had previous experience with repetitive, monotonous work reported that they received feedback that made the work more satisfying, easier or more fun.

Some users reported that the feedback they got was given orally.

One user reported the feedback they got was 'more game-alike'.

14 out of 30 users reported self-driven attempts to try to make the task more fun with things such as music.

Section 6, comments and demographics:

Condition C1 (file feedback):

2 users asked for features that had been created as gamification elements in the gamified feedback.

1 user asked for a 'counter, also per minute/hour so you can set yourself a goal'.

Condition C2 (non-gamified feedback):

1 user commented 'maybe make it so that it encourages to keep working'

3 users asked for some 'game' features, and 1 user even mentioned gamification.

4 users asked for competitive elements.

Condition C3 (gamified feedback):

While 1 user asked for live updates instead of only receiving feedback post-test, 1 user said 'you should not be able to get feedback too frequently'.

1 user asked 'Maybe make the rewards for the employee clearer, so if the employee would finish 500 annotations, what would he get?'.

1 user asked for more background information.

From some of the comments, likes and dislikes concerning the three versions it seems that not all users were aware that there is no 'right' or 'wrong' annotations in the scope of this test - some users asked for correctness indicators, despite being mentioned in the introductory material that this was not important.

Demographics

Users were aged between 19 and 30, with the majority being 21-25.

76.7% of participants had a bachelor level of education, 13.3% a high school level of education and 10% a masters level of education.

86.6% participants were students, 6.6% were employed students, 3.3% were self-employed or a business owner and 3.3% were employed.

73.3% of participants were male, 23.2% were female and 3.3% preferred not to say.

8.4 Limitations of the test

While the test did offer the opportunity to test the two hypotheses, there were some drawbacks both in the test design as well as the provided materials that are worth mentioning.

- Testing with non-users outside of the environment and conditions that the work is usually performed in compromises the conclusivity of the results in regard to the application of the prototype to the real workplace, used by real workers. The test can only make statements about the situation the task was applied in: experimental, mostly used by students in their 20's.
- During testing, many users verbally expressed their frustration with the task itself, asking why this is not done by intelligent automatic software, which might have lead them to give biased answers. Only real users, knowing the context of the data's use before and after annotation have the inherent knowledge to not be perturbed by the work.
- As the experiment was designed to only give feedback to a single user, none of the potentially effective (and test user requested) social elements could be tested.
- Measuring the amount of annotations done within the experiment is not as effective as when the prototype would be implemented in the workplace, and users would have the option to do annotation work or other given activities. A more accurate measurement would be comparing two (groups of) workers, one with the gamified feedback and one without it, and seeing how frequently over a long period of time workers pick up the annotation task and how many annotations they contribute.

8.5 Test conclusions

Users report being more challenged to do annotation work when given gamified feedback than when given non-gamified feedback. This speaks for the gamified system's quality to engage and motivate users (to work more), at least in comparison to the non-gamified system. Users anticipated seeing their feedback more when given the gamified feedback than in either other version - this indicates an improved user experience and engagement with the annotation task. It may be indicative that the feedback has greater meaning to users when visualized using the

present gamification elements (such as a data visualization of the annotation data, and the challenges and achievements).

While there is a significant difference in (dis-)agreement with the statement 'The feedback I received was childish', this statement was aimed at investigating the gamified feedback, which was not significantly different from either other version. It is safe to assume that neither the file-feedback nor the non-gamified feedback would be evaluated as 'childish' in any case. The gamified feedback received a significantly higher rating in the attribute 'frustrating vs. satisfying' in comparison to the non-gamified feedback, indicating that the gamified feedback may have positive effects on the user experience and user satisfaction. Similarly, a significantly higher score on the attribute 'visual value' may also indicate a more tolerable user experience than the non-gamified feedback.

However, these significant differences are between the two 'new' versions of feedback, and not between the original format, the file feedback, and the gamification prototype. This may be a side-effect of testing with inexperienced users, as showing them the resulting .ann files did not evoke the desired negative reaction which had come as the main problem statement for this work. This may be due to the fact that the log files are more apparently a functional and an essential part of the annotation software (the file in which annotations are saved), and thus already serve a real role, in contrast to an additional (non-gamified) 'feedback page'. This page, despite offering more information than the log files, does not have an integral role in the annotation process. Furthermore, the apparent lack of any design or interaction possibilities in the non-gamified feedback may have been more inviting to criticism than the functional log files. One could have expected that inexperienced users would not be able to reflect the same weariness and dissatisfaction that real users struggle with.

Further, while some users of the gamified feedback appreciated the various gamification elements such as the user levels and according achievements, one should not ignore the possible affinity that mostly male students in their 20s may have with (video-) game elements, in comparison to the older users who perform the annotation task at Mydatafactory.

8.6 Reflection on MoSCoW requirements

With the testing complete, the originally defined requirements (must have and should have) are reflected upon in their realization with the prototype:

Must have:

- The product must motivate the workers to perform (and keep performing) the annotation task.

This was partially realized - in the final prototype, elements intended on motivating users to do more work, such as the progress bar and level-based goals (contributing large amounts of work to the database) were implemented but a long-term implementation which would investigate their true potential was not possible in the scope of this work. Users reported a greater sense of challenge to do more annotation work when given the gamified feedback than when given the non-gamified feedback.

- The product must increase user satisfaction around the annotation task.
This was partially realized - the test showed that in terms of means, statements regarding user satisfaction, engagement and motivation were higher with the gamified feedback, but significant increases were only seen in a couple statements and attributes. Users had significantly higher agreement with the statement 'I looked forward to seeing my progress in the second round' with the gamified feedback than with the other two formats.
- The product must increase the amount of annotations workers contribute.
While the prototype did aim at this, no significant difference in the amount of work produced could be distinguished in the scope of the test.

Should have:

- The product should be integratable with the existing annotation software, as to reduce adaptation costs and thus maintain competence with the existing task.
This was realized - the product does not change the original method of or software for annotating product data that is currently performed at Mydatafactory, and can be directly implemented in the existing system.
- The product should offer feedback as close and relevant as possible to the real performance of the workers, to keep any kind of intervention relatable to performance and the task.
This was realized - besides the use of the gamification M&D 'loss aversion' in the 'streak' achievement, all other prototype elements are based purely on the work performed and resulting annotations.
- The product should not be invasive or demanding, and allow workers to autonomously decide when and how they contribute to the dictionary with annotation work.
This was not testable in the scope of the project - while no 'active' elements such as reminders, noises or notifications have been implemented, this can only be evaluated in the workplace by workers. The 'loss aversion' gamification M&D may be intrusive in motivating workers to maintain a daily participation by threatening to discard progress on their user level if they don't. Then again, this can only be tested in the workplace.
- The product should use appropriate gamification elements based on the context analysis of the task, the workers, and the goals of both.
This was realized - all gamification M&D implemented are based on improving worker motivation, showing progress and improving user experience.
- The product should not use gamification elements that are not relatable to the annotation task or inappropriate for the workplace or context in which the task is performed.
This was realized - unrelatable or inappropriate gamification M&D were discarded in the ideation phase of the development process.

Chapter 9: Discussion with client

The tested prototype was demonstrated at Mydatafactory, and feedback on it was collected. Additionally, new concepts building on ideas that could not be tested in the scope of this work were discussed, and some recommendations regarding future work were made. Within the discussion, one available worker gave feedback on the concepts and test results. This worker was a 44 years old man and a domain specialist and thus represents a different demographic and user group than the mostly male students in their 20s recruited for testing.

9.1 Prototype discussion

The concepts most valued by the worker who was interviewed were the visualization element of the annotation work and the progress indication on conceptual goals (the progress bar). The worker elaborated that seeing the annotations, and being able to group them in a visualization e.g. according to type, and being able to show additional information would be beneficial. The worker also favored the idea of having a screen in the office displaying the annotation database, and being able to see updates to it.

The achievement and level concepts were less interesting to the worker. Also, despite some test users being interested in having competitive elements as part of the gamification, the worker rejected these ideas, saying that only if there was a designated team they would make sense.

Regarding the visual design, the worker said that the design should be as neutral as possible - no 'game' elements, no sounds, no fantasy elements or resemblance to existing games.

9.2 Concept discussion

As the experimental test was limited and could not investigate gamification elements outside of the ones used, three concepts were developed:

- Extended visualization, focusing on the data and progress in the annotation task (figure 24);
- Competition, a leaderboard with various categories allowing workers to compete in different styles of work (figure 25);
- Communal goals, a progress bar and accountability display for how often the team as a whole reached the goals they had set for themselves (figure 26).

These were then presented and discussed at Mydatafactory to elucidate if these other, more thematic concepts than the general gamification developed in the scope of this project may have favorable traits or elements that could be useful in future products.

These concepts can be seen here:

Visualization concept:

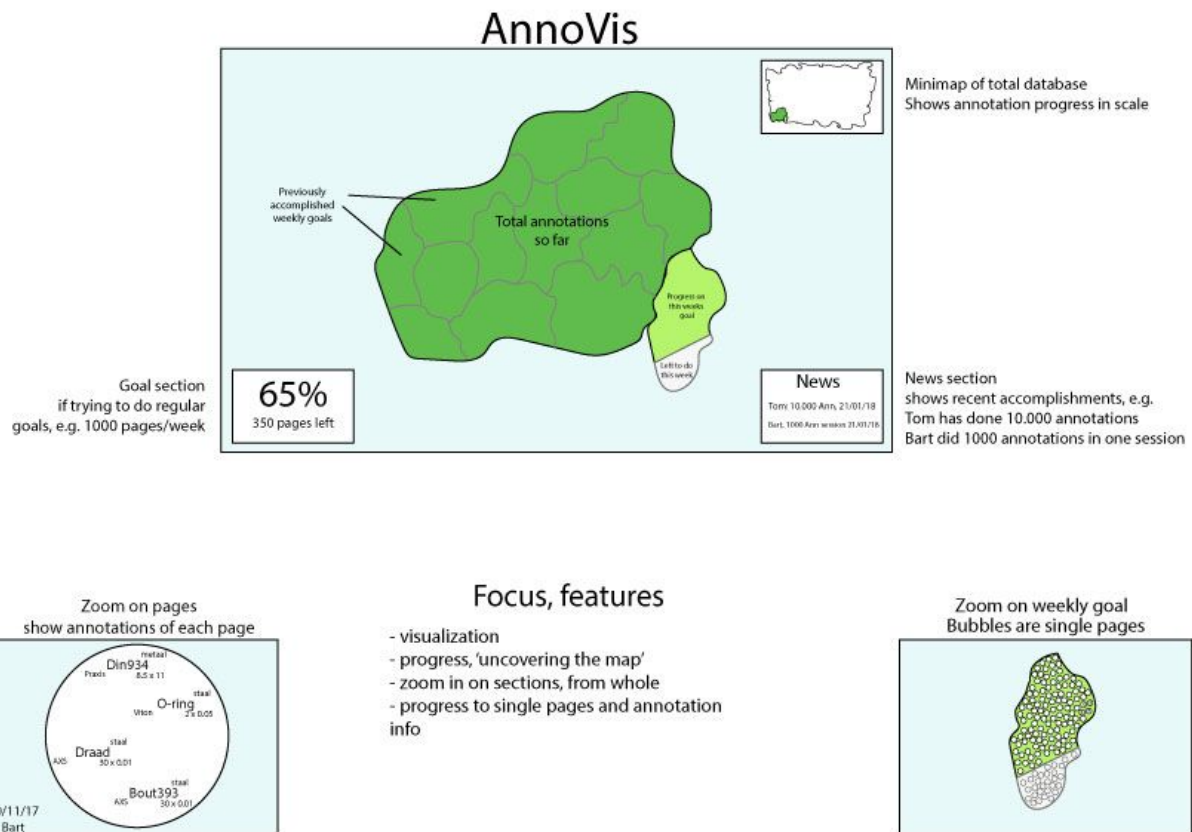


Figure 24. Visualization concept with descriptions of the elements; the large rectangle represents the screen, the two smaller ones on the bottom right and left are examples of zoom features to detail the annotation database.

This theme emphasizes visualization aspects more than behaviour-triggering gamification elements. There are however a couple additional elements built in: a progress indicator showing the completion percentage on conceptual goals, and a news section giving recognition to noteworthy achievements of office members.

Competition concept:

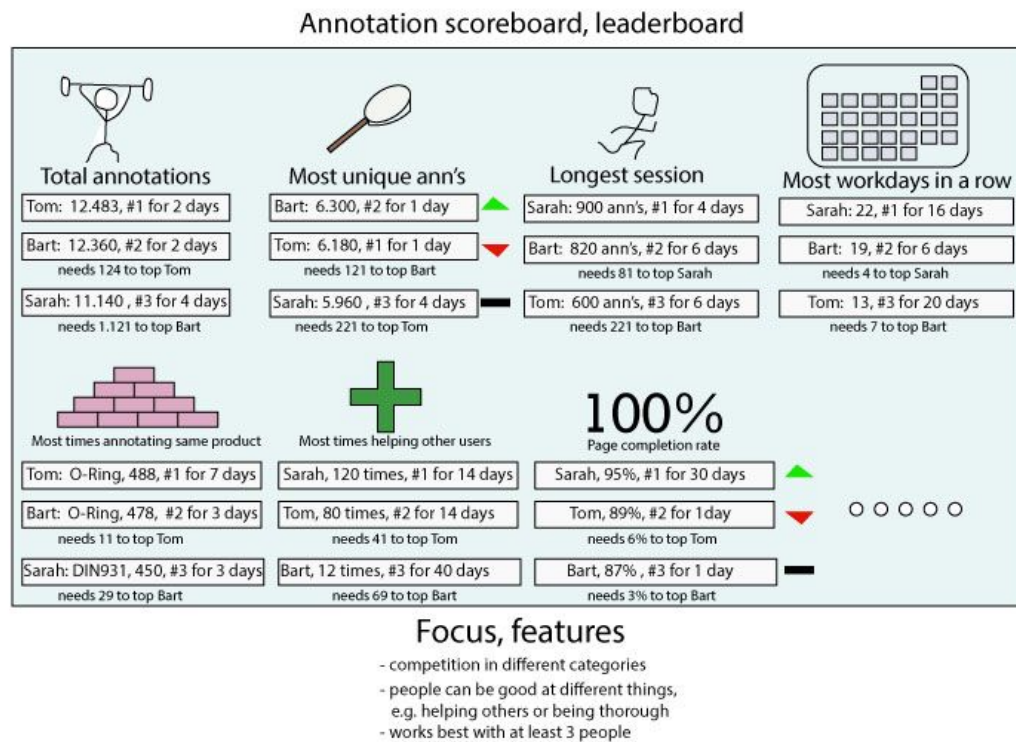


Figure 25, competition concept

In this concept, rankings and different workstyles are honored and ideally induce competition in the workplace. Users can choose to work in a specific manner, and be ambitious about keeping their position. If a competitive atmosphere can be sustained, it should ultimately lead to more work being accomplished. Leaderboards can also be reset for fresh starts in regular intervals.

Communal goal concept:

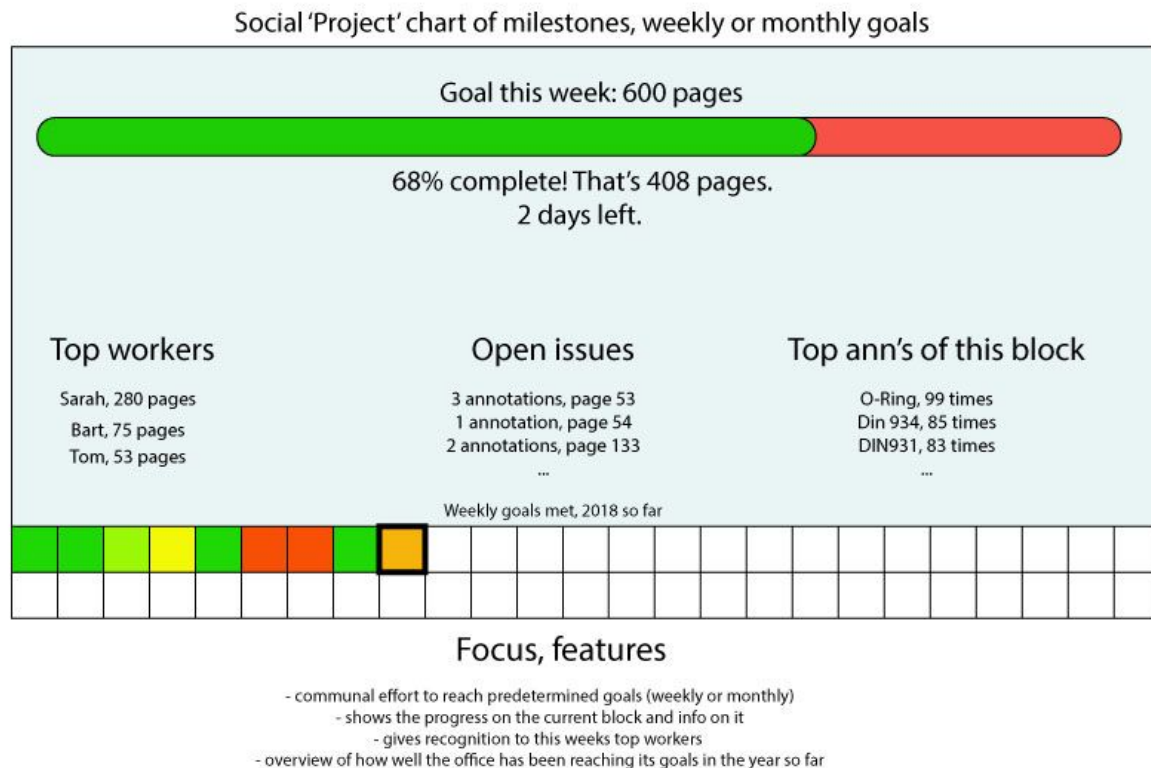


Figure 26, communal project concept

In this version the team as a whole focuses on getting a predetermined amount of work done in regular intervals, such as a week, a month or a quartile. It shows how much longer a team has to accomplish a specific amount of work. Only top contributors are mentioned, the ranking plays a minor role in the concept. As an additional function, a list of open issues can be displayed and information such as the most frequent terms of the block of work is offered. On the bottom, a timeline with colored squares (each representing one interval of time) indicate how often the team has come close to accomplishing the goals that are set. This ideally will trigger the gamification dynamic of 'loss aversion': if there are multiple red blocks indicating not coming close to the goal, the team may want to change that; if there multiple green blocks in a row, the team may be motivated to maintain the high level of completion.

Feedback from concept discussion:

The visualization and accountability features were most favored, and the competition concept was rejected. The worker noted that having a visualization with sorting functionalities would be rewarding and the best way of representing annotation work. Adding an accountability measure would also be a motivating mechanic, as long as the normal workload permits regular work on the annotation task (so in times when for weeks or months there are more urgent projects, time sensitive M&D can be paused to avoid demotivating effects).

The worker also noted that the issue tracker would be a helpful tool - being able to mark something as incomplete or uncertain in a way that other users could see it and try to solve it themselves was a desired feature.

9.3 Recommendations for future work

Summarizing the discussions in the company around both the prototype and the concepts, future work should focus on building a visualization (making use of the 'interaction concepts gamification M&D) which incorporates accountability elements (making use of the 'loss aversion' gamification M&D) and possibly functional improvements as well, such as the issue tracker shown in the communal project concept. These stand out as the most appealing features that a feedback system should have. These could be implemented in the form of an (office) publicly visible visualization, which could show (all the) annotation work done so far, and allow grouping of e.g. related terms or product families. This was also mentioned as a potentially beneficial demonstration product for clients who are visiting the company.

The use of any competitive elements should be omitted until a 'fair game' ground can be guaranteed (meaning that those assigned to the task can offer comparable amounts of time, regularly enough for any performance comparison to be fair). Furthermore, employees stressed that this is an extracurricular task which is not as pressing as day-to-day work, and should thus (continue to) be treated as such. Behaviour may be encouraged with the prototype, but should not be demanded, and should not distract from usual work. As availability to work on the dictionary comes and goes in phases, one should be able to pick up the annotation task after periods of time in which no work took place.

Chapter 10: Conclusion

Returning to the original research question, ‘Can the annotation task be enhanced with gamification?’, the answer is yes. Within the scope of this study, gamification mechanics and dynamics (as synthesized by Thiebes et al.) were applied to a computer based annotation task with a twofold goal: to motivate users to pick up the task on their own (and thus contribute more to the resulting dictionary of product names) by enticing them with challenges and game elements, and to offer an improved user experience around the otherwise monotonous, repetitive and unrewarding annotation task. This was to be achieved by applying gamification, a trending design method in which game elements are applied to non-game contexts [3]. After performing a context analysis and gaining an understanding and library of potential elements to implement in a prototype, appropriate mechanics and dynamics were applied and developed in the form of a feedback interface. This gives users feedback on their latest and total performance in the annotation task, and a visualization of their contributed annotations. In an experiment, the gamified feedback was measured against a non-gamified feedback and the original condition of the annotation task.

In terms of absolute means, the gamified feedback received higher agreement and ratings across almost all statements and attributes. However, significance tests showed that in most cases the differences were not significant. Thus, for most of the statements and attributes one can not claim that the gamified feedback was better or worse than either of the other feedback versions. There are however some significant differences: with the gamified feedback, users felt significantly more challenged to do more annotation work and looked significantly more forward to seeing the results of their work (in comparison to the non-gamified feedback). Furthermore, they were significantly more satisfied with the gamified feedback and found it significantly more aesthetically pleasing than the non-gamified feedback. These results indicate that, in these aspects, the gamification was successful in improving user satisfaction, engagement and motivation.

Within the scope of this work, the test results can only reflect the reception in the group it was tested in: primarily male students, recruited from university. As it was not feasible to develop and test with real users in the real workplace over a longer amount of time, compromises were made and the investigation into some potentially effective features was not possible. During presentation at the workplace, some of these features were discussed, and deemed promising, given that they be developed and tested adequately. Furthermore, some of the positively received features of the tested prototype (the challenges and user levels) were not as highly valued by the interviewed workplace users as test users.

Nonetheless, the developed prototype demonstrated that gamification can be applied to the annotation task in a meaningful, workplace appropriate way. Given appropriate development, it has the potential to fulfill its promises: enticing users to voluntarily dedicate more of their time and effort to the task.

Chapter 11: Future work

The developed prototype is only one of many possible ways in which gamification can be applied to this annotation task, and the same goes for each of the used interface elements - experimenting with each element in the system and its way of increasing motivation and user satisfaction may produce more effective results. This also goes for unexplored combinations between the various mechanics and dynamics that may have synergistic effects. Furthermore, the system may be expanded with new mechanics and dynamics that were not tested here, such as time pressure, reminders or the social influences outlined by Thiebes et al., such as reputation or conforming behaviour.

While the prototype developed in the scope of this project was focused on single user experience, in the future multiple users may be assigned to the task, and mechanics and dynamics that make use of social influences should not be ignored. As expressed early by the client and later during the company visit, the concept of a (workplace) publicly visible data visualization of the annotation dictionary so far could be motivating, especially if progress and outstanding contributions can be highlighted. As this particular workplace is (whenever possible) occupied with client work that is more important, the annotation task is an extracurricular activity and must continue to be treated as such. Through development and testing at the workplace, the line between intrusive and enticing must be distinguished and respected, or an intervention, such as a publicly visible leaderboard or progress tracker may have counterproductive effects. Thus, future investigation may ask ‘How can an office-visible implementation boost productivity without interfering with everyday duties?’, or ‘What level of participation can be evoked from employees [on a regular basis]?’.

Another aspect of the task that could not be investigated was maintaining and increasing the quality of annotations contributed. In testing with non-experts, this was not an option, and most implemented mechanics and dynamics focused on quantity. As the data being produced must be as accurate as possible, this should be encouraged with appropriate mechanics and dynamics. One worker at the client company expressed that an issue-tracker may be helpful in allowing users to highlight an entry or section of the database for review without interrupting or delaying their annotation session. That way it can be returned to later, and other workers can help, if needed. Thus, an appropriate research question to investigate this may be ‘How can workflow be supported for multiple users to increase quantity while maintaining high quality?’. Relevant examples of successful peer-review mechanics within gamified language resource software are mentioned in the state of the art (section 3.5) and can help guide a design process.

References

- [1] Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68-78.
<http://dx.doi.org/10.1037/0003-066X.55.1.68>
- [2] Ryan, R. and Deci, E. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, 25(1), pp.54-67.
- [3] Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From Game Design Elements to Gamefulness: Defining "Gamification." *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '11*, 2425.
<https://doi.org/10.1145/1979742.1979575>
- [4] <http://www.bunchball.com/gamification>, taken 12/10/2017
- [5] Deterding, S. (2012). Gamification: designing for motivation. *interactions*, 19(4), 14-17. DOI: 10.1145/2212877.2212883
- [6] Thiebes, S., Lins, S., & Basten, D. (2014). Gamifying information systems-a synthesis of gamification mechanics and dynamics.
- [7] J. Klasen. "Employees' Experiences and Perceptions of Work Gamification," Pepperdine University, Ann Arbor, 2016.
- [8] Jovanovic, Mladjan. "Gamifying knowledge maintenance." (2015).
- [9] Mader, A. H., & Eggink, W. (2014). A design process for creative technology. *The Design Society*.
- [10] Hunicke, R., Leblanc, M. and Zubek, R. (2004). MDA: A Formal Approach to Game Design and Game Research. In *Proceedings of the Challenges in Games AI Workshop, Nineteenth National Conference of Artificial Intelligence*, p. 1, AAAI, San Jose, USA.
- [11] Guillaume, Bruno, Karën Fort, and Nicolas Lefebvre. "Crowdsourcing complex language resources: Playing to annotate dependency syntax." *International Conference on Computational Linguistics (COLING)*. 2016.

- [12] Chamberlain, Jon, Massimo Poesio, and Udo Kruschwitz. "Phrase detectives: A web-based collaborative annotation game." *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*. 2008.
- [13] Venhuizen, Noortje J., et al. "Gamification for word sense labeling." *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Short Papers*. 2013.
- [14] Lazar, Jonathan, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.
- [15] Harper, Ben D., and Kent L. Norman. "Improving user satisfaction: The questionnaire for user interaction satisfaction version 5.5." *Proceedings of the 1st Annual Mid-Atlantic Human Factors Conference*. 1993.
- [16] Lund, Arnold M. "Measuring usability with the use questionnaire 12." *Usability interface* 8.2 (2001): 3-6.
- [17] Brooke, John. "SUS-A quick and dirty usability scale." *Usability evaluation in industry* 189.194 (1996): 4-7.
- [18] Donker, Afke. *Human factors in educational software for young children*. Diss. 2005.

Appendix

Pen and paper low-level questionnaire

During development, a pen-and-paper prototype was developed and evaluated (Report chapter 4), which can be seen in its digital form in figure A1:

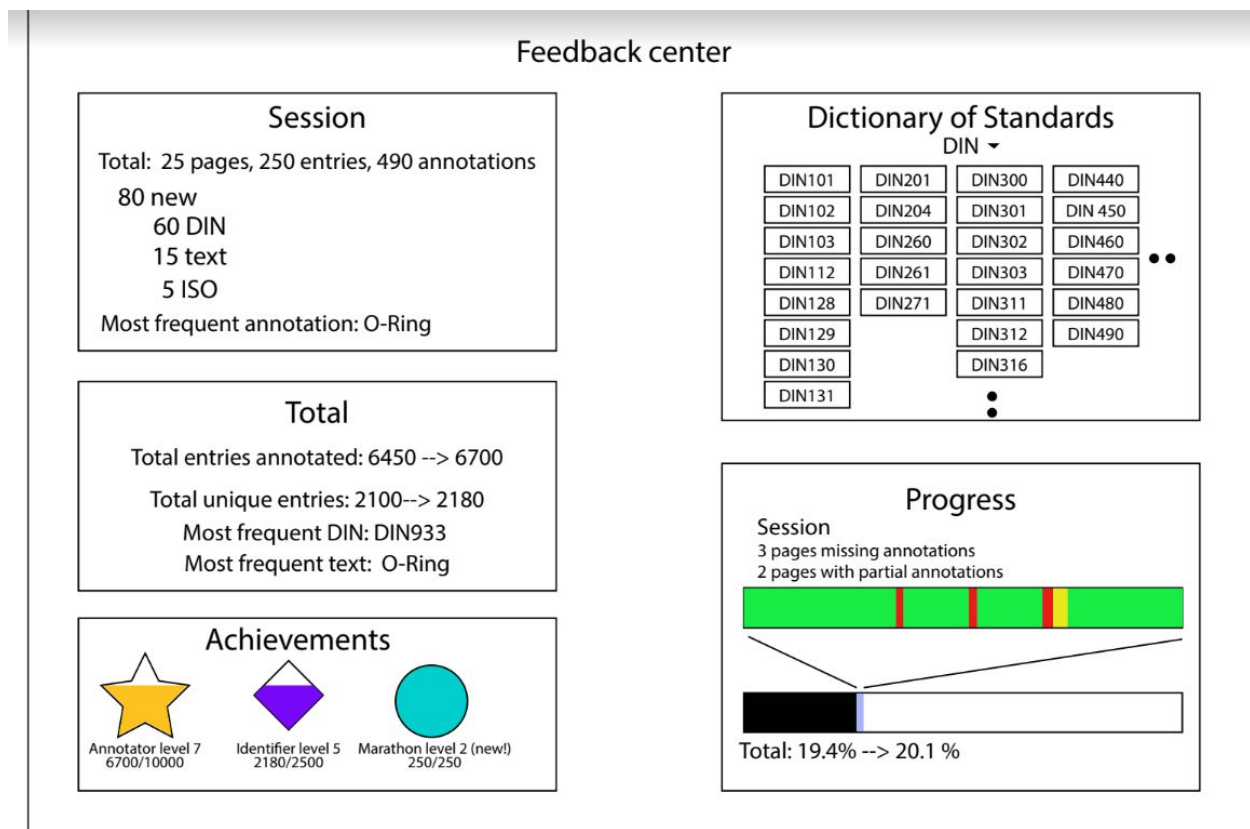


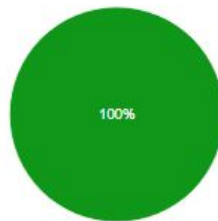
Fig. A1. Pen-and-paper prototype.

As described in the report, users were first shown only the left side ('Section 1'), then the right side ('Section 2'). Feedback was collected with a questionnaire - the questions and their answers can be seen here:

Section 1: Session feedback

The feature is insightful

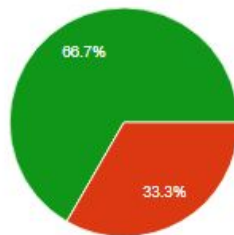
3 responses



- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

The feature is interesting

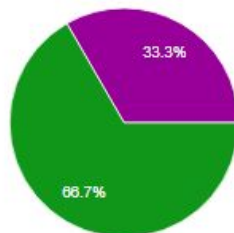
3 responses



- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

The feature is relevant & appropriate

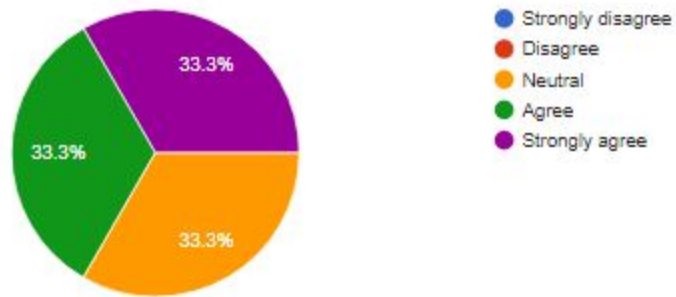
3 responses



- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

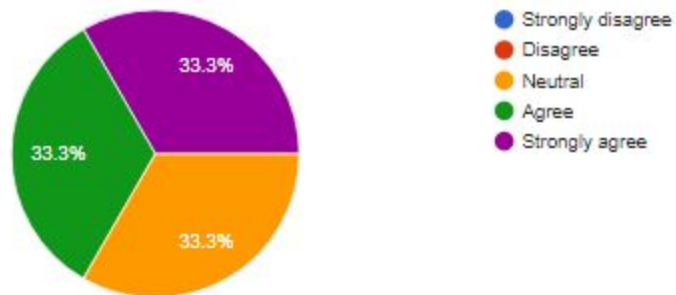
The feature is easy to understand

3 responses



The feature reflected my work efforts

3 responses



Was the feature missing something?

3 responses

no

maybe take time needed into account

I think this section needs a line that states whether you made a common mistake or all your fellow workers see you wrong on that.

Comments or questions on the feature

2 responses

nice insight

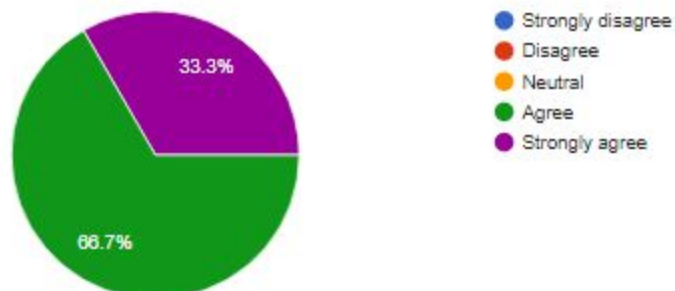
Also I think the most frequent isnt too important.

#

Section 1: Total feedback

The feature is insightful

3 responses



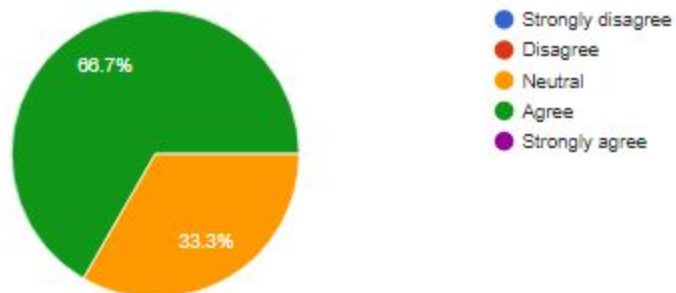
The feature is interesting

3 responses



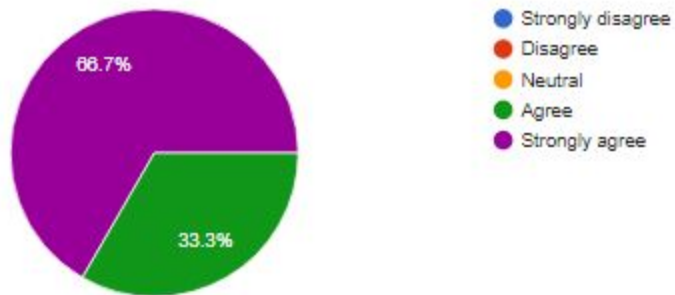
The feature is relevant & appropriate

3 responses



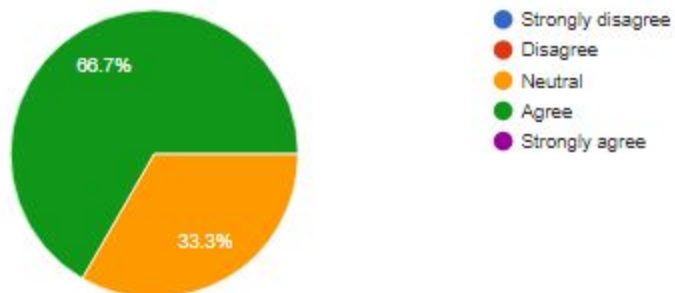
The feature is easy to understand

3 responses



The feature reflected my work efforts

3 responses



Was the feature missing something?

1 response



Comments or questions on the feature

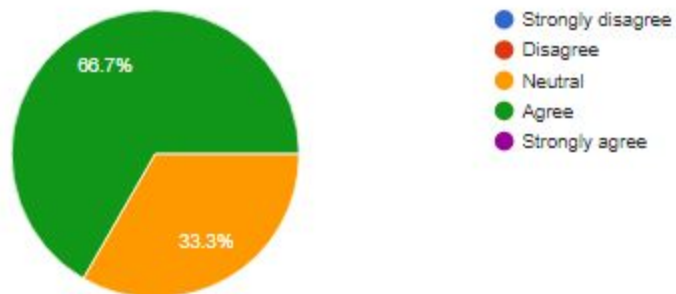
0 responses

No responses yet for this question.

Section 1: Achievements

The feature is insightful

3 responses



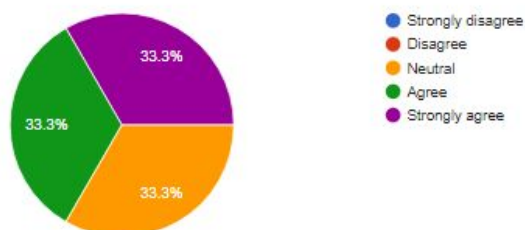
The feature is interesting

3 responses



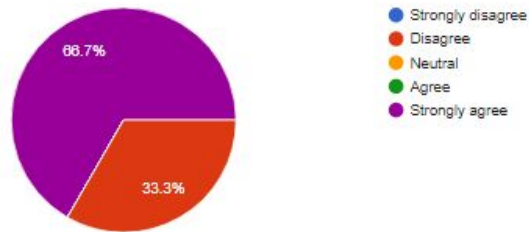
The feature is relevant & appropriate

3 responses



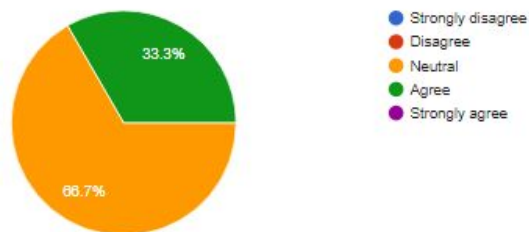
The feature is easy to understand

3 responses



The feature reflected my work efforts

3 responses



Was the feature missing something?

0 responses

No responses yet for this question.

Comments or questions on the feature

2 responses

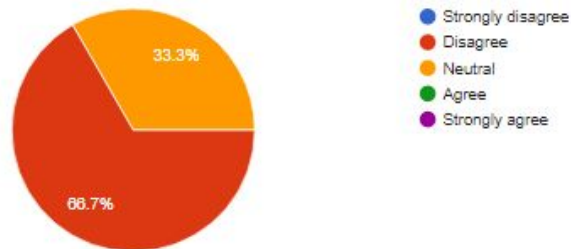
in my case 41/50 annot. means I identified 41 of possible 50 product names?
19/50 identifier, what does identifier mean in this context? by looking to the session block I figured it describes the amount of data entries
marathon = streak?

I would bind the circle element to the top section (Session) design wise.

Section 2: Dictionary

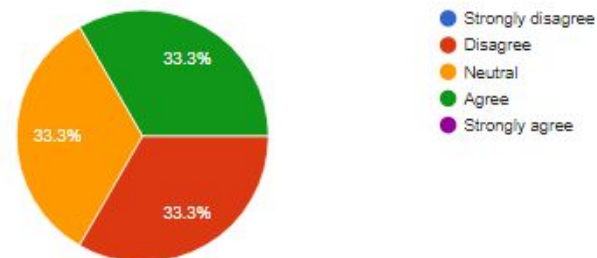
The feature is insightful

3 responses



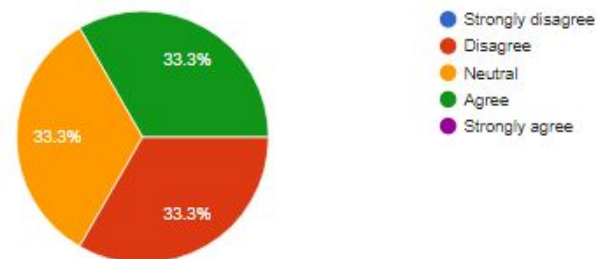
The feature is interesting

3 responses



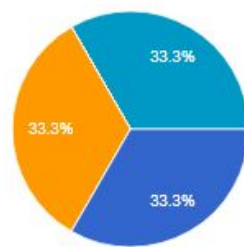
The feature is relevant & appropriate

3 responses



The feature is easy to understand

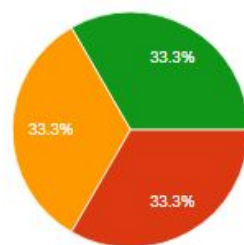
3 responses



- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree
- if the title would be different it would be easy to understand

The feature reflected my work efforts

3 responses



- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

Was the feature missing something?

1 response

Gamification

Comments or questions on the feature

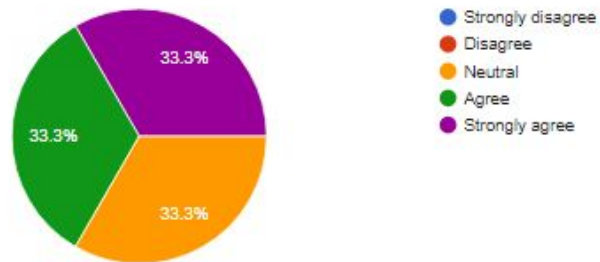
1 response

This section could be built up like a collection. So people can collect seldom item/DINs. But it seems not as relevant as the other sections.

Section 2: Progress

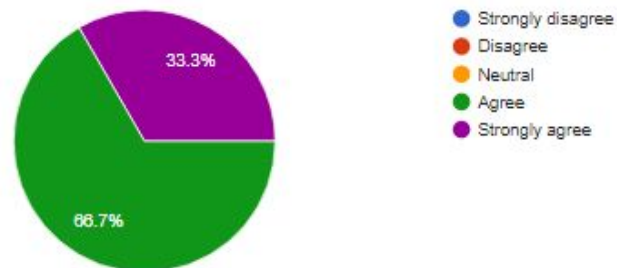
The feature is insightful

3 responses



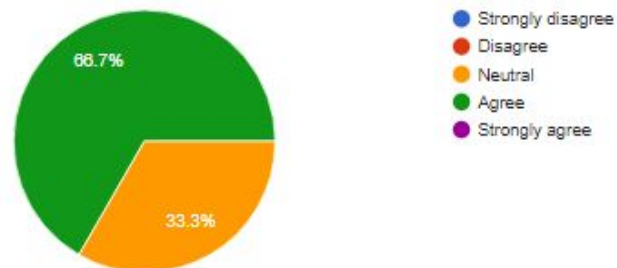
The feature is interesting

3 responses



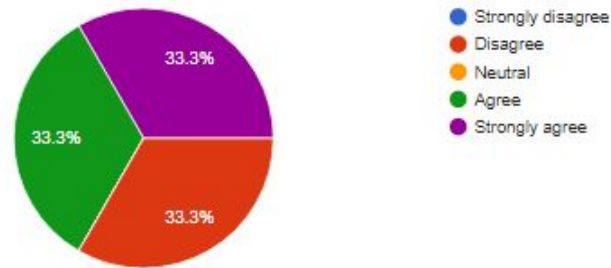
The feature is relevant & appropriate

3 responses



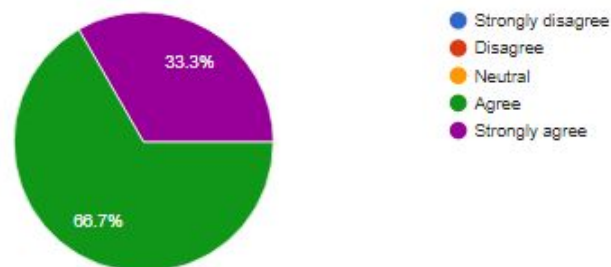
The feature is easy to understand

3 responses



The feature reflected my work efforts

3 responses



Was the feature missing something?

1 response

no

Comments or questions on the feature

1 response

what is good, bad and mediocre in this context? i'm guessing it evaluates how many product names I identified..

General

Which was your favorite feature, and why?

3 responses

Achievements, because it is motivating to level up

the second one. to me, this type of work was new. you could easily forget to highlight something, because the product names are very homogeneous. the second feature seems to outline, which product names are most recognizable for a worker.

I actually got two favorite sections: the Achievements and the Progress. I like them because they offer motivation and visual feedback.

Did you have a least favorite feature, and why?

2 responses

the fourth, because i don't understand it

Yes, the Dictionary of Standards. I think the name is misleading. Also it does not give me more relevant insight.

Were there any features that were superfluous?

3 responses

Maybe the Dictionary of Standards I'm not sure, then again it could be useful in a way to learn about the different DIN's and ISO's and thus being able to recognize them faster

no.

I guess the Dictionary might be.

As a feedback and progress system, is it missing anything?

3 responses

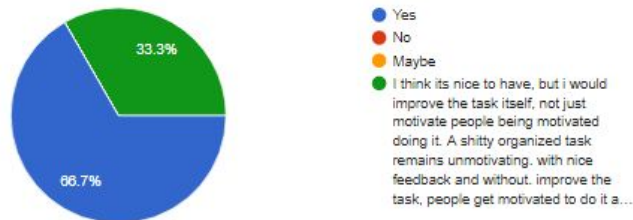
maybe statistics like the amount of time spend working

maybe display time management/time needed for a worker

I think a streak feature might reward people very good for being accurate. Several multipliers could be involved, for lots of rewards :)

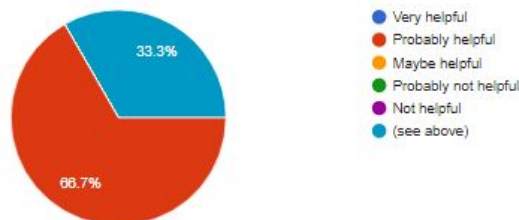
Is this kind of feedback appropriate for such a task?

3 responses



If you were tasked with doing this work regularly, for an open ended amount of time, can you see this system being helpful?

3 responses



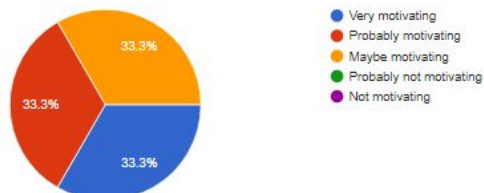
Which of the features, given the premise of doing this work long term, would you find most interesting, and why?

3 responses

Achievements and Progress, because it shows a long-term progress and 'leveling up' which feels rewarding
second, already mentioned the reason
Progress

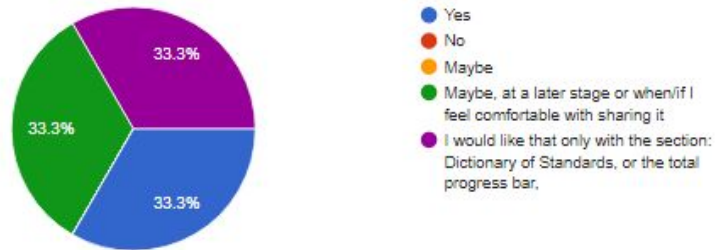
If you were tasked with doing this work regularly, for an open ended amount of time, can you see this system being motivating?

3 responses



Given that you would be doing this work in an office environment, would you be interested in showing your colleagues your progress (via a publicly visible visualization)?

3 responses



Any comments on the system, improvements, open questions...

1 response

You have to be careful, that the system is designed in a way, so that it does not consume the time of the workers and also is not too much of a feeling of superiority.

Prototype test: all questions and answers

In the following, the test results from the research experiment described in report chapter 7 are given in detail.

Averages of answers on questions of section 2 (statements, likert scale) and 3 (feedback evaluation)

Rows: C1 (file feedback), C2 (non-gamified feedback), C3 (gamified feedback)

Columns: statements (1-15, score 1: completely disagree, 7: completely agree) and feedback attributes (16-22, score 1: negative attribute, 7: positive attribute),

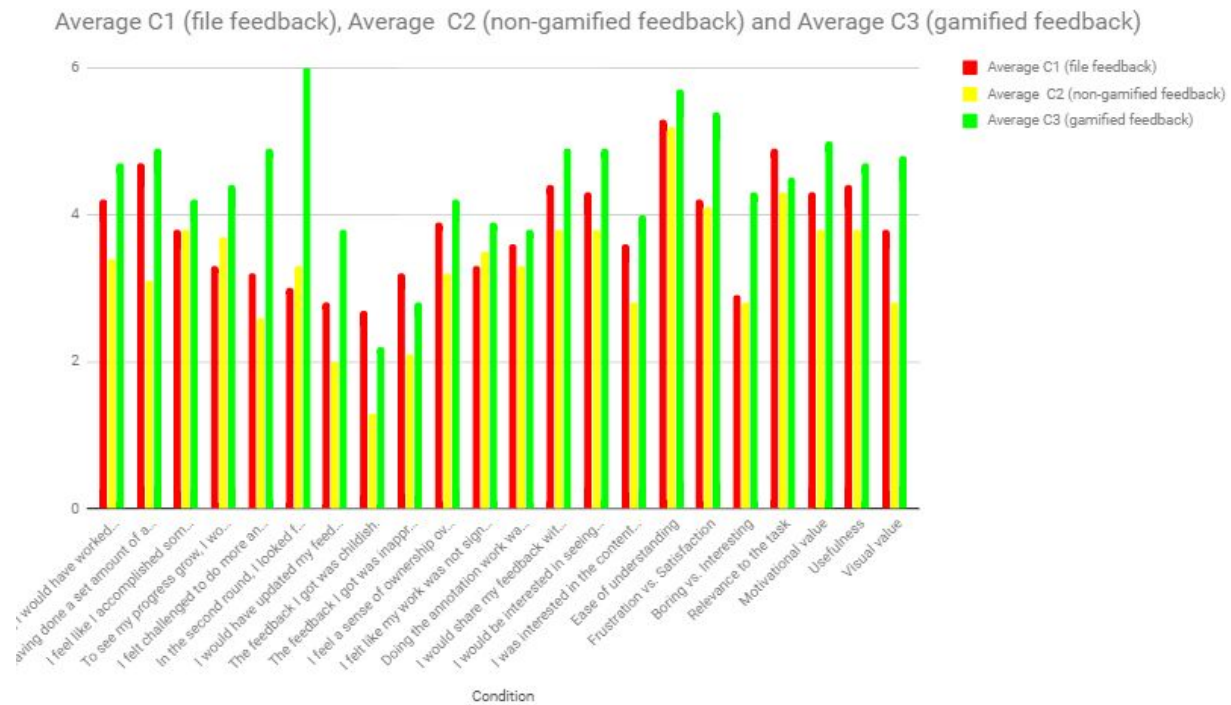
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		16	17	18	19	20	21	22
C1	4.2	4.7	3.8	3.3	3.2	3	2.8	2.7	3.2	3.9	3.3	3.6	4.4	4.3	3.6		5.3	4.2	2.9	4.9	4.3	4.4	3.8
C2	3.4	3.1	3.8	3.7	2.6	3.3	2	1.3	2.1	3.2	3.5	3.3	3.8	3.8	2.8		5.2	4.1	2.8	4.3	3.8	3.8	2.8
C3	4.7	4.9	4.2	4.4	4.9	6	3.8	2.2	2.8	4.2	3.9	3.8	4.9	4.9	4		5.7	5.4	4.3	4.5	5	4.7	4.8

Questions/feedback attributes of each column

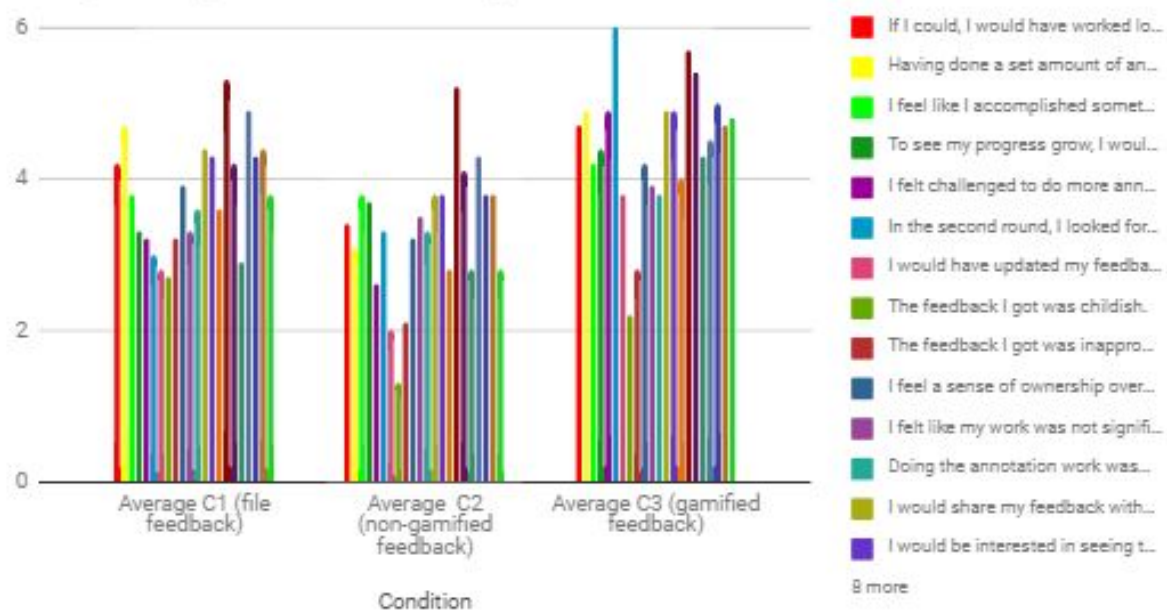
1. If I could, I would've worked longer to complete a set amount of annotations (e.g. 50, 100, 200, 500 etc.).
2. Having done a set amount of annotations (e.g. 50, 100, 200, 500) would give me satisfaction.
3. I feel like I accomplished something.
4. To see my progress grow, I would often pick up the annotation task often inbetween other tasks.
5. I felt challenged to do more annotation work.
6. In the second round, I looked forward to seeing the results of my work.
7. I would've updated my feedback more often, if I could've.
8. The feedback I got was childish.
9. The feedback I got was inappropriate for a workplace, such as an office.
10. I feel a sense of ownership over the work I've done.
11. I feel like my work was not significant.
12. Doing the annotation work was satisfying.
13. I would share my feedback with others who were tasked to do the same work.
14. I would be interested in seeing the feedback of others such as colleagues, who also did this annotation work.
15. I was interested in the content of the database.

-
16. Ease of understanding (1: very difficult to understand, 7: very easy to understand)
 17. Frustration vs. satisfaction (1: very frustrating to see, 7: very satisfying to see)

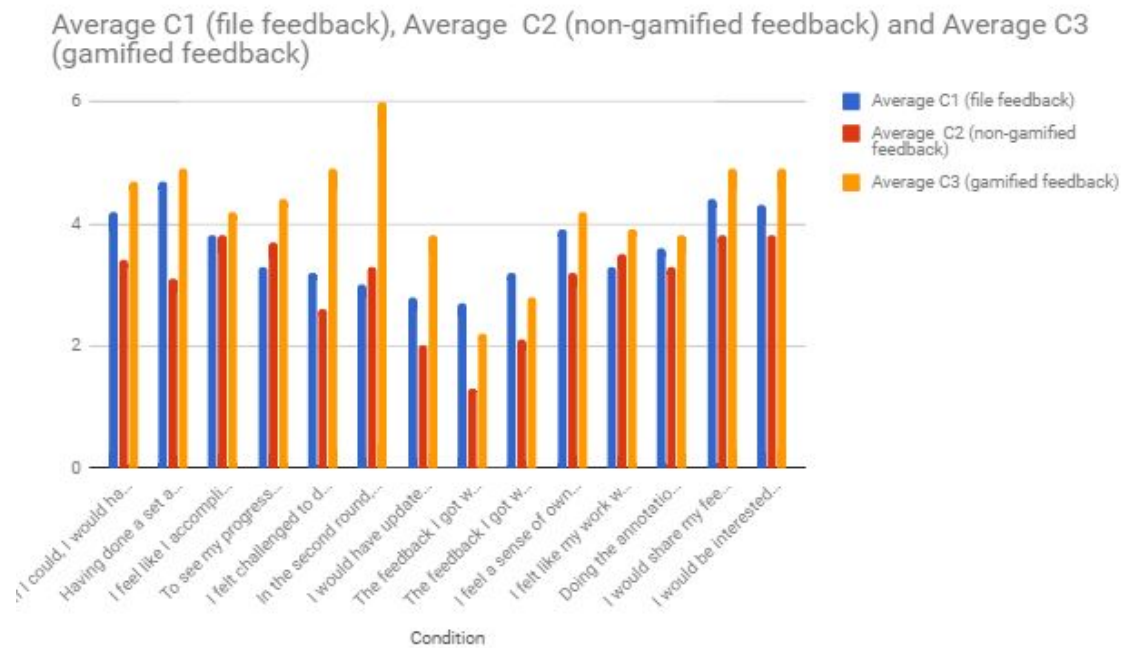
18. Boring vs. interesting (1: very boring/dull, 7: very interesting)
19. Relevance to the task (1: completely irrelevant to the annotation task, 7: highly relevant to the annotation task)
20. Motivational value (1: very demotivating, 7: very motivating)
21. Usefulness (1: completely useless, 7: very useful)
22. Visual value (1: ugly, aesthetically displeasing, 7: good looking, aesthetically pleasing)



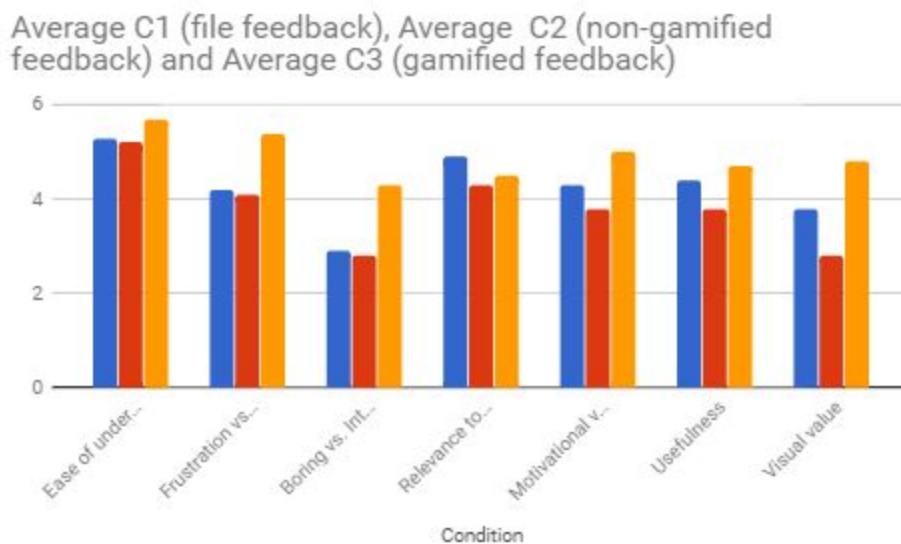
Average C1 (file feedback), Average C2 (non-gamified feedback) and Average C3 (gamified feedback)



Section 2 summary (statements, likert scale 1-7 with 1 is completely disagree and 7 completely agree)

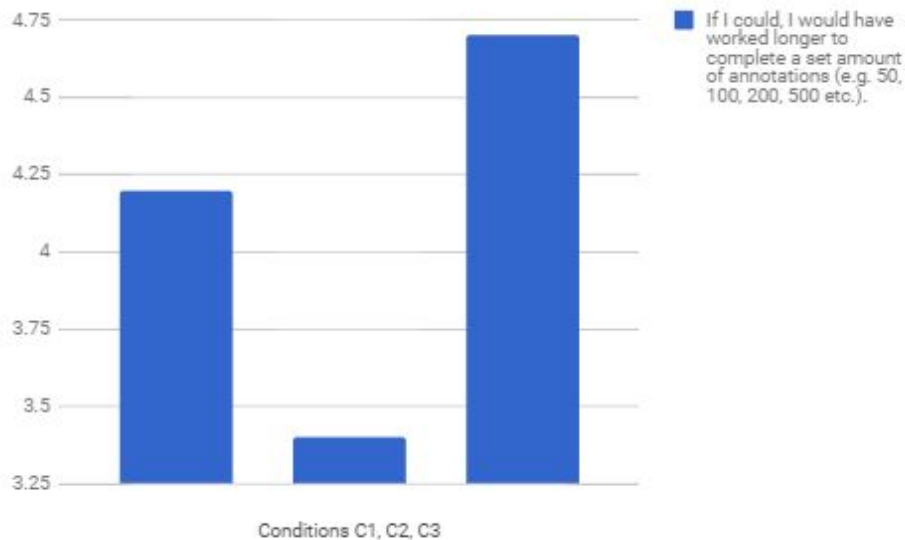


Section 3 summary (feedback evaluation, attributes 1-7 with 1 is negative attribute and 7 is positive attribute)

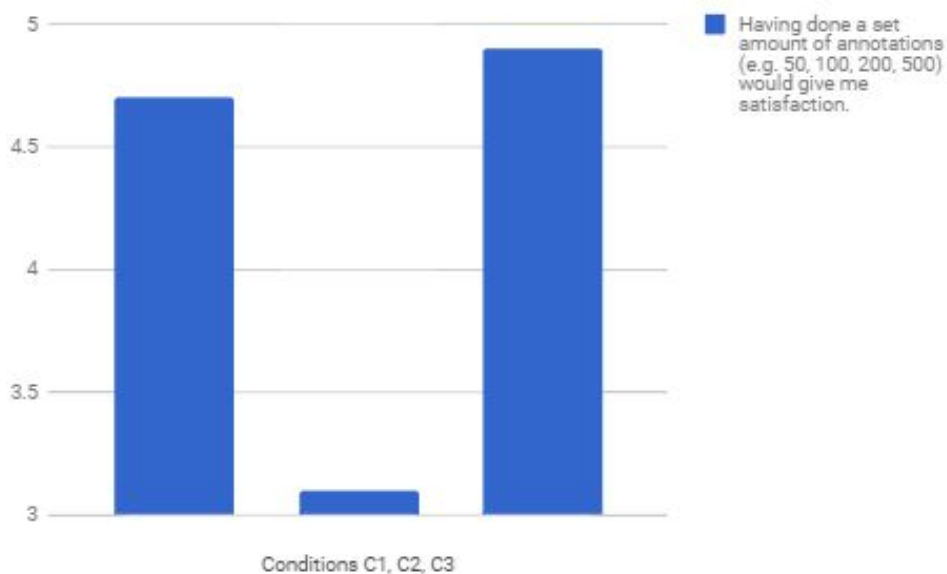


Single questions, section 2 (Statement questions, likert scale 1-7 with 1 completely disagree and 7 completely agree) averages

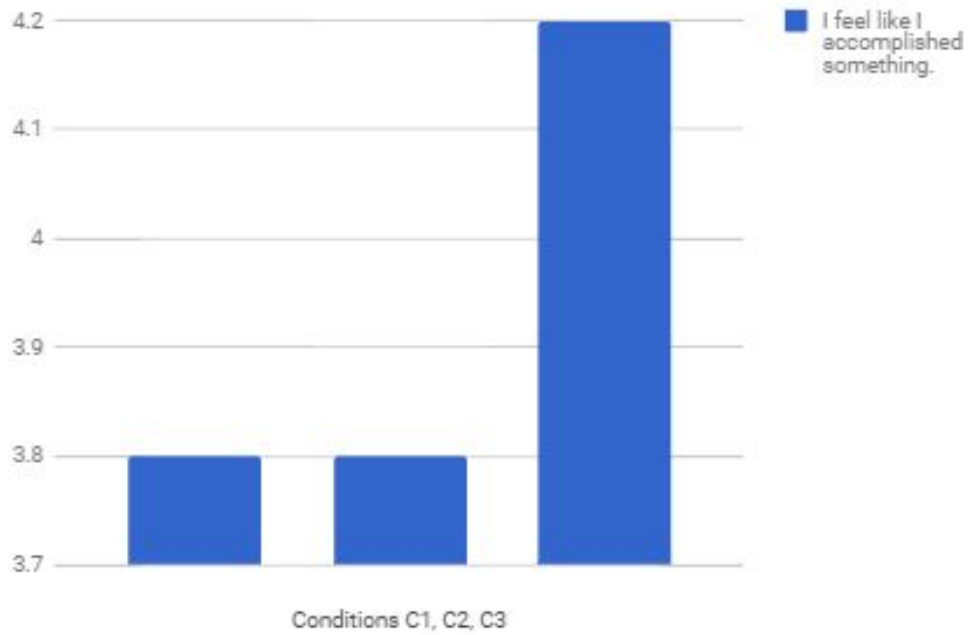
1. If I could, I would've worked longer to complete a set amount of annotations (e.g. 50, 100, 200, 500 etc.).



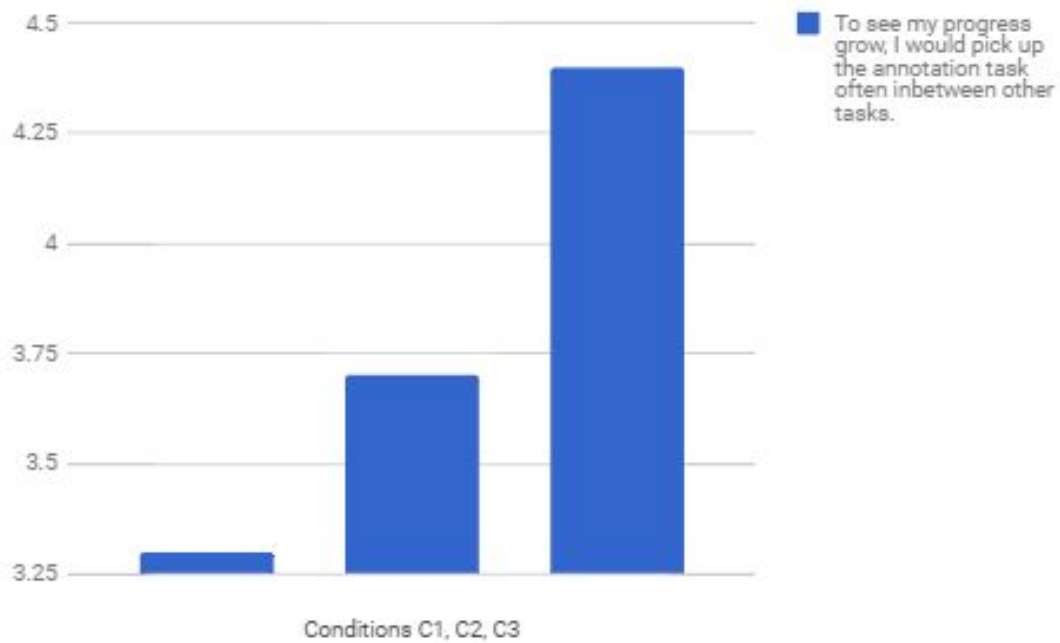
2. Having done a set amount of annotations (e.g. 50, 100, 200, 500) would give me satisfaction.



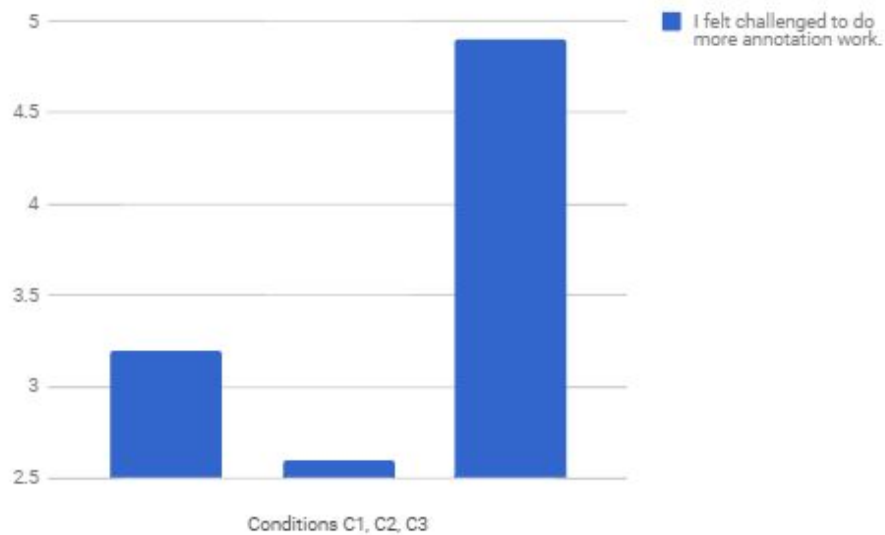
3. I feel like I accomplished something.



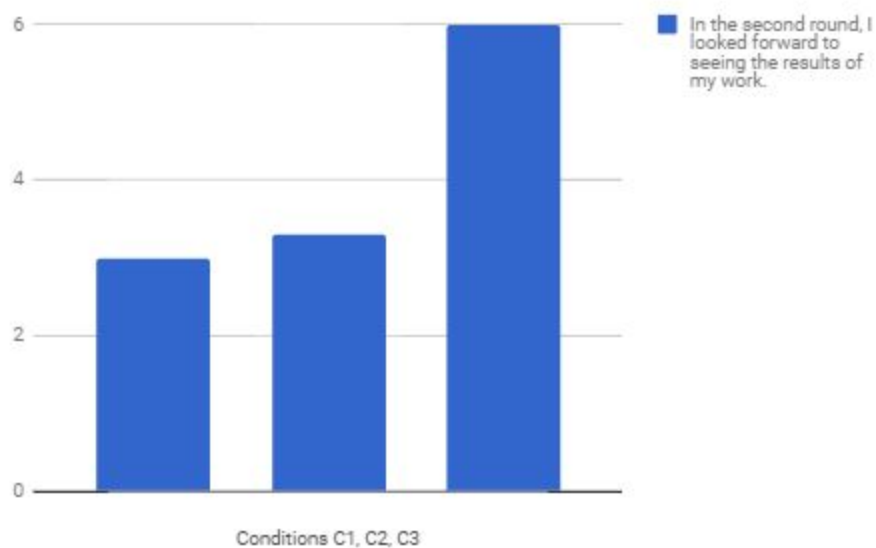
4. To see my progress grow, I would often pick up the annotation task often inbetween other tasks.



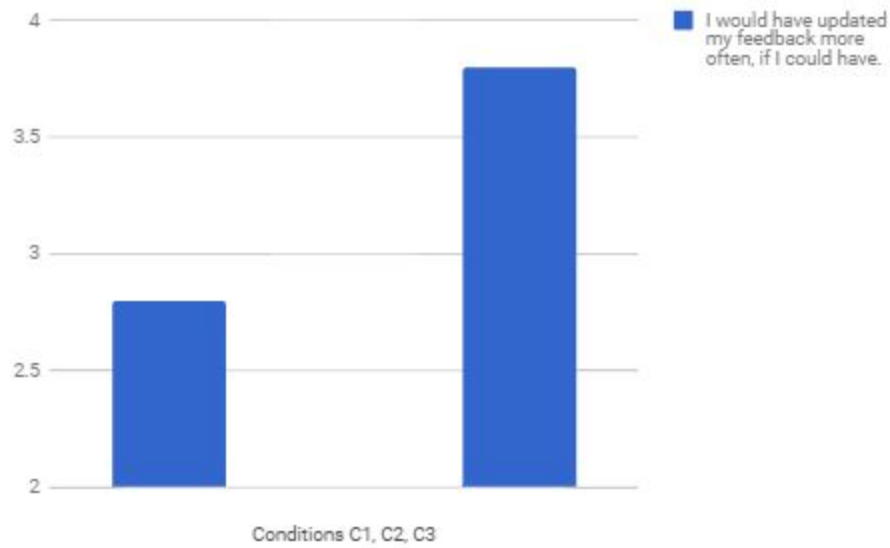
5. I felt challenged to do more annotation work.



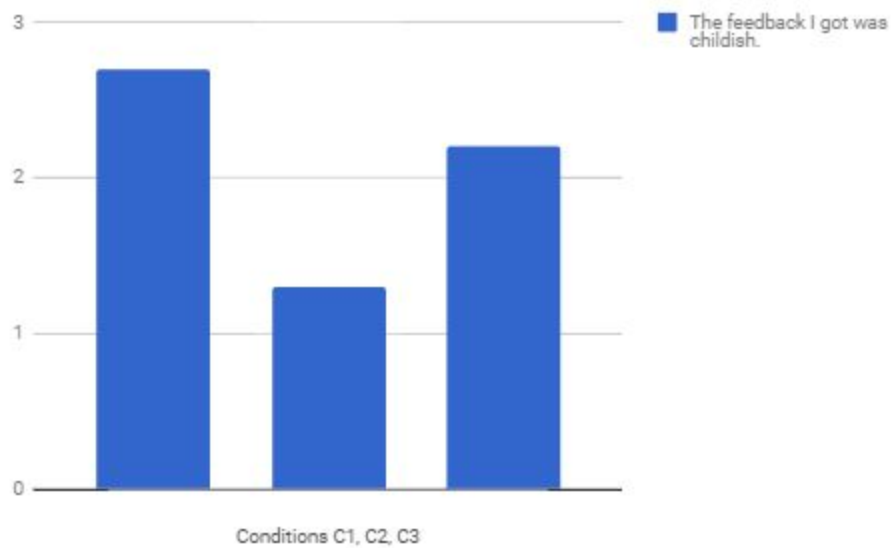
6. In the second round, I looked forward to seeing the results of my work.



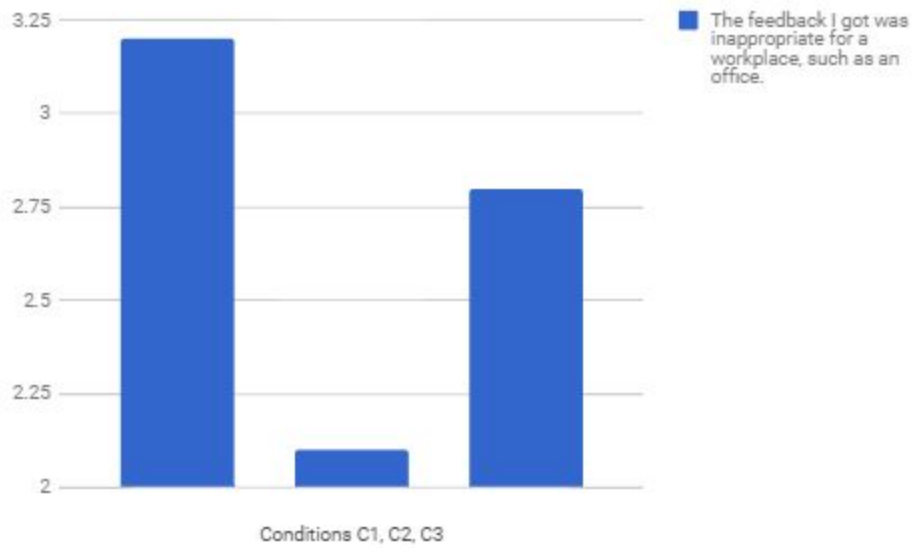
7. I would've updated my feedback more often, if I could've (C2 = 2).



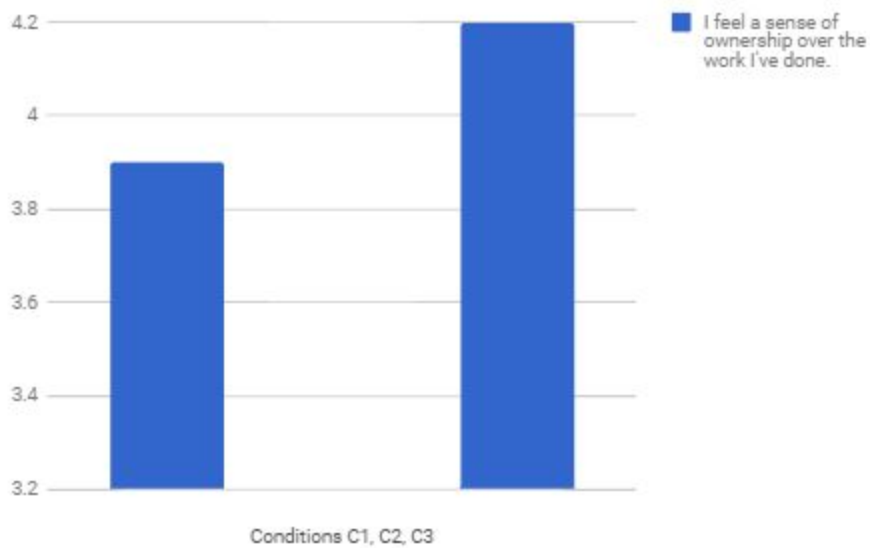
8. The feedback I got was childish.



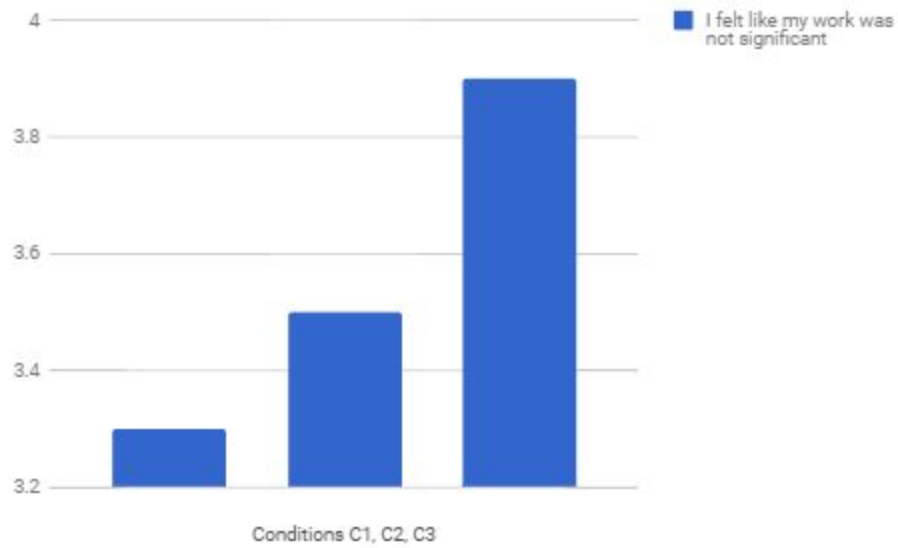
9. The feedback I got was inappropriate for a workplace, such as an office.



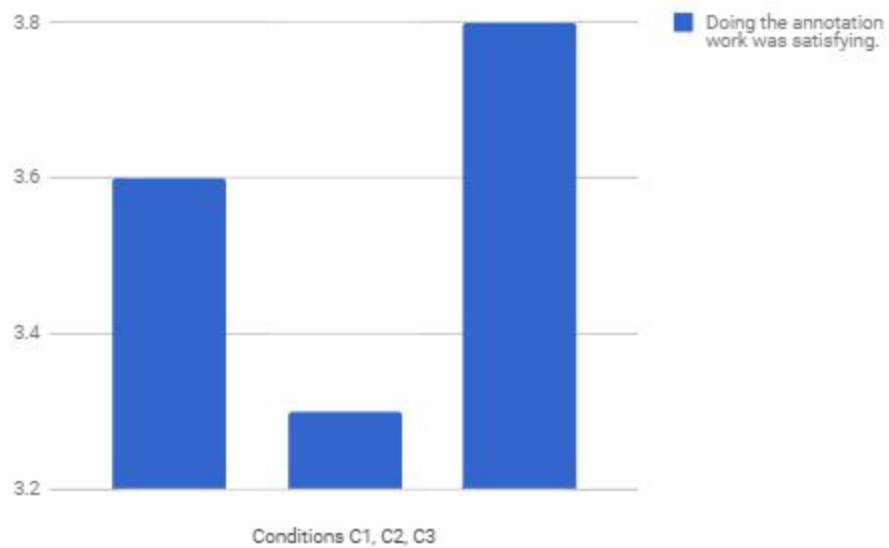
10. I feel a sense of ownership over the work I've done (C2 = 3.2).



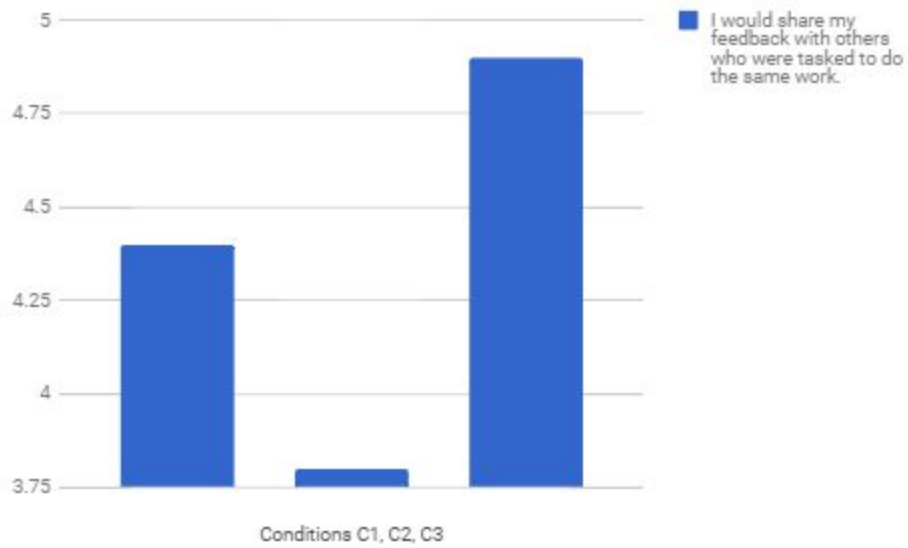
11. I feel like my work was not significant.



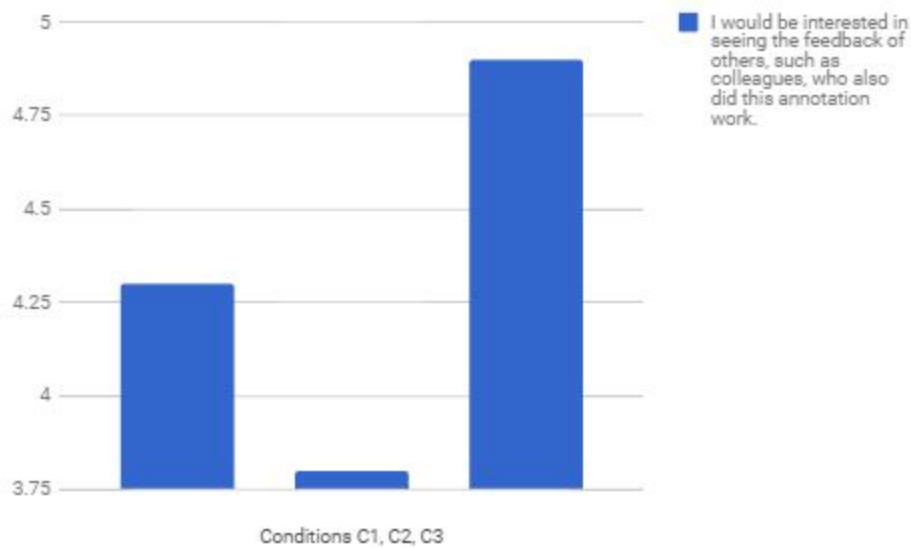
12. Doing the annotation work was satisfying.



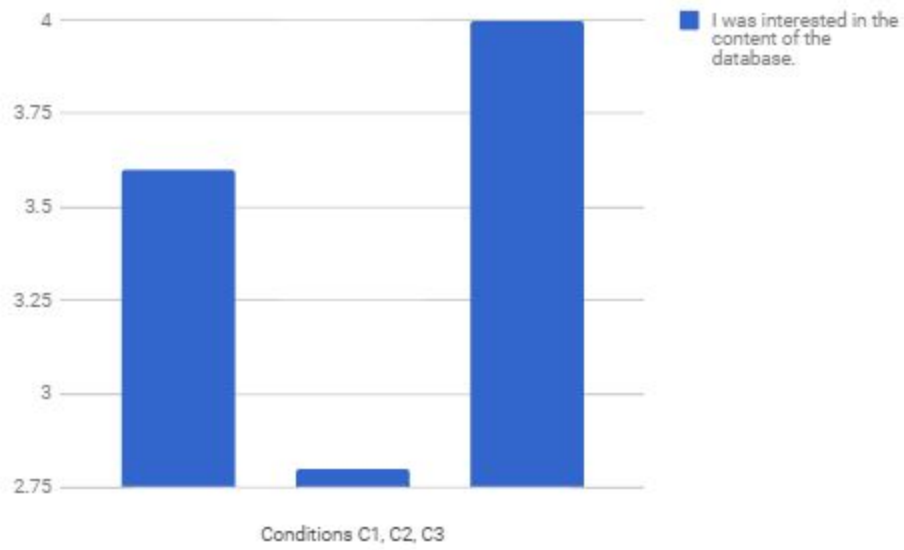
13. I would share my feedback with others who were tasked to do the same work.



14. I would be interested in seeing the feedback of others such as colleagues, who also did this annotation work.

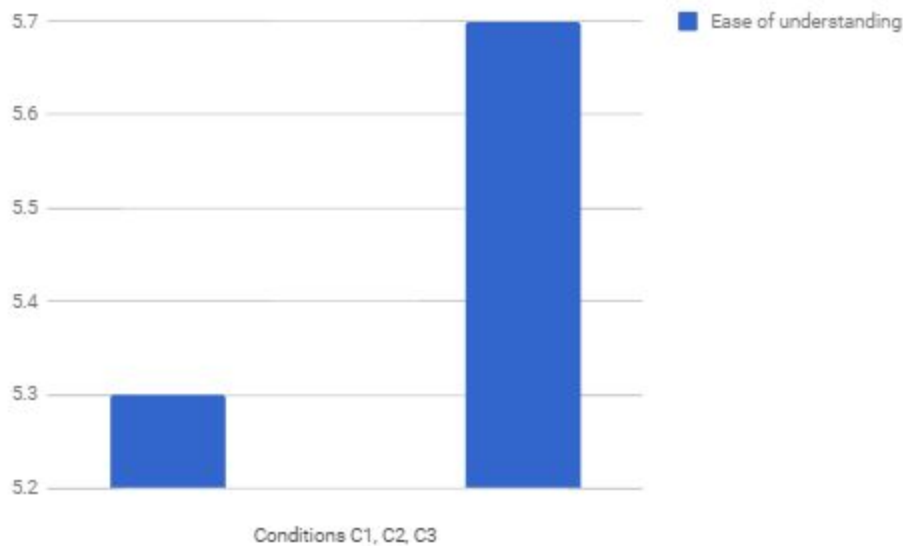


15. I was interested in the content of the database.

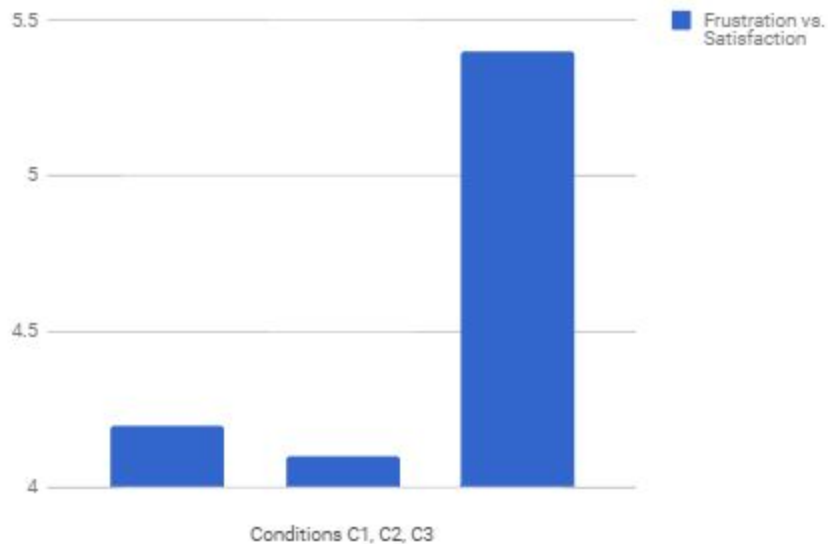


Single attributes, section 3 (feedback evaluation, attributes 1-7 with 1 is negative attribute and 7 is positive attribute)

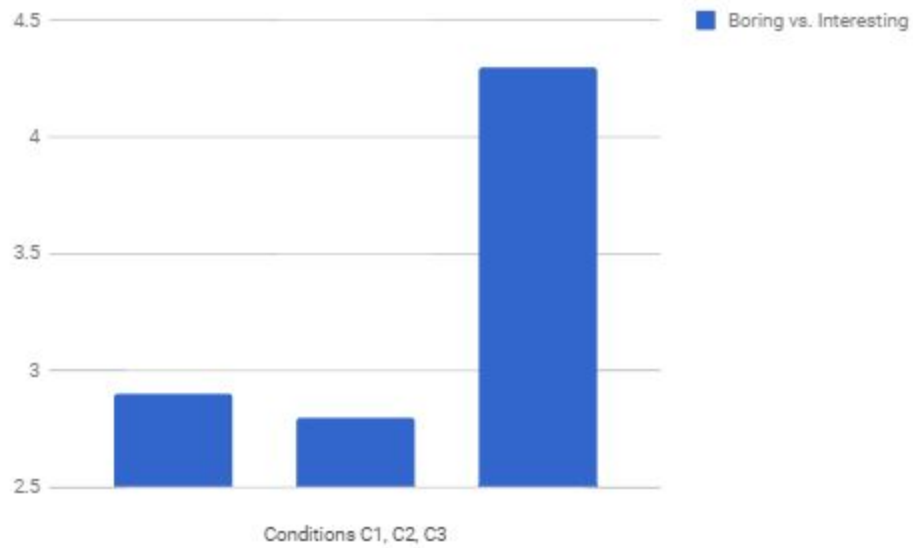
16. Ease of understanding (1: very difficult to understand, 7: very easy to understand) (C2 = 5.2)



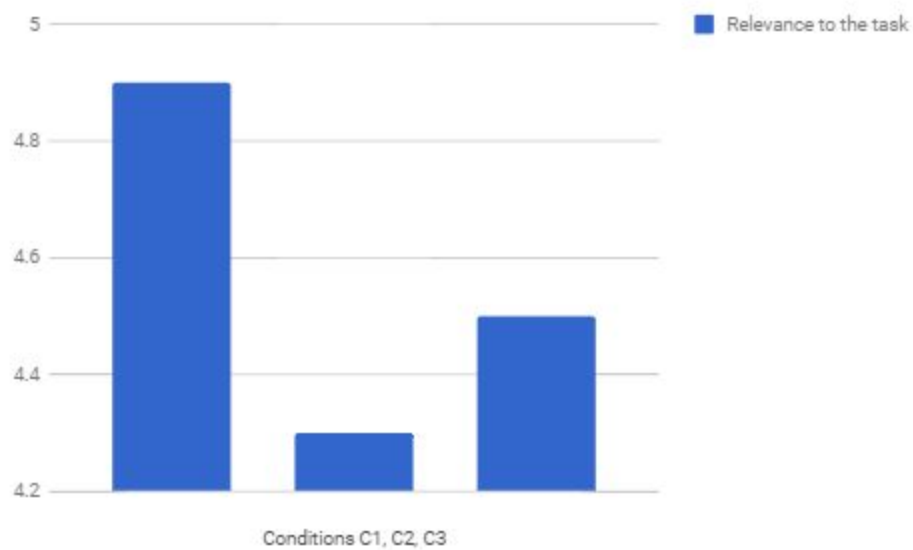
17. Frustration vs. satisfaction (1: very frustrating to see, 7: very satisfying to see)



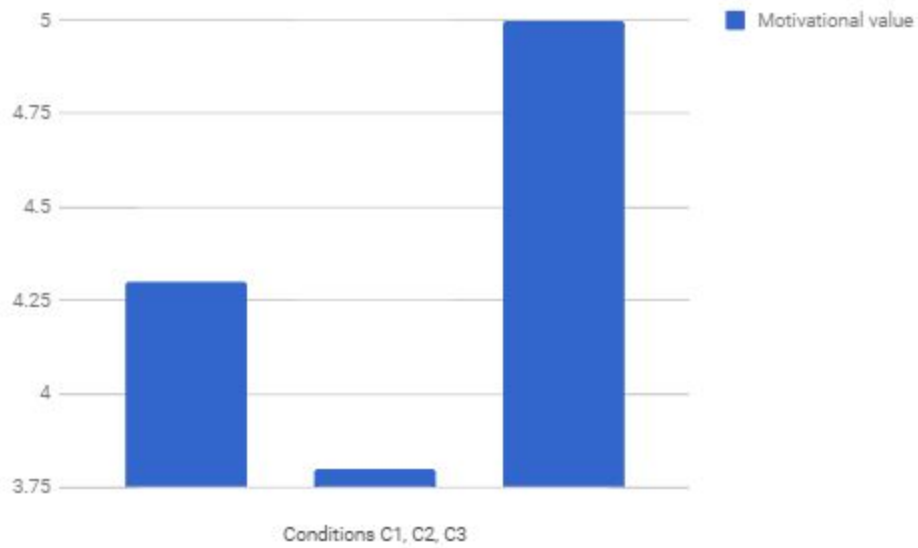
18. Boring vs. interesting (1: very boring/dull, 7: very interesting)



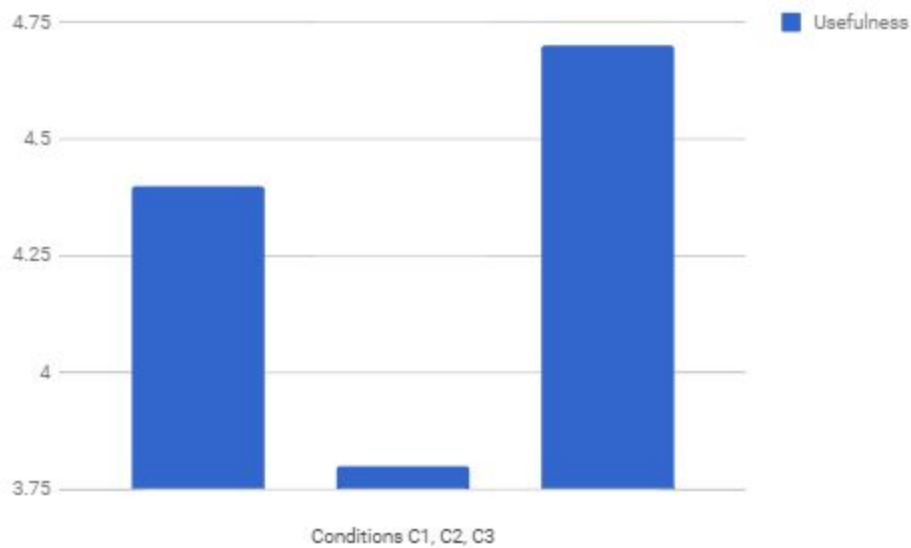
19. Relevance to the task (1: completely irrelevant to the annotation task, 7: highly relevant to the annotation task)



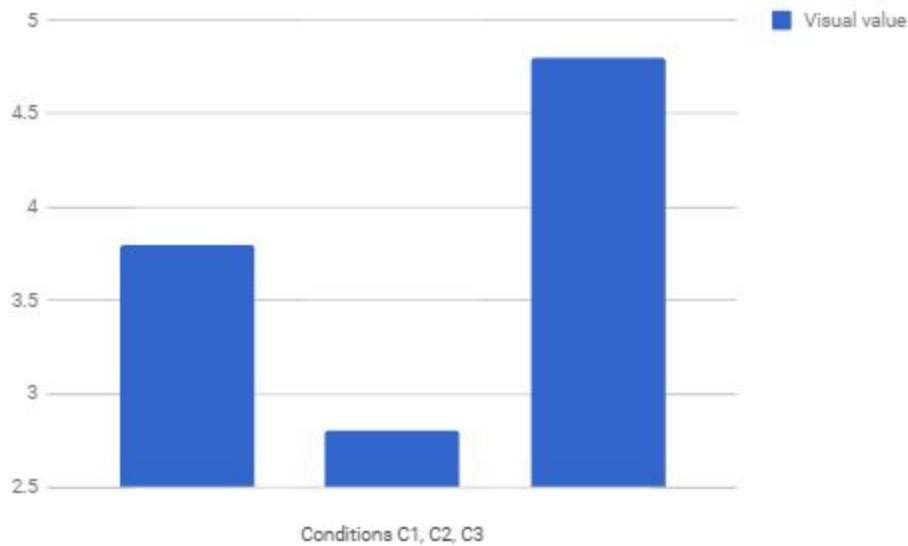
20. Motivational value (1: very demotivating, 7: very motivating)



21. Usefulness (1: completely useless, 7: very useful)



22. Visual value (1: ugly, aesthetically displeasing, 7: good looking, aesthetically pleasing)



Section 4: What did you like/dislike about the feedback?

Condition C1 (file feedback), rows are users, columns are like/dislike about feedback

What did you like about the feedback?	What did you not like about the feedback?
The clear view of all the items I annotated, which helps comparing different annotation to see differences or similarities	I did not know if I could edit it, which could be useful and better than coming back to the list webpage to change an annotation
It is nicely ordered, gives structure	very dry,
It showed me a summary of the annotations I found	The feedback did not show me product names that I did not identify or if my annotations were correct/wrong. After receiving the feedback I did not know If I did a good job on the annotation work.
It was good to see what has happened and what the work that was made created within the feedback sessions. In a way, one could conduct and see the accomplishment that one has made throughout the time the annotation was taking place.	I believe that the feedback does not need to be so regular. As it may distract from the actual task, as with the time you use for the back, it may take away your concentration and motivation for doing the task.

Clarity	Complicated language
Organized and neat	some numbers I could not figure out their meaning f.e. 92 14
It was systematically ordered and thus easily understandable	It didn't give me any insights on earlier progress I've made which in turn could've helped me to do a better job in future iterations. on top it could've been motivating to see development during the iterations. more vs less progress made ecetera
visually simple, good to follow	too little information
it seemed very structured	i didnt really knew what to do with it
That you could see how much work you have done	It was hard to see how much you actually have done, and it was just a bunch of words I'd already seen

Condition C2 (non gamified feedback), rows are users, columns are like/dislike about feedback

What did you like about the feedback?	What did you not like about the feedback?
I liked seeing how many things i was able to click in the given amount of time	the list of product names seemed to be a bit meaningless especially because i dont know what all the products are and why I should care about them
It showed me how many times I found a certain product which reaffirmed the method I used to deduce a product name.	It was a long list and maybe a graph of some sort would have been easier to take in at once instead of a long list of numbers.
it was short and pure.	it was too raw, not very appealing nor challenging
It faste gives you a feeling that you understand something you had no idea beforehand and it helps you to see minor mistakes quickly.	that I could not see if I were doing it right or wrong.
to see which articles I annotated the most.	Nothing at all
Shows you how many annotations you made, increases fast	Maybe should have had some sort of graphs to visualize progress.

I liked that I could see and compare the effectiveness of the two separate sessions: have a improved?	I did not like that I could not compare myself to others. I also would have liked to see my total progress...is there ever an end to this?
The numbers, they clearly indicate what I did	No visual elements, such as graphs or charts. Also there was not comparative feedback, so I do not know how well I did compared to others.
The feedback was only showing me random numbers and letter bits. I had no clue whatsoever what they mean. I was extremely frustrated doing the work. If I would have read Japanese, I think I would have understood about the same amount of information (I don't speak Japanese).	It had no explanation or scale on which to measure success/ failure. There was no order or classifications by which I could get a sense of the meaning. Maybe it is not nice for humans to see numbers as a feedback or be measured by the amount of clicks they could have done in 5 minutes. I dont see the value in life being measured by numbers or dull work like that (post capitalistic world view).
That you were able to check whether you marked sth you did not plan to mark or which did not make sense	it looked boring, not interactive

Condition C3 (gamified feedback), rows are users, columns are like/dislike about feedback

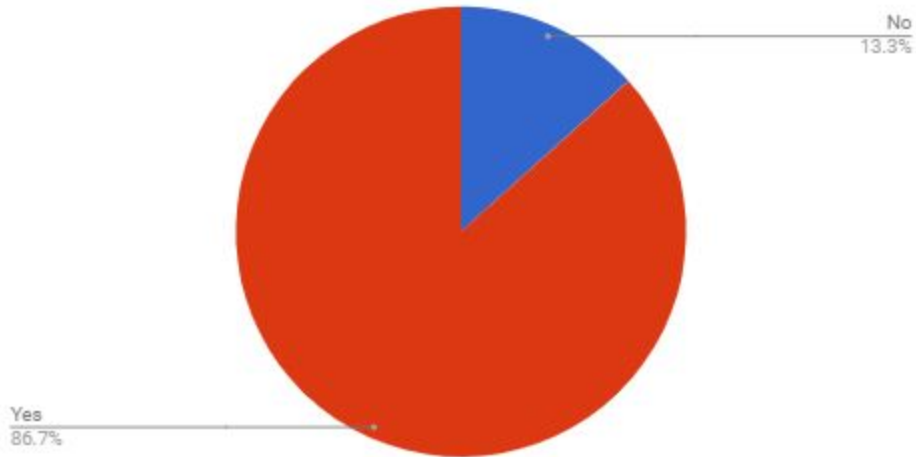
What did you like about the feedback?	What did you not like about the feedback?
The achievements	The bubbles with tags are not particularly relevant because the achievements are all about the total number of tags
simple overview of all the product names chosen	could not really see my progress
That I can compare the results I had before. Also I liked to keep an overview on how I am working	Although I did not have the feeling to work slower in the second round my achievements and results have been worse. It might be hard sometimes to compare your results especially if you do the feedback too frequently. I assume that working for a longer while shows more clearly a difference of productiveness.
The 'perks'/achievement concept	It was mostly based on the amount of completed annotations and newly added terms. It could've been more motivating if one knew how these terms would be organised in a later database step helping the user making connections between products, names and their functions (e.g. if a product is essentially the same but it has two different manufacturers).

It told me that I made a good marathon. That is stuff that makes me to want to be even better the next round-	Achieving a new level (marathon, sprint, etc.) could be animated (and with sound), so the emotional reward is more satisfying.
Interactivity of hovering, the level system, the 'badges' like a game achievement system	The colors of the bubbles, general feel was a bit like its unfinished, it needs a nice theme
Visual, easy to grasp.	Too minimal.
I liked that there was a level indication and that you got to go up levels by doing more annotating. It made the task more playful and if you like to compete with yourself a little, like I do, the feedback is pretty satisfactory.	I would have liked to get feedback about whether I did it right or how <i did compared to previous sessions.
the vividness and the visualizations used to show the information	It could come across as checking up on employees by the company, it could also make the workplace quite competitive
The feedback was nice to have, for such a short amount of time its nice to see how you did compared to the last time	All the different colors made it look a bit more childish than it could be.

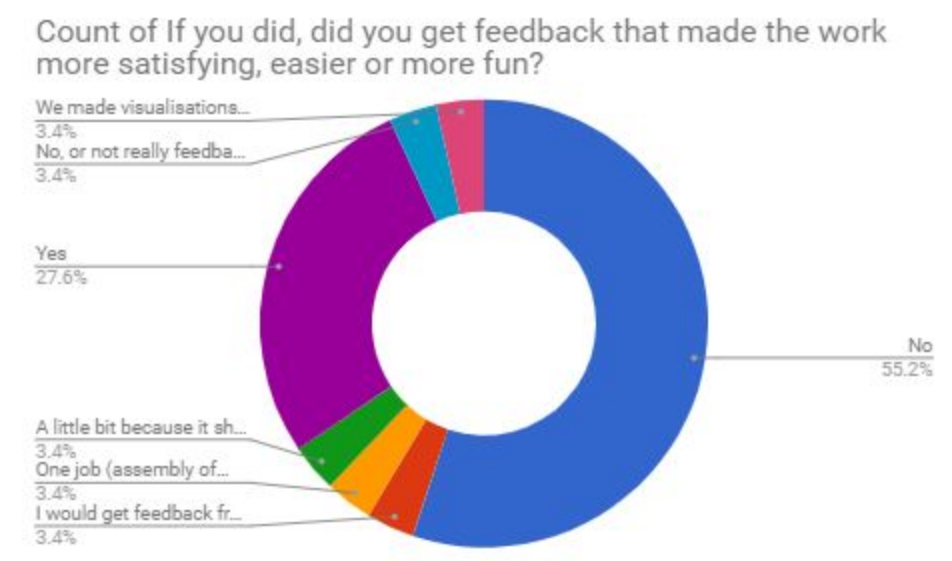
Section 5: Relatable experience, all testers

Do you have past experience doing things of the same nature as the annotation task: repetitive, linear, and not inherently interesting? Examples are: physical labor (factory or warehouse work, packaging, sorting, assembling), coding data or correcting or manipulating text files.

Count of Do you have past experience doing things of the same nature as the annotation task: repetitive, linear, and not inhere...



If you did, did you get feedback that made the work more satisfying, easier or more fun?



Answers expanded (each answer from one person):

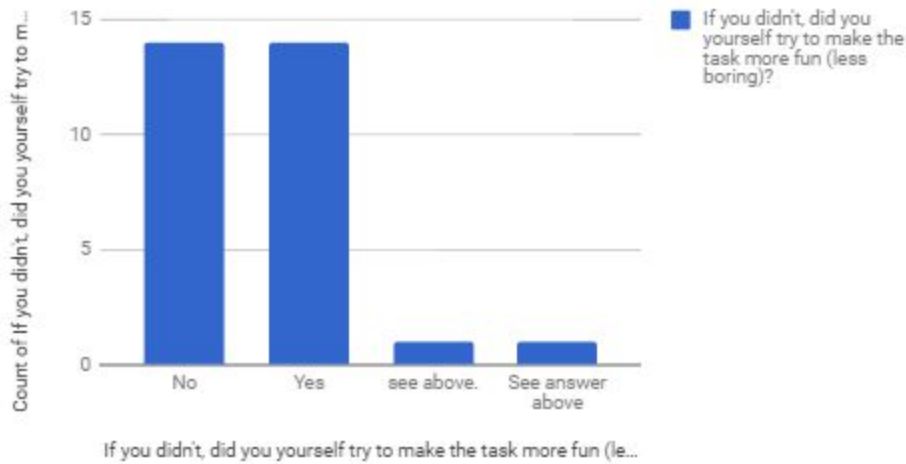
- I would get feedback from co-workers, not from software
- One job (assembly of machines): no, other job (delivering flyers): yes.
- A little bit because it showed I consistently found some names and serial numbers.
- No, or not really feedback, only that I should work faster or whatnot
- We made visualisations of the coded data, which showed some reward for the boring work.

What kind of feedback did you get that made it more satisfying, easier or more fun?
Positive feedback ("You did a good job!") from co-workers is always satisfying, maybe if the software would give me a positive feedback it would be funnier
approval of the result of the complete set of actions, the flyer company gave me Christmas cards which I would give to the houses and the people who live there gave me some money. This resulted in a Christmas bonus of op to 250 euros.
The high numbers of some of the product names that made me feel like I was doing well.
it was more game-alike

The achievements make you work harder and more motivated. Therefore performance appraisal is highly appreciated.
progress, how well have I done the task
none
Simply see progress in my work (how fast and efficient I got with time)
None
I got to upload it to a database which was quite fun to see
Worked in a oyster factory, opening oysters from a conveyor belt. Amount of oysters opened led to higher salary.
none- it was mostly just plain assembly line work
I got a daily goal. and feedback whether or not i had reached that daily goal.
.
customer interaction could tell you a fair bit about how well you are doing.
-
That I completed the job to the satisfaction of the employeeer (but only because that ment I could stay working there and would get salary)
Direct person-to-person congratulatory feedback, document feedback of total progress.

If you didn't [get any feedback], did you yourself try to make the task more fun (less boring)?

Count of If you didn't, did you yourself try to make the task more fun (less boring)?



If you tried to make it more fun, how did you do so?

BY trying to go as fast as I could, without losing time with miss-clicks

Distracting myself using music and audiobooks, since it was quite mindless work

-

Music

finding efficient working patterns

set time based goals, recite lyrics in my head..

trying to find the smoothest workflow possible

By constantly trying to improve the workflow. In that way you actually do not try to accomplish the task but instead improve it.

listening to podcasts

.
come up with challenges for myself, or talk with colleagues
-
listening to music while doing it
I was zoning out and going to my imaginary place of fun and trying to put on the autopilot to get the stuff done I had to do :D
Background music.
try to find a rhythm while doing it, doing it to the beat of music

All answers next to each other as a table

First Column question: Do you have past experience doing things of the same nature as the annotation task: repetitive, linear, and not inherently interesting? Examples are: physical labor (factory or warehouse work, packaging, sorting, assembling), coding data or correcting or manipulating text files.

Condition		If you did, did you get feedback that made the work more satisfying, easier or more fun?	What kind of feedback did you get that made it more satisfying, easier or more fun?	If you didn't, did you yourself try to make the task more fun (less boring) ?	If you tried to make it more fun, how did you do so?
2	No	No		No	
1	Yes	I would get feedback from co-workers, not from software	Positive feedback ('You did a good job!') from co-workers is always satisfying, maybe if the software would give me a positive feedback it would be funnier	Yes	BY trying to go as fast as I could, without losing time with miss-clicks
1	Yes	One job (assembly of machines): no, other job (delivering flyers): yes.	approval of the result of the complete set of actions, the flyer company gave me Christmas cards which I would give to the houses and the people who live there gave me some money. This resulted in a Christmas bonus of up to 250 euros.	Yes	Distracting myself using music and audiobooks, since it was quite mindless work
2	Yes	A little bit because it showed I consistently found some names and serial numbers.	The high numbers of some of the product names that made me feel like I was doing well.	No	
3	Yes	No		No	
1	No			No	

2	Yes	Yes	it was more game-alike	see above.	-
1	Yes	Yes	The achievements make you work harder and more motivated. Therefore performance appraisal is highly appreciated.	No	
3	Yes	No	progress, how well have I done the task	No	
3	Yes	No	none	Yes	Music
2	Yes	No	Simply see progress in my work (how fast and efficient I got with time)	See answer above	
2	Yes	No		No	
1	Yes	No	None	No	
1	Yes	Yes	I got to upload it to a database which was quite fun to see	Yes	finding efficient working patterns
1	Yes	Yes	Worked in a oyster factory, opening oysters from a conveyor belt. Amount of oysters opened led to higher salary.	Yes	set time based goals, recite lyrics in my head..
3	Yes	No	none- it was mostly just plain assembly line work	Yes	trying to find the smoothest workflow possible
3	Yes	No		Yes	By constantly trying to improve the workflow. In that way you actually do not try to accomplish the task but instead improve it.
2	Yes	No		Yes	listening to podcasts
2	Yes	Yes	I got a daily goal. and feedback whether or not i had reached that daily goal.	No	
1	Yes	Yes	.	Yes	.

2	Yes	Yes	customer interaction could tell you a fair bit about how well you are doing.	Yes	come up with challenges for myself, or talk with colleagues
3	Yes	No		No	
1	No	No	-	No	-
1	Yes	No		Yes	listening to music while doing it
2	Yes	No, or not really feedback, only that I should work faster or whatnot	That I completed the job to the satisfaction of the employee (but only because that meant I could stay working there and would get salary)	Yes	I was zoning out and going to my imaginary place of fun and trying to put on the autopilot to get the stuff done I had to do :D
2	Yes	No		No	
3	Yes	Yes	Direct person-to-person congratulatory feedback, document feedback of total progress.	Yes	Background music.
3	No	No		No	
3	Yes	We made visualisations of the coded data, which showed some reward for the boring work.		No	
3	Yes	No		Yes	try to find a rhythm while doing it, doing it to the beat of music

Section 6: Comments and demographics

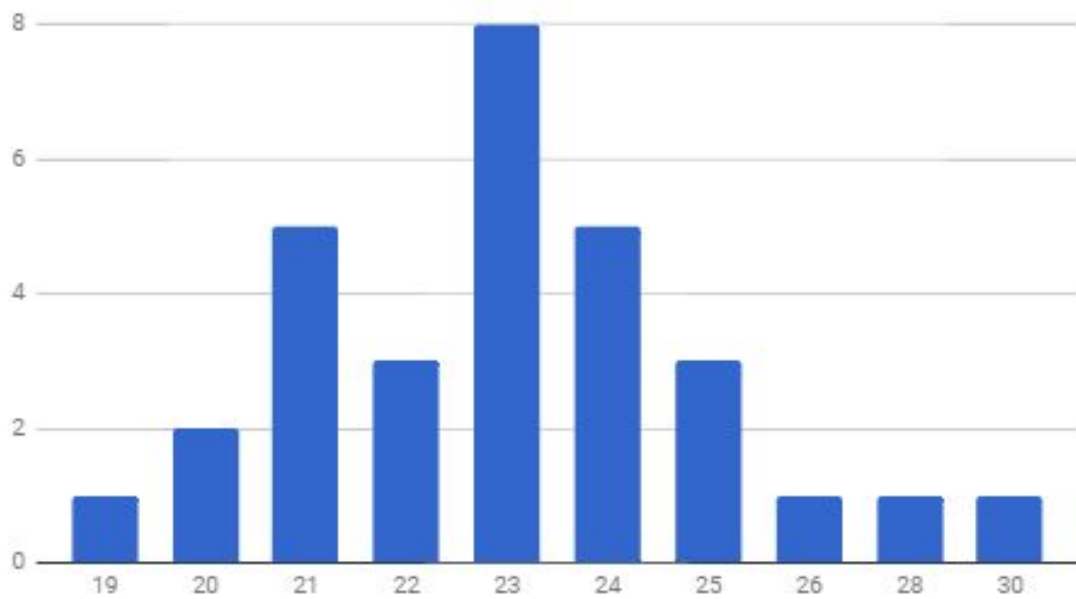
C	Assuming you have to keep working with this program: if you could improve the feedback system you got today, how would you? What would you add, take away or change?	Any final questions or comments?	Age:	Gender	Level of education	Employment status
1	I would make the font bigger to avoid wrong selections, I would make the program a bit more colorful	It is hard to perform such a job with very little information on the nature of the products beforehand	26	Male	Bachelor	Student with part time jobs
1	Add functionality to change entries you have made. Make the feedback more visually pleasing. you can implement learning so that the program can actually give suggestions.	.	21	Male	Bachelor	Student
1	I would add something that shows me the percentage of how many product names of a given set I identified correctly and which product names I missed.	.	24	Male	Bachelor	Student
1	Make it less often, and make it so that one must work towards a certain goal.		23	Female	Bachelor	Student
1	Delete duplicates	Software is easy to use and understand	22	Male	Bachelor	Employed student
1	column headers		23	Male	Bachelor	Student
1	add progress-graphs/ comparison scoreboards, add prizes for certain goals	selection process could be optimized by placing buttons for entering and deletion close to each other	30	Male	Bachelor	Student
1	add total counter, also per minute/hour - so you can set yourself a goal	does someone review my feedback?	21	Male	Bachelor	Student
1	maybe also some information about what i did wrong	-	21	Female	Bachelor	Student
1	I would pressure to improve the automatic algorithm, so that there is less work		19	Male	Bachelor	Student
2	maybe make it so that it encourages to keep working		22	Male	Master	Student

2	Make it more visual, maybe gamified.	The program interface was a little dull and the font unattractive.	22	Male	Bachelor	Student
2	i'd like to see something more appealing. More game-alike, so that i feel more motivated and challenged to do more, quicker whilst having fun beating myself and my colleagues.		24	Female	Bachelor	Student
2	Add already used product-names and give a feedback of right or wrong		24	Male	Bachelor	Student
2			20	Male	High school	Student
2	maybe add graphs that show progress, maybe show accuracy		23	Male	Bachelor	Student
2	I would make my colleague's feedback visible to create a sense of competition	good luck	24	Prefer not to say	Bachelor	Self-employed or business owner
2	more visual and comparative elements. Also, maybe a final score	annotation work is strangely relaxing	23	Male	Bachelor	Student
2	Most of the time it is motivational for humans to have stuff like this set up in a "game" sorta thing. Like if it would be a game or competition among co-workers, maybe it would be more fun. For example, pc games sometimes also include repetitive "clicking" tasks, but they are way more fun to do, also because of the visual appealing. Feedback should provide room to grow for individuals and be understandable and not only with numbers. The human should gain a value out of feedback and no pressure to "score" more clicks per minute (like pressure)	why the hell did I have to do this weird work?	25	Female	Master	Student
2	sound or colour or nicer button , some arrows etc	no	25	Female	Master	Student
3	Live updates during the work instead of only reviewing afterwards.		24	Male	Bachelor	Student
3	I would add whether the chosen product names have been chosen correctly		21	Female	High school	Student

3	You should not be able to get feedback to frequently since the feedback might be not really refer to the actual productivity.	no	25	Male	High school	Student
3	Give users more background info on what happens with the data and how it is organised.		23	Male	Bachelor	Student
3	Add an animation to achieving new levels and stages.		23	Male	High school	Student
3	maybe small sounds, ' buying' stuff with your points like in a mobile game, just follow all the steps of freemium games, making the money be your annotations		23	Male	Bachelor	Student
3	Would add feedback on accuracy or correctness of the work done.		28	Male	Bachelor	Employed
3			21	Female	Bachelor	Student
3	Maybe make the rewards for the employee clearer, so, if the employee would would finish 500 annotations, what would he get?		20	Male	Bachelor	Student
3	reduce the amount of colours on the page. The color of the bubbles was fine, but keep the rest a bit down		23	Male	Bachelor	Student

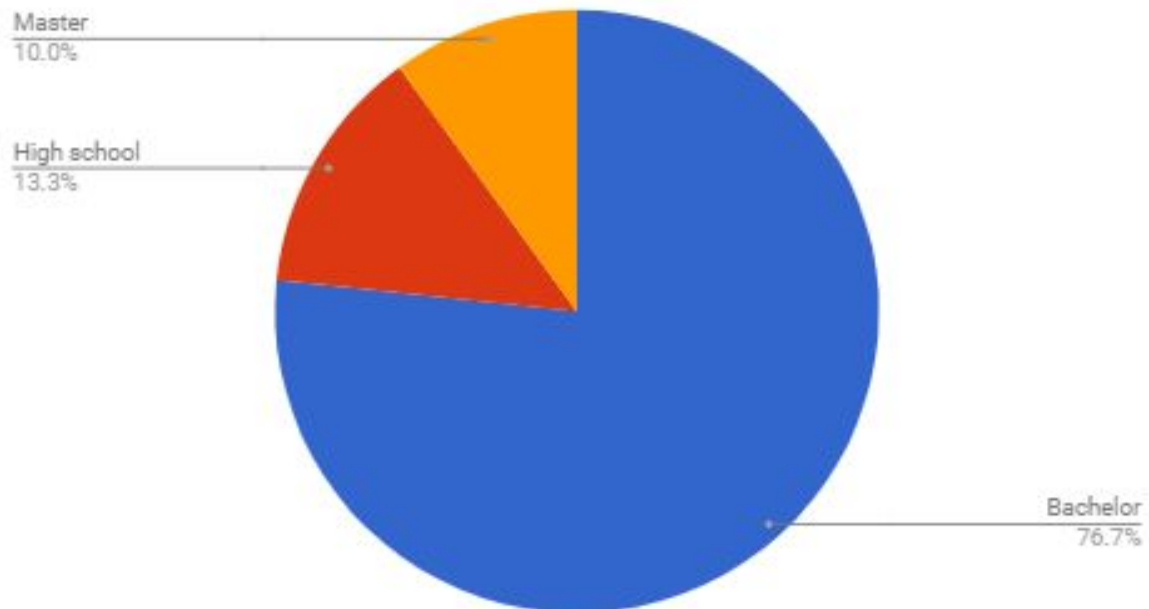
Tester ages:

Age distribution of testers



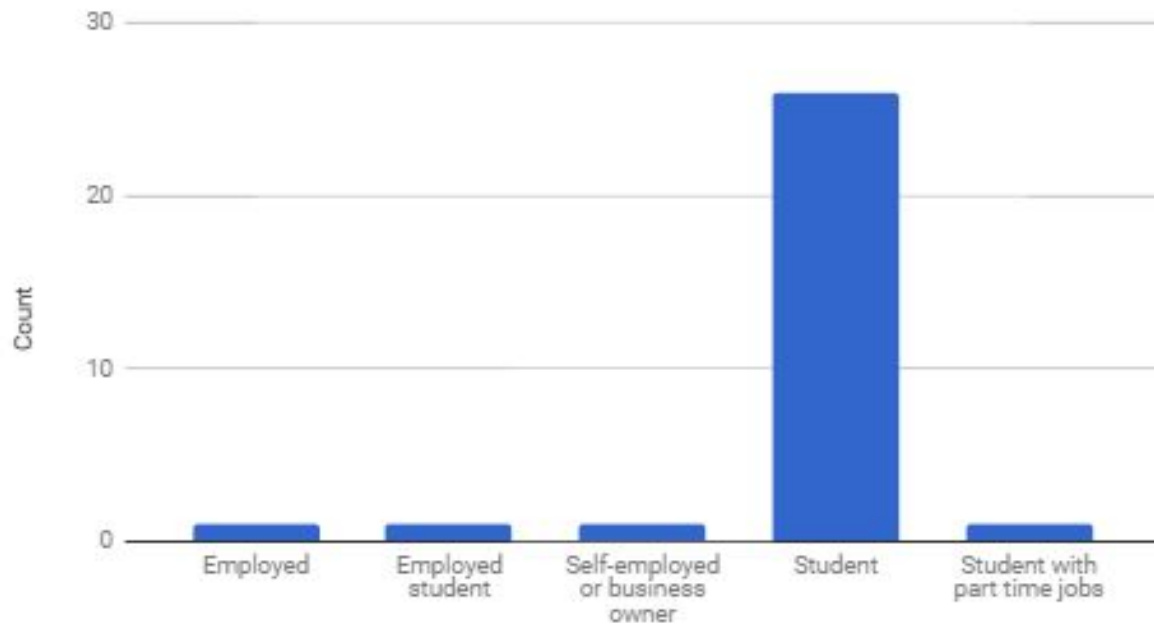
Tester education

Tester education levels



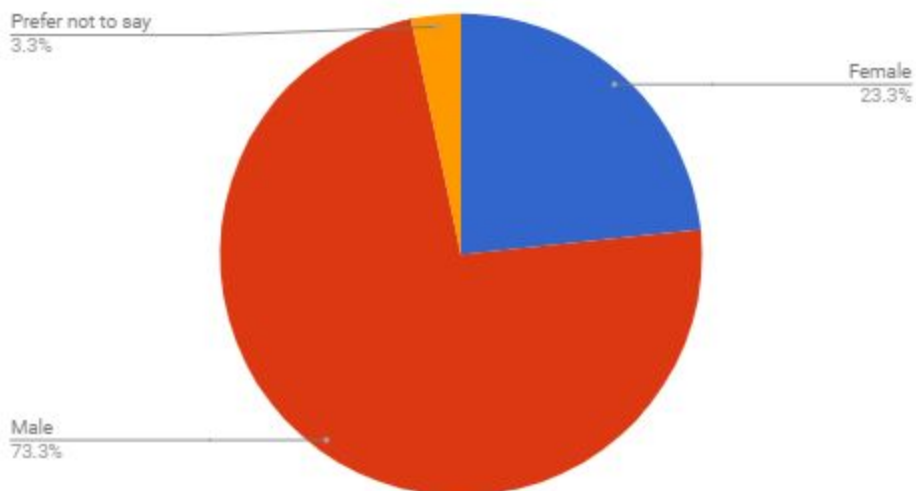
Tester employment status

Tester employment status



Tester gender distribution

Tester gender distribution



One-way ANOVA significance tests and follow-up Tukey tests

All results calculated with the online one-way ANOVA test tool at:

(<http://www.socscistatistics.com/tests/anova/default2.aspx>)

All Tukey test results calculated with the online Tukey HSD test tool at:

http://astatsa.com/OneWay_Anova_with_TukeyHSD

Condition 1 (Treatment 1/A) = file feedback

Condition 2 (Treatment 2/B) = non-gamified feedback

Condition 3 (Treatment 3/C) = gamified feedback

Test on amount of annotations done in each condition:

	Treatments					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	1187	1157	1259			3603
Mean	118.7	115.7	125.9			120.1
ΣX^2	150139	147835	175665			473639
Std.Dev.	32.0453	39.3984	43.6615			37.5631

Result Details				
Source	SS	df	MS	
Between-treatments	549.6	2	274.8	$F = 0.18379$
Within-treatments	40369.1	27	1495.1519	
Total	40918.7	29		

The F -ratio value is 0.18379. The p -value is .83314. The result is not significant at $p < .05$.

At $F(2,27) = 0.184$, $p = 0.833$, the result is not statistically significant.

There is no statistically significant difference in the amount of annotations made when users are given different formats of feedback.

Statements and attributes

1. If I could, I would've worked longer to complete a set amount of annotations (e.g. 50, 100, 200, 500 etc.).

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	42	34	47			123
Mean	4.2	3.4	4.7			4.1
ΣX^2	206	152	233			591
Std.Dev.	1.8135	2.0111	1.1595			1.7291

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	8.6	2	4.3	$F = 1.48656$
Within-treatments	78.1	27	2.8926	
Total	86.7	29		

The f -ratio value is 1.48656. The p -value is .244079. The result is *not* significant at $p < .05$.

At $F(2,27) = 1.487$, $p = 0.244$, the result is not statistically significant.

There is no statistically significance in whether users would've worked longer to complete a set amount of annotations.

2. Having done a set amount of annotations (e.g. 50, 100, 200, 500) would give me satisfaction.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	47	31	49			127
Mean	4.7	3.1	4.9			4.2333
ΣX^2	241	139	251			631
Std.Dev.	1.4944	2.1833	1.1005			1.7943

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	19.4667	2	9.7333	<i>F</i> = 3.55616
Within-treatments	73.9	27	2.737	
Total	93.3667	29		

The *f*-ratio value is 3.55616. The *p*-value is .042571. The result is significant at $p < .05$.

At $F(2,27) = 3.556$, $p=0.043$, the result is statistically significant.

Users report a lower agreement with the statement “Having done a set amount of annotations (e.g. 50, 100, 200, 500) would give me satisfaction.” when given feedback in the non-gamified format.

POST-HOC TUKEY TEST

As taken from the online Tukey test tool, the condition comparison results are as follows:

Tukey HSD results			
treatments pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
A vs B	3.0583	0.0961406	insignificant
A vs C	0.3823	0.8999947	insignificant
B vs C	3.4406	0.0551924	insignificant

The results indicate that amongst no pairs of the conditions are significant differences in the answers to the statement. Thus, no difference in agreement with the statement can be claimed.

3. I feel like I accomplished something.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	38	38	42			118
Mean	3.8	3.8	4.2			3.9333
ΣX^2	158	194	186			538
Std.Dev.	1.2293	2.3476	1.0328			1.596

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	1.0667	2	0.5333	$F = 0.1978$
Within-treatments	72.8	27	2.6963	
Total	73.8667	29		

The f -ratio value is 0.1978. The p -value is .821711. The result is *not* significant at $p < .05$.

At $F(2,27) = 0.198$, $p=0.822$, the result is not statistically significant.

There is no statistically significant difference in how users agreed to the statement "I feel like I accomplished something".

4. To see my progress grow, I would often pick up the annotation task often inbetween other tasks.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	33	37	44			114
Mean	3.3	3.7	4.4			3.8
ΣX^2	133	175	214			522
Std.Dev.	1.6364	2.0575	1.5055			1.7499

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	6.2	2	3.1	$F = 1.01332$
Within-treatments	82.6	27	3.0593	
Total	88.8	29		

The f -ratio value is 1.01332. The p -value is .376405. The result is *not* significant at $p < .05$.

At $F(2,27) = 1.013$, $p=0.376$, the result is not statistically significant.

There is no statistically significant difference in how users agreed to the statement “ To see my progress grow, I would often pick up the annotation task often inbetween other tasks”.

5. I felt challenged to do more annotation work.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	32	26	49			107
Mean	3.2	2.6	4.9			3.5667
ΣX^2	128	98	261			487
Std.Dev.	1.6865	1.8379	1.5239			1.9061

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	28.4667	2	14.2333	$F = 4.9974$
Within-treatments	76.9	27	2.8481	
Total	105.3667	29		

The f -ratio value is 4.9974. The p -value is .01424. The result is significant at $p < .05$.

At $F(2,27)=4.998$, $p=0.0142$, the result is statistically significant.

Users report a higher agreement with the statement “I felt challenged to do more annotation work” when given feedback in the gamified format.

POST-HOC TUKEY TEST

As taken from the online Tukey test tool, the condition comparison results are as follows:

treatments pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
A vs B	1.1243	0.6964440	insignificant
A vs C	3.1854	0.0803593	insignificant
B vs C	4.3097	0.0136981	* $p < 0.05$

In condition comparison, there is a significant difference in the agreement on the statement between conditions C2 (B) and C3 (C). Users given gamified feedback had significantly higher agreement with the statement, and users given the non-gamified feedback had significantly lower agreement with the statement.

6. In the second round, I looked forward to seeing the results of my work.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	30	33	60			123
Mean	3	3.3	6			4.1
ΣX^2	114	165	364			643
Std.Dev.	1.633	2.4967	0.6667			2.187

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	54.6	2	27.3	$F = 8.76457$
Within-treatments	84.1	27	3.1148	
Total	138.7	29		

The F -ratio value is 8.76457. The p -value is .001166. The result is significant at $p < .05$.

At $F(2,27)=8.765$, $p=0.001$, the result is statistically significant.

Users report a higher agreement with the statement “ In the second round, I looked forward to seeing the results of my work. ” when given feedback in the gamified format.

POST-HOC TUKEY TEST

As taken from the online Tukey test tool, the condition comparison results are as follows:

treatments pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
A vs B	0.5375	0.8999947	insignificant
A vs C	5.3753	0.0020899	** p<0.01
B vs C	4.8378	0.0054855	** p<0.01

In condition comparison, there is a significant difference in the agreement on the statement between conditions C3 (C) and the other two conditions. This means that users given the gamified feedback had significantly higher agreement with the statement than users given either other format of feedback.

7. I would've updated my feedback more often, if I could've.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	28	20	38			86
Mean	2.8	2	3.8			2.8667
ΣX^2	102	70	174			346
Std.Dev.	1.6193	1.8257	1.8135			1.852

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	16.2667	2	8.1333	$F = 2.63942$
Within-treatments	83.2	27	3.0815	
Total	99.4667	29		

The f -ratio value is 2.63942. The p -value is .089747. The result is *not* significant at $p < .05$.

At $F(2,27) = 2.639$, $p=0.089$, the result is not statistically significant.

There is no statistically significant difference in how users agreed to the statement “ I would've updated my feedback more often, if I could've”.

8. The feedback I got was childish.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	27	13	22			62
Mean	2.7	1.3	2.2			2.0667
ΣX^2	93	21	52			166
Std.Dev.	1.4944	0.6749	0.6325			1.1427

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	10.0667	2	5.0333	$F = 4.88849$
Within-treatments	27.8	27	1.0296	
Total	37.8667	29		

The f -ratio value is 4.88849. The p -value is .015422. The result is significant at $p < .05$.

At $F(2,27)=4.889$, $p=0.015$, the result is statistically significant.

Users report a lower agreement with the statement “The feedback I got was childish” when given feedback in the non-gamified format.

POST-HOC TUKEY TEST

As taken from the online Tukey test tool, the condition comparison results are as follows:

treatments pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
A vs B	4.3630	0.0125139	* $p < 0.05$
A vs C	1.5582	0.5213273	insignificant
B vs C	2.8048	0.1356006	insignificant

In condition comparison, there is a significant difference in the agreement on the statement between conditions C1 and C2. This means that users given the file feedback had significantly higher agreement with the statement than users given the non-gamified feedback.

9. The feedback I got was inappropriate for a workplace, such as an office.

	Treatments					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	32	21	28			81
Mean	3.2	2.1	2.8			2.7
ΣX^2	126	79	100			305
Std.Dev.	1.6193	1.9692	1.5492			1.7251

Result Details				
Source	SS	df	MS	
Between-treatments	6.2	2	3.1	$F = 1.04494$
Within-treatments	80.1	27	2.9667	
Total	86.3	29		

The f -ratio value is 1.04494. The p -value is .365505. The result is *not* significant at $p < .05$.

At $F(2,27) = 1.045$, $p=0.365$, the result is not statistically significant.

There is no statistically significant difference in how users agreed to the statement "The feedback I got was inappropriate for a workplace, such as an office".

10. I feel a sense of ownership over the work I've done.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	39	32	42			113
Mean	3.9	3.2	4.2			3.7667
ΣX^2	173	140	194			507
Std.Dev.	1.5239	2.044	1.3984			1.675

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	5.2667	2	2.6333	$F = 0.9343$
Within-treatments	76.1	27	2.8185	
Total	81.3667	29		

The f -ratio value is 0.9343. The p -value is .405196. The result is *not* significant at $p < .05$.

At $F(2,27)=0.934$, $p=0.405$, the result is not statistically significant.

There is no statistically significant difference in how users agreed to the statement “ I feel a sense of ownership over the work I've done”.

11. I feel like my work was not significant.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	33	35	39			107
Mean	3.3	3.5	3.9			3.5667
ΣX^2	131	177	165			473
Std.Dev.	1.567	2.4608	1.1972			1.775

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	1.8667	2	0.9333	$F = 0.28156$
Within-treatments	89.5	27	3.3148	
Total	91.3667	29		

The f -ratio value is 0.28156. The p -value is .756791. The result is *not* significant at $p < .05$.

At $F(2,27)=0.282$, $p=0.757$, the result is not statistically significant.

There is no statistically significant difference in how users agreed to the statement “ I feel like my work was not significant”.

12. Doing the annotation work was satisfying.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	36	33	38			107
Mean	3.6	3.3	3.8			3.5667
ΣX^2	150	129	156			435
Std.Dev.	1.5055	1.4944	1.1353			1.3566

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	1.2667	2	0.6333	$F = 0.32822$
Within-treatments	52.1	27	1.9296	
Total	53.3667	29		

The F -ratio value is 0.32822. The p -value is .723041. The result is *not* significant at $p < .05$.

At $F(2,27)=0.328$, $p=0.723$, the result is not statistically significant.

There is no statistically significant difference in how users agreed to the statement “Doing the annotation work was satisfying”.

13. I would share my feedback with others who were tasked to do the same work.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	44	38	49			131
Mean	4.4	3.8	4.9			4.3667
ΣX^2	234	200	275			709
Std.Dev.	2.1187	2.4855	1.9692			2.1732

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	6.0667	2	3.0333	$F = 0.62567$
Within-treatments	130.9	27	4.8481	
Total	136.9667	29		

The F -ratio value is 0.62567. The p -value is .542481. The result is *not* significant at $p < .05$.

At $F(2,27)=0.626$, $p=0.542$, the result is not statistically significant.

There is no statistically significant difference in how users agreed to the statement "I would share my feedback with others who were tasked to do the same work".

14. I would be interested in seeing the feedback of others such as colleagues, who also did this annotation work.

	Treatments					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	43	38	49			130
Mean	4.3	3.8	4.9			4.3333
ΣX^2	227	198	277			702
Std.Dev.	2.1628	2.4404	2.0248			2.1867

Result Details				
Source	SS	df	MS	
Between-treatments	6.0667	2	3.0333	$F = 0.61765$
Within-treatments	132.6	27	4.9111	
Total	138.6667	29		

The F -ratio value is 0.61765. The p -value is .546657. The result is *not* significant at $p < .05$.

At $F(2,27)=0.618$, $p=0.547$, the result is not statistically significant.

There is no statistically significant difference in how users agreed to the statement “I would be interested in seeing the feedback of others such as colleagues, who also did this annotation work”.

15. I was interested in the content of the database.

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	36	28	40			104
Mean	3.6	2.8	4			3.4667
ΣX^2	160	118	200			478
Std.Dev.	1.8379	2.0976	2.1082			2.0126

Result Details				
Source	SS	df	MS	
Between-treatments	7.4667	2	3.7333	$F = 0.91636$
Within-treatments	110	27	4.0741	
Total	117.4667	29		

The F -ratio value is 0.91636. The p -value is .412053. The result is *not* significant at $p < .05$.

At $F(2,27)=0.916$, $p=0.412$, the result is not statistically significant.

There is no statistically significant difference in how users agreed to the statement “I was interested in the content of the database”.

Feedback evaluation, attributes

16. Ease of understanding (1: very difficult to understand, 7: very easy to understand)

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	53	52	57			162
Mean	5.3	5.2	5.7			5.4
ΣX^2	305	318	335			958
Std.Dev.	1.6364	2.2998	1.0593			1.6938

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	1.4	2	0.7	$F = 0.23105$
Within-treatments	81.8	27	3.0296	
Total	83.2	29		

The f -ratio value is 0.23105. The p -value is .795252. The result is *not* significant at $p < .05$.

At $F(2,27)=0.231$, $p=0.795$, the result is not statistically significant.

There is no statistically significant difference in how users rated their feedback in the category 'ease of use'.

17. Frustration vs. satisfaction (1: very frustrating to see, 7: very satisfying to see)

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	42	41	54			137
Mean	4.2	4.1	5.4			4.5667
ΣX^2	188	185	296			669
Std.Dev.	1.1353	1.3703	0.6992			1.2229

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	10.4667	2	5.2333	$F = 4.29483$
Within-treatments	32.9	27	1.2185	
Total	43.3667	29		

The f -ratio value is 4.29483. The p -value is .024018. The result is significant at $p < .05$.

At $F(2,27)=4.295$, $p=0.024$, the result is statistically significant.

Users given the gamified feedback rated their feedback in the category 'frustration vs. satisfaction' closer to the attribute 'satisfying'.

POST-HOC TUKEY TEST

As taken from the online Tukey test tool, the condition comparison results are as follows:

treatments pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
A vs B	0.2865	0.8999947	insignificant
A vs C	3.4377	0.0554333	insignificant
B vs C	3.7242	0.0356633	* $p < 0.05$

In condition comparison, there is a significant difference in evaluation of the attribute between conditions C2 and C3. This means that users given the non-gamified feedback rated the attribute significantly lower than users given the gamified feedback.

18. Boring vs. interesting (1: very boring/dull, 7: very interesting)

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	29	28	43			100
Mean	2.9	2.8	4.3			3.3333
ΣX^2	101	102	195			398
Std.Dev.	1.3703	1.6193	1.0593			1.4933

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	14.0667	2	7.0333	$F = 3.75296$
Within-treatments	50.6	27	1.8741	
Total	64.6667	29		

The f -ratio value is 3.75296. The p -value is .036462. The result is significant at $p < .05$.

At $F(2,27)=3.753$, $p=0.036$, the result is statistically significant.

Users given the gamified feedback rated their feedback in the category 'boring vs. Interesting' closer to the attribute 'interesting'.

POST-HOC TUKEY TEST

As taken from the online Tukey test tool, the condition comparison results are as follows:

treatments pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
A vs B	0.2310	0.8999947	insignificant
A vs C	3.2340	0.0749265	insignificant
B vs C	3.4650	0.0531982	insignificant

The results indicate that there is no statistical significance between any of the pairs of conditions. Thus, no difference in evaluation of the attributes can be claimed.

19. Relevance to the task (1: completely irrelevant to the annotation task, 7: highly relevant to the annotation task)

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	49	43	45			137
Mean	4.9	4.3	4.5			4.5667
ΣX^2	259	213	217			689
Std.Dev.	1.4491	1.767	1.2693			1.4782

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	1.8667	2	0.9333	$F = 0.40976$
Within-treatments	61.5	27	2.2778	
Total	63.3667	29		

The f -ratio value is 0.40976. The p -value is .667871. The result is *not* significant at $p < .05$.

At $F(2,27)=0.409$, $p=0.668$, the result is not statistically significant.

There is no statistically significant difference in how users rated their feedback in the category 'relevance to the task'.

20. Motivational value (1: very demotivating, 7: very motivating)

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	43	38	50			131
Mean	4.3	3.8	5			4.3667
ΣX^2	195	162	266			623
Std.Dev.	1.0593	1.3984	1.3333			1.3257

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	7.2667	2	3.6333	$F = 2.24485$
Within-treatments	43.7	27	1.6185	
Total	50.9667	29		

The f -ratio value is 2.24485. The p -value is .125353. The result is *not* significant at $p < .05$.

At $F(2,27)=2.245$, $p=0.125$, the result is not statistically significant.

There is no statistically significant difference in how users rated their feedback in the category 'motivational value'.

21. Usefulness (1: completely useless, 7: very useful)

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	44	38	47			129
Mean	4.4	3.8	4.7			4.3
ΣX^2	222	166	225			613
Std.Dev.	1.7764	1.5492	0.6749			1.4179

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	4.2	2	2.1	$F = 1.04806$
Within-treatments	54.1	27	2.0037	
Total	58.3	29		

The f -ratio value is 1.04806. The p -value is .36445. The result is *not* significant at $p < .05$.

At $F(2,27)=1.049$, $p=0.364$, the result is not statistically significant.

There is no statistically significant difference in how users rated their feedback in the category 'usefulness'.

22. Visual value (1: ugly, aesthetically displeasing, 7: good looking, aesthetically pleasing)

	<i>Treatments</i>					
	1	2	3	4	5	Total
N	10	10	10			30
ΣX	38	28	48			114
Mean	3.8	2.8	4.8			3.8
ΣX^2	170	112	246			528
Std.Dev.	1.6865	1.9322	1.3166			1.808

Result Details				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	
Between-treatments	20	2	10	$F = 3.60963$
Within-treatments	74.8	27	2.7704	
Total	94.8	29		

The f -ratio value is 3.60963. The p -value is .040809. The result is significant at $p < .05$.

At $F(2,27)=3.609$, $p=0.041$, the result is statistically significant.

Users given the gamified feedback rated their feedback in the category 'visual value' closer to the attribute 'aesthetically pleasing'. Users given the non-gamified feedback rated closer to the attribute 'ugly, aesthetically displeasing'.

POST-HOC TUKEY TEST

As taken from the online Tukey test tool, the condition comparison results are as follows:

treatments pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
A vs B	1.8999	0.3853631	insignificant
A vs C	1.8999	0.3853631	insignificant
B vs C	3.7998	0.0316390	* $p < 0.05$

In condition comparison, there is a significant difference in evaluation of the attribute between conditions C2 and C3. This means that users given the non-gamified feedback rated the attribute significantly lower than users given the gamified feedback.