INSTITUT FÜR INFORMATIK
der Ludwig-Maximilians-Universität München

Enhancing Motivation through Gamification and Feedback

An Application to Text Entity Evaluation Tasks

Felix Wanie

Masterarbeit

Aufgabensteller	Prof. Dr. François Bry
Betreuer	Prof. Dr. François Bry, Martin Bogner
Unterstützung	Dr. Christian Bauer, Glanos GmbH
Abgabe am	2. Oktober 2018

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst habe und keine anderen als die angegebenen Hilfsmittel verwendet habe.

München, den 2. Oktober 2018 Felix Wanie

Abstract

Many processes in modern business rely on huge amounts of data that have to be entered, processed and kept up-to-date. However, the tasks that relate to data maintenance are often perceived as uninteresting and monotonous. Therefore, gamification and feedback can be used to increase the motivation of the employees assigned to these tasks.

This master's thesis applies this strategy to a text analysis software called Gold Standard Annotator which was developed by the Munich-based company Glanos. More specifically, it introduces motivational enhancements to a process which requires users to evaluate entities found in large text corpora. The goal is to enhance motivation by introducing gamification and feedback and subsequently increase the number of processed entities and the quality of the evaluations within the process.

In order to achieve this goal, the Gold Standard Annotator is analyzed for existing motivational affordances and potentials for additional motivation enhancement. The findings are backed up by interviews which were conducted with two of Glanos' employees. These interviews document the current state of the system and inquire into the users' view on their work and their current motivation.

With this information available, it is possible to chose a selection of game design elements which address the psychological needs for competence and social relatedness that are currently neglected. Providing progress bars, badges and a team challenge is an attempt to satisfy these needs and thus create a more motivating environment.

To test whether these elements work as expected, an online study is conducted which splits participants into two groups. Each participant is introduced to the topic and the tasks first. They then evaluate entities according to the evaluation process but while one group is exposed to gamification and feedback, the other group does not receive any such stimuli. Finally, all participants are forwarded to a questionnaire which records their thoughts and feelings towards the process they just participated in. Thus, the study not only gathers data on the user behavior during the evaluation phase but also collects information on their personal opinions.

This data is then analyzed for differences in motivation, number of evaluations and quality of the output. As expected, users who are subject to gamification state higher scores on motivation. Moreover, strong indications for a greater amount of processed entities can be found. However, no changes in evaluation quality can be detected. The thesis concludes by discussing possible reasons for these results and proposing further research on the long-term effects of the chosen solution.

Zusammenfassung

Viele Prozesse in der modernen Geschäftswelt bauen auf großen Datenmengen auf, die eingepflegt, verarbeitet und aktuell gehalten werden müssen. Die Aufgaben die mit der Datenpflege einher gehen werden jedoch oft von den damit beauftragten Angestellten als uninteressant und monoton empfunden. Deshalb setzt man Gamification und Feedback ein um die Motivation der Mitarbeiter zu steigern, die mit diesen Aufgaben betraut sind.

Diese Masterarbeit wendet diese Strategie auf eine Textanalysesoftware an, die Gold Standard Annotator genannt wird und von der Münchner Firma Glanos entwickelt wurde. Genauer gesagt werden motivationssteigernde Maßnahmen in einen Prozess eingebaut, bei dem die Nutzer Entitäten evaluieren müssen, die in großen Textkorpora gefunden wurden. Das Ziel ist die Motivation durch die Einführung von Gamification und Feedback zu stärken und damit sukzessiv auch die Anzahl der verarbeiteten Entitäten und die Qualität der Evaluierungen innerhalb des Prozesses zu steigern.

Um dieses Ziel zu erreichen wird der Gold Standard Annotator auf bereits existierende motivationale Anreize und Potentiale für zusätzliche Motivationsunterstützung untersucht. Die Erkenntnisse werden durch Interviews mit zwei Angestellten der Glanos GmbH bestätigt. Sie dokumentieren den aktuellen Stand des Systems und befragen die Nutzer zu Ihren Ansichten bezüglich ihrer Arbeit und ihrer aktuellen Motivation.

Mit diesen Informationen ist es möglich eine Auswahl an Spiel-Design-Elementen zusammenzustellen, die die psychologischen Bedürfnisse nach Kompetenzerleben und sozialer Eingebundenheit ansprechen. Diese werden aktuell vernachlässigt. Das Bereitstellen von Fortschrittsbalken, Auszeichnungen und einer Teamaufgabe ist ein Versuch diese Bedürfnisse zu stillen und so eine motivierendere Umgebung zu schaffen.

Um zu testen ob diese Elemente so funktionieren wie erwartet, wird eine Onlinestudie durchgeführt die die Teilnehmer in zwei Gruppen aufteilt. Jeder Teilnehmer wird zuerst in das Thema und die Aufgaben eingeführt. Dann evaluieren die Teilnehmer Entitäten wie im Evaluationsprozess vorgesehen, aber während eine Gruppe Gamification und Feedback angezeigt bekommt erhält die andere Gruppe keine solchen Stimuli. Zum Schluss werden alle Teilnehmer auf einen Fragebogen weitergeleitet der ihre Gedanken und Gefühle bezüglich des Prozesses, an dem sie gerade teilgenommen haben, festhält. Damit erfasst die Studie nicht nur Daten zum Nutzerverhalten während der Evaluationsphase sondern sammelt auch Informationen über die persönlichen Ansichten der Nutzer.

Diese Daten werden dann auf Unterschiede in Motivation, Anzahl von Evaluierungen und Qualität des Arbeitsergebnisses hin untersucht. Wie erwartet geben Nutzer auf die Gamification angewandt wurde bessere Werte im Bereich Motivation an. Zudem enthalten die Daten starke Hinweise auf eine größere Anzahl an verarbeiteten Entitäten. Allerdings können keine Unterschiede in der Qualität der Evaluierungen gefunden werden. Die Arbeit schließt mit der Diskussion der möglichen Gründe für diese Ergebnisse und regt zu weiterer Forschung bezüglich der Langzeiteffekte der gewählten Lösungen an.

Acknowledgments

While a Master's thesis might look like a one-man job from the outside it would not be possible without the helping hands of many.

I want to thank Prof. Bry for his support and guidance throughout this project. His many ideas and insightful suggestions were beyond helpful and the enthusiasm with which he supported this thesis motivated me to do my very best.

A special thanks goes to Martin Bogner for his considerate monitoring of my progress and his help on so many levels. His calm approach to distinguishing the possible from the impossible and his precise feedback helped me to stay on track and improve my work continuously. Thanks again!

Moreover, I want to express my thanks to the head of Glanos GmbH, Dr. Christian Bauer, who encouraged me to choose this topic and provided the necessary data to conduct the study. I also thank Matthias Groch and Sofian Latreche for giving me their insight into the state of the DataSphere and answering all my questions. I extend these thanks to all my (former) colleagues at Glanos who came up with tips, explanations and ideas for every problem I encountered.

I also want to thank my parents Elke and Klaus Wanie for their continuous support and their believe in me. A warm thanks goes to my wonderful girlfriend Julia Wunderlich, who did not stop encouraging me even though she had to put up with an overworked version of myself for far too long. Finally, a big thanks to all my friends and family for their support along the way.

Contents

CHAPTER 1

Introduction

Due to the introduction of modern information technology into almost every section of today's business world, an increasing amount of data has to be entered, managed and kept up to date on an everyday basis. While these tasks are vital for most businesses to function successfully, the employees assigned to these tasks often perceive them as uninteresting, tedious and repetitive. This lack of motivation naturally leads to inputs of lower quality which then influences the company's daily business and repeatedly necessitates corrections.

This issue also concerns the company Glanos. The Munich-based firm has developed advanced text analytics software called DataSphere which finds entities of various types in almost all kinds of text. With their programs, it is possible to automatically extract client names, subjects, and addresses from thousands of emails for example. However, it is unavoidable to go through the text and its entities by hand from time to time in order to ensure the correctness of the results. The contexts in which the entities are located might change over time and additional cases will be found during productive use. Thus, checks by humans are required for quality insurance and continuous development. On the other hand, manually scrolling through a text and checking hundreds or even thousands of found entities can be tiring and demotivating. The chances for making mistakes increase, which will in turn influence the continuous development of the underlying programs and might even lead to wrongly extracted entities in the future.

One solution to improve this situation is to employ adequate feedback mechanisms and means of gamification. The basic idea is to enrich the editors with game elements to enhance the user's motivation and thus receive better quality results. This strategy has been successfully used in Human Computation-based systems like ARTigo[1] or Eyewire[2].

However, the gamification concepts for business software inevitably differ from the concepts used in open source projects because they were implemented with another context in mind. Most Human Computation platforms are publicly available and try to process huge quantities of data with crowed-based approaches (i.e. in the form of games). Glanos' software, on the other hand, is available for paying customers only and requires specifically instructed workers to process the data. Common issues of open source projects like

[1] ARTigo is a platform to tag artwork. It therefore uses a variety of games with a purpose. For more information, visit `https://www.artigo.org/`.

[2] Eyewire is a citizen science project which focuses on mapping the brain. See `https://blog.eyewire.org/about/` for more information.

recruiting a sufficient amount of volunteers are not applicable to a company with its payed workforce. Additionally, working environments have their own social rules which have to be considered. Enhancing competition might be undesirable, for example, if the team members are intended to solve the task cooperatively. By contrast, introducing competition to an Human Computation platform can be beneficial for the volunteers' motivation (see for example the Old Weather project [**?**]). Due to these differences, transferring these platforms' gamification mechanisms to a business software like the DataSphere might not result in the desired effects.

This thesis puts forth a concept for applying gamification and feedback to the DataSphere's entity evaluation process. This concept considers the task as well as the circumstances that surround it. The proposed solutions are implemented and a user study is conducted to measure their impact on user motivation, quantity of processed entities, and evaluation quality.

After specifying the research question and its hypotheses, the paper summarizes important related work and provides the necessary theoretical background. Apart from defining the key terms, it intends to give an overview on the basic principles of motivation, the mechanics of feedback and the concepts of gamification.

A detailed case analysis is next. It starts by describing Glanos' business model and explaining how the text analysis works. It also highlights the potential for motivational affordances in the system. These claims are substantiated by interviews on the user experience of the current version, which were conducted with two Glanos employees.

The proposal for adequate feedback and gamification methods is outlined in chapter five. Therefore, the findings from the analysis are mapped to psychological needs which are disregarded by the current version of the system. By proposing adequate game design elements which address these needs, the overall motivational affordance of the system is enhanced.

The sixth chapter elucidates the concept for the user study. It describes the study's setup and determines a set of requirements for the application the study runs in. It also outlines the construction of the questionnaire which is intended to collect the data on the participants' inner feelings and attitudes. Additionally, the statistical methods for analyzing the resulting data are presented.

The actual implementation including the program's architecture, the data processing and its user flow are summarized hereafter. Apart from describing the basic data and its preprocessing, the chapter intends to give an insight on the concepts that guided the implementation. Additionally, it outlines major decisions on issues like security, data privacy and user experience design.

Finally, the results of the study are presented and interpreted in chapter eight. It also evaluates the approach critically. The thesis finishes with summarizing the contents and addressing potential future work.

CHAPTER 2

Research question and terminology

2.1 Research question and hypotheses

As already mentioned, this thesis' central topic is the conflict that arises when employees who work in data maintenance perceive their tasks as tedious and uninteresting even though these tasks are essential for a company's business processes. More specifically, it examines this conflict within the entity evaluation process in Glanos' DataSphere.

During this process, employees scroll through large quantities of text which contain highlighted words wherever an entity was found. They then click on these highlightings and accept or reject the entity by clicking a button. If they come across an entity that has not been found by the software, they add the highlighting manually. It is easy to see that this monotonous task is necessary for quality insurance but boring or demotivating for the employees assigned to it.

This thesis attempts to improve the situation by introducing adequate feedback and gamification strategies to the process. The goal is to both enhance user motivation and process more data correctly. More specifically, this means to evaluate greater quantities of entities in a better quality. The overall research question reflects these goals:

Does introducing feedback and gamification to the entity evaluation process in the DataSphere have a positive effect on its participants' motivation and its results?

A valid approach to answering this question is to focus on only one of its aspects at the time. Each aspect is formulated into a hypothesis which can then be proven or disproven through collecting and analyzing data. Accepting or rejecting the hypotheses will then support finding an answer to the overall research question.

The first aspect in the question is the topic of enhancing motivation. The decision to apply gamification and feedback mechanisms is usually based on the assumption that there should be a measurable difference in motivation caused by these methods. In other words, it must be possible to evaluate whether the implemented methods work as intended. This is the topic of the first hypothesis:

> ***Hypothesis 1: Users who are exposed to motivation-enhancing methods while evaluating entities report to be more motivated than their peers who are not subject to these methods.***

Even though this statement sounds simple and logical, it reveals to be far more complicated upon closer examination. There is the problem of how to objectively measure motivation of a person, for instance. In addition to that, participants can only give subjective information about this inner state of mind which makes it difficult to compare. It is therefore necessary to find a strategy which allows to record this subjective state in a comparable form.

In contrast, comparing quantities of evaluated annotations is rather simple. Nevertheless, a goal of this project is applying motivation-enhancing strategies in order to get more evaluated entities:

> ***Hypothesis 2: Users who are subject to gamification and feedback evaluate larger quantities of entities than users who lack these stimuli.***

The underlying idea is that gamification and feedback uphold the user's motivation and concentration which lets them work through greater number of entities. They also help them focus on certain types of entities which they might have ignored otherwise. The expected result is therefore that they not only evaluate a larger amount of entities but also a greater variety of entity types.

Apart from processing more entities than users who are not exposed to motivation-enhancing mechanisms, the participants are expected to produce evaluations of higher quality:

> ***Hypothesis 3: The overall quality of the evaluations is better for users who receive motivation-enhancing stimuli.***

Here, it is necessary to provide a definition for the term quality in this context. More specifically, it has to be clarified what a high-quality annotation is and how these definitions are passed on to the users. Only if the users share a common understanding of the standards they have to adhere to, the overall result's quality will be measurable. Otherwise, it will be difficult to assess whether the result was caused by diverse definitions or a difference in motivation and focus.

It is notable that these hypotheses are neither in conflict with each other nor are they redundant. Even if users feel more motivated, as postulated in the first hypothesis, that does not necessarily entail that they process more entities and that these entities are of higher quality. A larger amount of processed entities can still be of the same (or lower) quality. And even if the participants produce more or better evaluations, they might still not feel motivated by the gamification and feedback methods. Thus, each of the hypotheses is an independent statement.

In order to prove the hypotheses, a study has been conducted during which data has been collected from both the evaluation process and by asking the users themselves. This requires building an application which behaves like the current evaluation process at Glanos but can be enhanced with gamification and feedback methods. To ensure comparability, the participants have to be split into two groups. Depending on their group, they are either using the gamified version of the application or its non-gamified counterpart. After having evaluated entities in their respective environment, the participants fill out a questionnaire which records their experiences during the evaluation process. Thus, the study gathers both their evaluation data and their personal state of mind. Analyzing the recorded data will then lead to accepting or rejecting the hypotheses.

2.2 A note on terminology

Before going into detail on how to prove each of the hypotheses, it is worthwhile to clarify the meaning of some terms that occur regularly throughout this thesis. These terms are

study and *application* or *program*. Even though they are clearly connected, they are not interchangeable.

Within the given context, *study* refers to the entire process of gathering volunteers, letting them evaluate entities on a corpus and then recording their experiences with this evaluation process. It also includes the analysis of the results. The volunteers who partake in this process are referred to as *participants*.

Application or *program*, on the other hand, denotes the website that the participants use to enter the data. It is a product of a development process and designed to record and save specific data. The people interacting with it are called *users*.

Although this distinction seems a little pedantic at first, it is necessary when talking about how both terms are related. While the *study* is an abstract concept which aims at proving or disproving the hypotheses by applying a sound scientific process, the *application* is one concrete realization of this concept. *Participants* of the *study* are hence necessarily *users* of the *program*. What distinguishes both terms, though, is their focus. The *application* has to face practical issues like software architecture and data security, for instance, whereas the *study* is concerned with theoretical problems like the validity of the results. This example clearly depicts why the two terms cannot be used synonymously.

In the following, this distinction helps to differentiate between the aspects of the conceptual scientific process and its practical implementation. The use of the correct terms will clarify which aspect are currently in focus and thus allow the reader to follow the development of the arguments more easily.

CHAPTER 3

Definitions and related work

Before analyzing the case in detail, it is necessary to introduce the basic concepts that refer to enhancing motivation. Therefore, this chapter defines motivation and explains the psychological mechanisms involved. Based on this, the terms feedback and gamification are introduced. Additionally, the chapter outlines how these techniques are connected to motivation and mentions the requirements for using their potential to enhance motivation. It finishes by taking a closer look at the specifics of implementing gamification in enterprise environments.

3.1 Motivation

Even though motivation is a concept that everyone understands on an everyday basis, it is surprisingly hard to describe it universally. An intuitive description can be found in Deci and Ryan: "To be motivated means *to be moved* to do something." [**?**, p. 54]. However, this simplistic view explains neither the origin of motivation nor its underlying mechanics.

To understand motivation, it is necessary to understand why humans act the way they do [**?**, p. 2]. A coherent explanation can be found in Puca and Schüler [**?**, p. 225]. According to the authors, humans usually act organized and goal-oriented with the goals depending on their needs, their previous experiences, and the stimuli available in their environment. The willingness to react to these stimuli in order to fulfill the need is called *motive*. For example, if employees anticipate to be promoted for fulfilling a difficult task, they will probably start working harder to solve the problem. The expectation to meet their goal is making them behave in a way they assume will bring them closer to the anticipated result. Naturally, motives are developed differently in every person depending on how rewarding the individual deems the fulfillment of a certain need (e.g. how important the promotion is to them). Motives are thus stable individual qualities which are decisive for the types of stimulus.

The combination of stimulus and motive creates *motivation* which triggers seemingly appropriate behavior for a certain time [**?**, p. 225]. In the example above, the possibility for a promotion acts as stimulus. Together with the employee's performance motive it creates the motivation to work harder in order to fulfill the task. It is notable, though, that motivation as such cannot be observed from the outside but has to be inferred by observing the individuals behavior [**?**, p. 14]. It also depends on the intensity of the stimulus and the

manifestation of the motive [?, p. 225]. Hence, the term motivation is an abstract umbrella term for a variety of processes that determine a continuous active behavior towards a goal [?, p. 14]

Due to the complexity of the concept, psychologists have devised a variety of models to explain how motivation is created and how it manifests in action. One of the most popular models with over forty years of research supporting it [?, p. 600] was proposed by Deci and Ryan and is called *Self-Determination Theory* [?]. Within this theory, the authors identify three basic psychological needs which humans strive to satisfy and which are thus a potential source for motivation [?, p. 68]. The need for *competence* is the desire to experience the efficacy of one's own actions [?, p. 427] which means that individuals want to feel that their actions have an effect on them and their environment. The second factor is *autonomy*, which is the "desire to experience self-regulation and integrity" [?, p. 87]. In other words, individuals want to be able to make their own decisions. Complementary to that, *relatedness* is the need to belong to a social group and being accepted by others [?, p. 87].

Like many other models, self-determination theory distinguishes between *intrinsic* and *extrinsic motivation*. Deci and Ryan define intrinsic motivation as "the doing of an activity for its inherent satisfactions rather than for some separable consequence" [?, p. 56]. Intrinsically motivated individuals act because they are inherently interested in an activity and not due to some external reward [?, p. 56]. The authors claim that this type of motivation originates in the needs for competence and autonomy because individuals who are able to act as they please and then experience the results of their action are inherently satisfied by this causality [?, 427]. This form of motivation has been widely recognized as powerful momentum and important factor in learning and working [?, p. 57]. However, it is difficult to elicit or enhance this form of motivation from the outside because it depends on the individual's motives and the stimuli provided by its surroundings [?, p. 58]. Therefore, it is only possible to facilitate intrinsic motivation by providing conditions where the individuals can act autonomously and competently [?, p. 70].

Extrinsic motivation, on the other hand, is "construct that pertains whenever an activity is done in order to attain some separable outcome" [?, p. 60]. An action which is extrinsically motivated is performed because of some external regulations that have been imposed on the individual [?, pp. 88–89]. Ultimately, the person acts to either achieve a certain goal, to gain a reward or to avoid punishment [?, pp. 88–89]. However, Deci and Ryan differentiate between different forms of extrinsic motivation depending on how well individuals accept the external regulations for themselves [?, p. 71]. The degree of internalization of a requested behavior causes a variety of reactions from unwillingness to passive compliance to even active commitment [?, p. 71]. Here again, the reaction depends on how much the three psychological needs and especially the need for relatedness is fulfilled by the request [?, p. 73]. Humans tend to internalize behavior far better if they are modeled by their social group or if they want to please people that are important to them [?, p. 73]. Thus, when trying to foster extrinsic motivation, it is important to analyze in how far the addressees of the motivation-enhancing methods are impaired in their autonomy, relatedness and competence [?, p. 64].

3.2 Feedback and motivation

Feedback is an ambiguous term that is used in a variety of contexts[1]. In human-computer interaction, it is understood as "communication with a user resulting directly from the user's action" [?]. This definition includes any form of output that a software system emits to its

[1] At the time of this thesis, the English Wikipedia page for *feedback* lists definitions for nine different fields, including biology, mechanical engineering and software (see `https://en.wikipedia.org/wiki/Feedback`).

users as a response to their input [**?**, pp. 105–106]. These outputs are generally delivered as audio signals, text outputs or visualizations.

The main purpose of feedback is to keep the user informed about the state of the system and its operations [**?**, p. 106]. The intention is to make the users understand what they are doing and how the system reacts to their inputs [**?**, p. 106]. Only if they are aware of that, they can work effectively and efficiently [**?**, p. 106]. Therefore, the ultimate goal of feedback is to increase the individual's performance [**?**, pp. 87–88]. It does that by providing adequate information concerning the process towards the goals that have been defined for the task the user is working on [**?**, pp. 87–88].

Both the idea of keeping the users informed and enhancing their performance is important when considering feedback to elicit motivation. From a psychological perspective, feedback facilitates motivation if it responds to the psychological needs [**?**, pp. 58–59]. By informing the individual about their progress, self-efficacy is supported which is essential when satisfying the need for competence [**?**, p. 463]. Thus, adequate feedback is a valid way to enhance motivation.

3.3 Principles of gamification

Similar to feedback, the term *gamification* is widely popular and used in different contexts and notions [**?**, p. 9]. A common definition was proposed by Deterding et al.: ""Gamification" is *the use of game design elements in non-game contexts*." [**?**, p. 2]. More specifically, it takes concepts that are usually included in games and applies them to a context where such elements are not commonly used [**?**, p. 276]. A simple example is the use of points and scoreboards in fitness apps. By providing rewards in the form of points the apps turn the subject of physical activity into a competition, which is a classical concept in games. This technique has been applied in a variety of fields, including health, education and in the workplace [**?**, p. 371].

It is important to understand that the result of gamification is not a full game as it is the case with serious games and games with a purpose [**?**, p. 276]. Designing a game with a purpose (GWAP) means to transform a problem into a game which solves the problem and entertains its players at the same time [**?**, p. 276]. For example, the platform ARTigo[2] provides GWAPs which make players describe artworks and tags the images through their inputs. Hence, the issue of tagging vast amounts of images is transformed into an enjoyable game. Serious games have a second purpose apart from entertainment [**?**, pp. 817–818]. It is often educational and intends to impart skills or knowledge while playing the game [**?**, pp. 817–818]. Stieglitz names specific training of firemen as an example [**?**, p. 817]. Contrary to that, gamification leaves the core service of the target application untouched but extends it for enhancing services [**?**, pp. 18–19]. These enhancing services are not strictly required for functionality but they make the system more attractive for its users and set it apart from competing products [**?**, p. 19].

Gamification is inherently a persuasive technology which intentionally utilize game-design elements to activate individual motives and thus influence user behavior [**?**, p. 2]. Introducing gamification methods aims at altering the environment so that at least one of the three psychological needs is addressed [**?**, p. 374]. In the example with the fitness app, the success of earning enough point to get a higher rank on the scoreboard relates to the need for competence. Hence, gamification motivates users to engage in an activity by satisfying their needs which the core service would otherwise ignore [**?**, p. 374].

Interestingly, gamification tries to elicit intrinsic motivation through extrinsic methods like rewards or social recognition [**?**, p. 277]. According to Deci and Ryan, it is possible to activate an individual's intrinsic motives by providing extrinsic stimuli [**?**, pp. 63–65].

[2] `https://www.artigo.org/`

Based on McGonigal [**?**], Blohm identifies four strategies for that: increasing user satisfaction, conveying optimism, facilitating social interaction and providing meaning [**?**, p. 277]. While increasing user satisfaction is manly achieved by highlighting the user's own performance, an optimistic view on their goals is usually communicated by emphasizing self-determination of the user's actions [**?**, p. 277]. By supporting social exchange within a community of users, gamification also provides the opportunity to support social interaction [**?**, p. 277]. Finally, introducing game design elements can work towards providing necessary context or a narrative which attributes additional meaning to the task [**?**, p. 277]. Implementing gamification elements according to these four strategies has been shown to be an effective way to raise intrinsic motivation [**?**, pp. 374–375].

Even though its positive effects are widely recognized, gamification has been criticized for reducing gaming experience to simply collecting points and badges [**?**, p. 18]. Moreover, concerns have been raised that the competition caused by levels or leaderboards might actually harm intrinsic motivation [**?**, p. 66]. However, there is currently no empirical evidence for such negative impacts [**?**, p. 72].

3.4 Gamification in enterprise environments

A special type of gamification is called *enterprise gamification* and describes the introduction of game design elements to working and learning processes within a company [**?**, p. 817]. According to Stieglitz, the central characteristic of this form of gamification is that it targets a company's employees instead of its customers [**?**, p. 817]. Nevertheless, the goal is the same: eliciting motivation and enhancing employee engagement by making a task enjoyable and thus ultimately increase performance [**?**, p. 7] [**?**, p. 346]. A famous example for this type of gamification is the SAP Community Network which is a platform for SAP professionals to exchange knowledge, seek help and solve problems [**?**, p. 25]. It successfully motivates users to actively contribute to the network by providing badges, points and leaderboards [**?**, p. 89–90].

It might not seem very intuitive to bring elements from games into the serious business environment. However, good gamification design does not distract the employees from their actual tasks but supports them through their enhancing functionality [**?**, p. 15]. Additionally, Kumar and Herger argue that people are already used to gamification in their everyday lives [**?**, p. 15–16]. Thus, employees will probably appreciate the motivating design resulting from gamifying work software [**?**, p. 22–23].

As working environments have their own social setting which already includes features like hierarchies and team work, the game design elements have to be tailored to this environment [**?**, pp. 818-819]. Moreover, extrinsic monetary incentives are an inherent part of this environment [**?**, p. 347]. Legal regulations like data privacy may also not be infringed [**?**, pp. 32–33].

Apart from these external issues, gamifying an enterprise software poses challenges in itself. It can only develop its motivational potential if it is adequate for the employees, the tasks they are engaged with and the goal of these tasks [**?**, pp. 30–31]. If an employee has to insert one hundred datasets into a database, for example, gamification could introduce intermediate goals to address the individual's achievement motive. Due to the fact that different types of game design elements appeal to different people, it is sensible to implement a variety of mechanics which address several different motives [**?**, pp. 374–375]. In the example with the datasets, accomplishing an intermediate goal can be advertised through a message to the team (social recognition) and by showing an animation with a positive message (achievement). Thus, applying gamification requires detailed analysis of the scenario and careful implementation with the target group in mind [**?**, pp. 823–824].

CHAPTER 4

Use case analysis

4.1 Use case description

4.1.1 Glanos, the DataSphere and the Gold Standard Annotator

Glanos is a company specialized on big data and text analytics. It offers several services including anonymization of documents, content tagging and document management.

Almost all of Glanos' services are based on finding and extracting entities from large text corpora. In order to anonymize a set of letters, for example, all person name and address entities (streets, postal codes, etc.) have to be found and then removed. While removing them is easy, retrieving them can be quite difficult. Glanos does that by employing sets of rules which describe typical contexts or grammatical environments of an entity. For instance, a postal code is often combined with a location or city, which makes the city the postal code's context (and vice versa). These rule sets, or grammars, are then included in a predictor and run on the corpus to extract all entities the text contains.

To monitor and improve this process, Glanos provides an application called the DataSphere. It is a collaborative tool with a web-based graphical interface which enables the users to define the semantics of new entities, specify basic rules, and run the predictors.

Within the DataSphere, the results of an entity extraction are displayed in an editor called the Gold Standard Annotator (or GSA for short). Herein, all found entities are highlighted as annotations. If the users click on an annotation they are provided with detailed information like the entity type, specifics on the extraction and additional meta data from other sources. They can also add annotations for entities that have not been found by the predictors yet. These additional annotations serve as feedback to the developers so that they can improve the predictors according to the user's requirements. Thus, computer linguists and customer users can work as a team to adjust the result to the customer's needs.

4.1.2 The necessity of human participation

While the process as described above can be automated in many ways, there are two parts which require human participation: finding the contexts to define the rules and evaluating the resulting entities' semantic correctness. The first one is almost self-explanatory: the system cannot find a certain entity without knowing what to look for. Thus, human computer

linguists have to provide and update input in the form of grammars.

The necessity to evaluate entities manually is probably less obvious. As an example, let us assume that one rule has the structure *[company entity][indicator of possession][product]*. Thus, the phrase "Google's phone Pixel" is matched by the rule with *Google* as the company entity, the *'s* being the indicator and *Pixel* being the product (*phone* would be an additional specifier to the product). However, the phrase "Google's pixel rendering technology" will also be matched, even though *pixel* does not refer to the product here. The underlying problem is that a set of syntactic structures such as the rule in the example cannot necessarily be mapped to a specific semantic concept like the product. As explained in the last section, the rules of a grammar define patterns which typically denote a certain entity. However, they cannot assure that the match will always belong to the semantic concept. The example shows nicely that just because both sentences fulfill the syntactic requirements to be matched by the rule the results do not necessarily belong together semantically. It is therefore necessary to check if the entities found by the grammars do also belong to the defined concept.

As becomes obvious from the example above, it is fairly easy for users to decide whether or not a match belongs to a certain entity. As long as they possess adequate language capabilities and a good understanding of its semantic concept, differentiating between the product and the technology will not be a problem. This, however, yields another issue: as the next section will show, the task is perceived as uninteresting, repetitive and boring. The resulting lack of motivation and concentration inevitably heightens the potential for mistakes. In turn such mistakes might impair the rest of the process and could result in extractions with lower accuracy than required.

It might seem like an option to apply machine learning to further automatize the process and thus avoid user mistakes. By training an adequate machine learning model, for instance, semantic differences could be resolved without human interaction. However, this approach would require providing a sufficient amount of training data which again would have to be annotated by human programmers or users. Moreover, the training would have to be done for different use cases and in a variety of languages which implies even more manually annotated data. In consequence, this strategy would increase the development process by another work-intensive tool while not solving the problems associated with unmotivated users entering potentially wrong data.

Instead of spending extensive efforts on training a machine to do a human's task, it seems to be more sensible to make the task more attractive for the people who are assigned to it. The approach chosen here does that by supporting motivation and giving adequate feedback.

4.2 Analyzing potentials for motivational support

A key question to introducing motivational support systems in the Gold Standard Editor is which of its parts require the most attention. It can be answered partly by analyzing the current state of the system for already existing motivation-enhancing methods and potential improvements therein. However, this analysis cannot include the issues that occur while actually using the editor. Thus, it is necessary to ask the users of the system for their experiences to get the motivational problems and usability issues that root in real-life usage. Both of these steps are made hereafter.

4.2.1 System analysis

Before focusing on the motivational support of the Gold Standard Annotator, its is necessary to describe its appearance and basic use. Its interface is divided into three areas (see

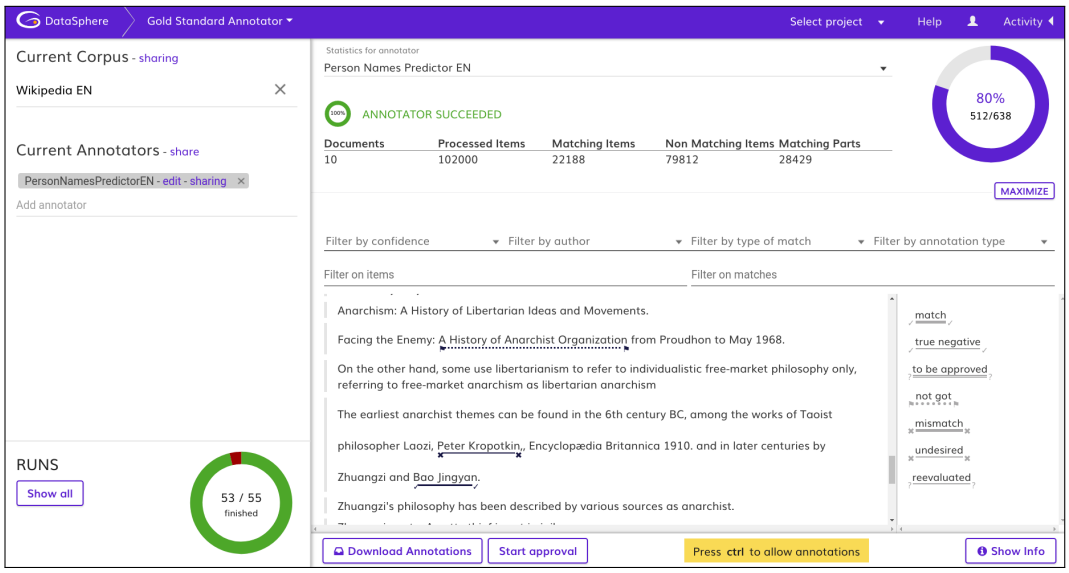

Figure 4.1: Main view in the Gold Standard Annotator

also figure 4.1). On the left, users can choose a text corpus and an annotator. Alternatively, they are able to choose a combination of those two from the latest run menu by clicking on the ”Show all” button under ”Runs”. The major part of the screen is reserved for displaying the corpus in a scroll panel (bottom right) and navigating through the results (top right).

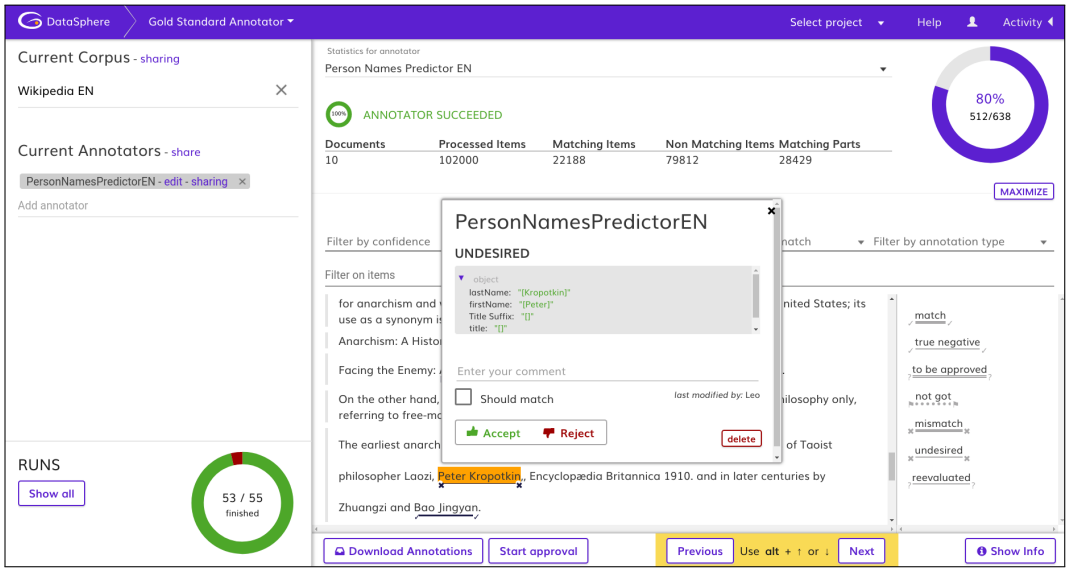

Figure 4.2: Evaluation panel with extraction details

Having selected the corpus and the annotator, the user can scroll through the corpus and inspect the annotations which are marked by underlining the respective text parts. As figure 4.2 shows, clicking on one of these snippets opens a dialog which contains details on the extraction and allows the user to accept or reject the annotation. Performing one of these actions changes the underlining to have either ticks or crosses at both ends. This

markup is meant to indicate to other users whether the annotation has been deemed correct or not.

The corpus can also be filtered for a certain type of annotation by entering values into the fields above the scroll panel. Thus, it is possible to search for annotations with a certain keyword or of a certain type for example.

Finally, the top right part show statistics concerning the last annotator run on the corpus. Apart from indicating whether the annotator ran successfully through the entire text, the number of processed documents and items (e.i. sentences) is listed. It also shows how many items contain a match and how many do not. The last number on the right, titled "matching parts", is the total number of all found matches.[1] In the top right corner, there is also a diagram giving a percentage. Its corresponding tooltip reads "This shows the quality of the annotations, calculated as the ratio of the matched annotations to the total number of annotations".

This diagram and the changing underlining are the most obvious forms of feedback that currently exist. While the markup focuses on indicating whether a single annotation is correct, the diagram intents to visualize how many of the intended matches are currently found by the annotator and how many are not found or mismatched.

What is difficult about this representation, though, is its message. It does not give feedback on the percentage of annotations that have been evaluated, but on how good the annotator is doing its job. Misunderstanding that, the users might not initially grasp how to influence that percentage. Intuitively, they might try to evaluate as many annotations as possible to get closer to 100%. However, it also requires the annotator to not mismatch elements to reach that goal and that is something the users cannot influence. Thus, accepting or rejecting might actually decrease the percentage, as more mismatches might occur. This, of course, can be highly demotivating for the users.

The statistics can also be considered as a type of feedback because the numbers might change from one annotator run to another. Thus, the user can be informed on changes in the annotator's performance. Unfortunately, there is no historic data so that this change could be visualized. It is important to stress that here again the feedback is not on the users' efforts but on the annotator's behavior.

Considering all this, the feedback on the users' performance is quite limited. Their only indicator on the work still to be done seems to be whether they can still find annotations to evaluate. Also the only goal they are given can be described as annotating the entire corpus. This task often includes reviewing thousands of sentences which in itself might hardly be considered motivating. It seems like the application still has potential for improvements on motivational affordances. However, before taking any further steps, it requires to ask the current users of the system whether their impressions align with this analysis.

4.2.2 Interviews

A common way to record user experiences is through interviews. This approach was chosen here as well. The semi-structured interviews consist of a series of core questions which were brought into a logical sequence. Thus the topic could move from general information to typical work flows and finally to an evaluation of the system. Nevertheless, there would still be the possibility to rearrange or skip questions during the interviews.

As every participant was interviewed separately with only one interviewer present, recording the answers had to be done in short notes. This method was chosen to not impair the communication during the interview was while still being able to gather all relevant points. The notes were extended directly after the interview to keep external influences minimal and preserve all information that could not be recorded during the conversation.

[1] Naturally, the number of matches is higher than the number of items containing matches as one item (or sentence) might contain more than one match.

Two of Glanos' employees were interviewed.[2] Both work with the Gold Standard Annotator (GSA) on a regular basis but their perspective is quite different. The first interviewee is Glanos' Scrum Manager who is mainly responsible for coordinating the developers' tasks and for communicating with management and customers. He was chosen for the interview as he is a regular user of the Gold Standard Annotator himself but only has a conceptual view on the development of the system. Thus, it could be assumed that he would not get caught up in implementation details but rather focused on the intended use of the editor. Additionally, his perspective on the usability and motivational elements includes feedback from the customers. Therefore, he would probably be more focused on improving the system based on his own experience and on customers reports.

The second participant is a computer linguist whose main task is to develop new grammatical rules so that specific entities can be found. After formulating suitable rules describing a certain pattern, he checks whether his rules work as expected. Therefore he runs them (within an annotator) on the given corpus and then evaluates the resulting annotations. This process repeats until all required entity types are found. Consequently, the linguist uses the Gold Standard Annotator extensively which made him an interesting candidate for an interview on the motivational affordances of the system.

Both interviewees had a similar user profile. At the time of the interview, they had been using the GSA for about 18 months but their time spent with it varied as it depends on the projects they were working on. Nevertheless, both stated that they used it with the intent of evaluating existing annotations and finding contexts in which specific entities appear regularly. Therefore, they most commonly used the keyword search and the filters.

Although both participants emphasized the importance of the task, they also admitted that it was exhausting, tiring, and not very interesting. Neither of them defined another goal apart from processing as much of the corpus as possible. Consequently, they stopped when they did not discover new contexts or annotations, or when they felt that they have put in enough effort. The scrum manager additionally expressed that the missing document structure in the corpus made the content sometimes look arbitrary. Additionally, it was sometimes hard for users with no background in computer linguistics to assess whether they were annotating the corpus correctly. This led to an insecurity which made it even harder to concentrate on the task for a long time.

The interviewees confirmed that their primary source of feedback was the change in the annotation markup. They also initially had trouble understanding the diagram in the top right corner and sometimes were still not sure how to manipulate it as they desired. Thus, they deemed it little helpful in respect to personal feedback. Overall, they stated that the amount of feedback was insufficient for a users to evaluate their own efforts.

Another point the participants agreed on was that it would be interesting to see other team members' contributions. As they currently coordinating the work via annotation type, it was proposed to split the work into blocks and then assign them to different members to further distribute the workload among the team. They also favored some sort of visualization either via live-updates or diagrams. Both expressed that they welcomed any additional ideas supporting teamwork.

One particularly alarming point was the scrum manager's answer to the question whether customers liked working with the Gold Standard Annotator. He stated that they would do the annotating because they know that it is important and necessary but that he is not under the impression that they would really like it. After all, it is a tiring procedure in which they cannot be sure to do it correctly which he blamed partly on missing definitions and partly on the knowledge gap between the grammar developers and the customers. The problem with this point is that users might wander of to products they enjoy using even if these are not working quite as well. Therefore, an unsatisfactory user experience might

[2]The notes on both interviews can be found in the appendix.

even be a threat to Glanos' business in case a competing product is placed on the market.

As all of the points brought up in the analysis were confirmed through the interviews, it is save to say that the Gold Standard Annotator could greatly benefit from a variety of improvements concerning motivational affordances. These should primarily target the individual user's motivation to keep them interested in the task at hand. However, it should also include cooperative elements as the interviewees expressed their interest in teamwork. This would also support the Gold Standard Annotator's intended purpose as a multi-user tool.

CHAPTER 5

Conceptualizing motivation-enhancing methods

As applying gamification to the Gold Standard Annotator seems to be a valid approach overall, the following proposes a concept for introducing gamification and feedback to the Gold Standard Annotator (GSA). The first section outlines the approach used for designing adequate motivation-enhancing methods. It also maps the findings from the case analysis to the psychological needs that are currently left unfulfilled by the GSA. Hereafter, guidelines for choosing adequate design elements are set up and several ideas for gamification methods are described and discussed.

5.1 Foundations for enhancing motivation in the GSA

5.1.1 Approach

There are multiple models for creating adequate motivation-enhancing methods for a specific scenario (see for example Stieglitz [**?**] or Ruhi [**?**]). This thesis roughly follows the approach proposed by Kumar and Herger [**?**] which consists of five steps:

1. understand the players by getting familiar with their context, their tasks, and their preferences
2. understand the mission by analyzing the business processes, their desired results, and the tasks involved
3. understand motivation and the human drivers that enable it
4. create and apply the game mechanics (the UI elements that the user interacts with such as progress bars and badges)
5. manage, monitor and measure the motivation, the outcome and adjust if necessary.

Fortunately, some of these points have already been addressed in the last chapters. The interviews (section 4.2.2) compiled all necessary information to understand the users' situation and behavior. Section 4.1 described the purpose of the entity evaluation process, outlined its expected outcome, and summarized the necessary tasks. Thus, the case analysis already addressed the first two steps.

The processes for eliciting and facilitating motivation was already mentioned in the related work part. However, it should be discussed in more detail for the case at hand.

5.1.2 Psychological needs involved

The key to designing mechanisms for enhancing motivation in a specific context is to identify which of the three psychological needs competence, autonomy, and social relatedness are currently left unfulfilled [**?**, p. 374]. By increasing the satisfaction of these needs users have positive attitudes towards the application and feel motivated to keep doing the activity [**?**, p. 374].

In case of the Gold Standard Annotator, the results of the analysis indicate that the users' need for competence is widely ignored. The interviewees said that it was difficult for a user to know whether the evaluations were correct, especially if they had no background in linguistics. The correctness of many entities depends on their exact definition. For example, users might face a match like *architect Frank Lloyd Wright* which has the type *personname* and lists *architect* as *title*. While it is easy to see that the match indeed belongs to the group of personnames, it depends on the definition of the entity whether *title* includes job titles or is reserved for nobility and academic titles only. To make a competent decision, users need to be aware of these definitions.

Another element that has a negative effect on the feeling of competence is the predictor performance diagram in the top right corner. As already explained, the feedback from this diagram is easily misinterpreted by the users. Moreover, the interviews revealed that its message is not intuitively understandable and neither is how to influence the percentage displayed. These factors are major impediments on experiencing self-efficacy which is of key importance for satisfying the need for competence. Therefore, the diagram should be removed or replaced.

The most severe factor in terms of competence, however, is the missing definition of goals. Consequently, no progress towards these goals can be displayed. Here again, self-efficacy can only be experienced if the users know what to do and get regular updates concerning their efforts on reaching the target.

The case analysis also shows some deficits concerning the need for social relatedness. While the interviewees clearly state that they would appreciate coordination and increased collaboration with their colleagues, the Gold Standard Annotator does not support any of this. It might even happen that two people process the same annotations as they are not made aware of the evaluations their coworkers put in. Despite the fact that the result of the evaluation process is obviously a cooperative accomplishment, the missing support for social interaction during the process somewhat undermines the editor's intended purpose as a collaborative tool.

In terms of the third psychological need, autonomy, no explicit infringements were brought up during the interviews. None of the interviewees felt that they were limited on how to approach the task. Both stated that they focused on the vague goal of evaluating a sufficient number of entities but that they were not limited on how to achieve it. Moreover, they both expressed the importance of the task. As meaningfulness and decision freedom are key requirements for autonomy [**?**, p. 374] and none of these two have been reported as problematic, there is no reason for implementing gamification methods which focus on this need.

5.2 Potential methods: selection, design and evaluation

With the motivational premises being clear, it is time to tackle the fourth step in the approach and start conceptualizing motivation-enhancing methods for the Gold Standard

Annotator. After outlining some restrictions and general guidelines, a selection of concrete methods is presented.

5.2.1 Restrictions and guidelines

Before starting to collect ideas for adequate motivation-enhancing methods it is necessary to clarify the boundaries for the resulting game design elements. Following the idea of gamification as an enhancing service [**?**, pp. 18–19] (see also 3.3), the changes are restricted to implementing new elements on top of the existing interface. It also means that the entity evaluation process may not be restructured or otherwise changed to fulfill the requirements of the gamification methods. Hence, the motivation-enhancing methods have to work with the data that is provided by the editor and its environment. Alternatively, it can measure or collect data on its own, of course, but it may not rely on the core service to issue additional data for its own purposes.

Unfortunately, this restriction makes it almost impossible to increase the users' confidence in the correctness of their evaluations. Techniques that involve confirming already evaluated entities to gain rewards, for example, require adjustments in the process and the addition of an extra state to the entity object. An entity would then transition from "matched" to "evaluated" to "confirmed". However, these changes affect the core service and are thus not allowed.

Although this restriction limits the range of methods to apply, it still does not answer the question which kind of gamification and feedback techniques should be introduced into the GSA. As a first step, it is sensible to distinguish between team level and individual level. While the team level primarily targets the need for relatedness through social interaction and cooperation, the individual level focuses on the competence and performance of a single user. Thus, the techniques for the individual level visualize goals and progress. Optionally, they can also aim at triggering competition of the individual with themselves by encouraging them to trump their last number of annotations, for instance.

However, the concept will refrain from introducing competition on the team level. Even though competitive techniques like leaderboards have been reported to be successful (for a comprehensive list of examples see Hamari et al. [**?**]) they are hardly appropriate for a collaborative task like the one at hand. The overall goal of the process is to evaluate a large number of annotations correctly and to keep the number of errors as small as possible. Introducing competition within the team means to create not only a motivational enhancer but also social pressure. This feeling of having to keep up with the team members can result in stress and thus be demotivating. It also might lead to lower-quality output as users might try to move up in rank with all means possible. If the leaderboard focuses on number of evaluations, for example, users could try to process as many annotations as possible without paying attention to the quality of their output. To avoid such effects, gamification methods which are based on competition between team members will not be used here. Instead, the team methods should focus on supporting collaboration and distributing the workload to multiple employees.

Having set up these guidelines and restriction helps to chose adequate methods for motivation-enhancement.

5.2.2 Specific method design

After reviewing existing gamification methods in a variety of publications such as Hamari et al. [**?**], Kumar and Herger [**?**], and Raftopoulos et al. [**?**] a collection of thirteen potential game-design elements was established. Six of these were deemed suitable according to the restrictions and guidelines specified above.

The first of these elements is the progress bar. It is probably the most straight-forward way to simultaneously define a goal (making the bar reach 100%) and give feedback on the effort that has already been put in. In terms of motivation, it conveys the self-efficacy to the users who see the progress bar moving each time they contribute to the goal. It also bears the promise of success once the target value is reached. Working towards a defined goal and meanwhile monitoring the own progress can therefore conduce to satisfying the need for competence. Another advantage of progress bars is that they can be used both at team level and at individual level. As long as the progress towards a goal is displayed and team members are not ranked by their contributions, the element can unfold its motivational potential.

Within the Gold Standard Annotator, the progress bar is used to display the number of entities that have been evaluated. The goal is to process as many annotations as is necessary to reach the predefined target number. However, if the target number is too large, it might be highly frustrating to see the progress bar move so slowly. Thus, intermediate goals are introduced that allow to put in only a fraction of the effort and still get a feeling of success. Games like Candy Crush[1] have been using this technique for years.

Badges are the second game-design element that focus on enhancing motivation through conveying competence. They are virtual representations of a player's achievements and therefore address the motives of achievement and collection [**?**, p. 72]. The advantage of using badges is that each batch can have its own distinctive goal. These goals can be self-competitive, relative to an input, or fixed to a specific value. Within the Gold Standard Annotator, they are used to gain a certain fixed number of manual annotations but also to make users evaluate a range of different types of entities. Badges are also awarded for beating the number of evaluated annotations from the last run. By applying this one form of gamification, it is thus possible to define a variety of different goals and diversify the challenges for the users which makes it more interesting for them to play.

The need for social relatedness is addressed through a team challenge. Its visualization is a combination of three progress bars, each with a target number for a specific entity type. The overall goal is to collectively evaluate enough entities of these three types to reach the 100% mark on each progress bar. The target values will usually be quite large so that the goal can only be reached if each team member contributes their part. Thus, this game design element intends to foster cooperation and task coordination within the team. The motivation is elicited by showing the efficacy of a collective effort and by sharing success with the other members.

Finally, the feeling of success upon reaching a goal does only emerge if the user recognizes his achievement. Thus, each achievement has to be highlighted through a distinctive message. This message praises the user's achievement in an optimistic tone. It also should be clear which goal has been reached. Hence, this form of feedback contributes to the motivational affordances of the editor through conveying optimism and emphasizing the accomplishment.

There are two other game-design elements that had to be discarded for technical reasons later but are still mentioned here to keep the original selection complete. One is a graph that visualizes the contribution based on the entity hierarchy. The *address* entity, for instance, is are built by combining basic entities like *street* and *location*. Thus, evaluating the base entities indirectly contributes to confirming the high-level matches. The graph would have visualized these dependencies and the user's contribution with in the hierarchy. The idea was to raise awareness for the importance of the task and hence attribute additional meaning to the users work which can have a motivating effect. Unfortunately, the data in the study did not contain enough of these composite entities to make this method effective. The second design element which was not implemented is a team competition.

[1] Candy Crush is a casual game developed by King.com Limited. For more information visit `https://king.com/de/game/candycrush`.

Even though the guideline specifies that no competitive strategies should be used within the team, it would have been possible to let each teams rival with others for higher ranks or better performance. This might have had an additionally motivating effect especially for competitive players. However, it also entailed simulating a second team within the study which requires complex implementation but is still not the same as having a real team to compete with.

Conceptualizing these motivational techniques and design elements concludes step four of Kumar and Herger's approach to a great extend. The only missing details on the actual implementation are discussed later (see section 7.4). The next step is therefore to work out an approach for measuring to what extend these methods actually improve motivation and whether these improvements have an impact on the quantity or quality of evaluated entities.

CHAPTER 6

Concept for the study and methodological approach

While implementing the concept for motivation-enhancing methods is technically feasible, the question remains whether the techniques are effective in the way the research question intends them to be. Thus, a study is conducted in order to answer this question. This chapter lays out the concept of the study. It also lists the technical requirements for the study's application, outlines the questionnaire design, and presents the methodology for evaluating the resulting data.

6.1 Study concept

Both the research question and the hypotheses are making claims which compare individuals who are exposed to motivation-enhancing techniques to individuals who are not subject to such methods. The study must reflect this duality by providing two comparable groups of participants. Therefore, a comparative study design is used.

The minimum number of participants was set to 16 so that each group has at least 8 members. Smaller groups would diminish the meaning of the results and make them less generalizable. The participants were recruited from volunteers. Although it would be more realistic to have Glanos' employees or other users of the DataSphere participate, it is questionable whether enough participants could be gathered. Glanos' team is small and the company would probably refrain from asking their paying customers to pass on the study to their staff. Thus, using voluntary participants was the better option.

In order to answer the hypotheses, the study has to gather both the interaction data from the entity evaluation process and information provided by the participants concerning their inner state of mind. Only if it records how many entities are evaluated by each participant and whether these entities were evaluated correctly it is possible to analyze for differences in quantity and quality. However, none of these values indicate whether the motivation differed. Thus, it is also necessary to ask directly about the participants' thoughts and feelings.

The necessity for both types of data is met by structuring the study in three three major phases: tutorial, entity evaluation (with or without gamification), and a questionnaire. The first phase trains the participants by explaining the basic process and showing them the actions they have to perform. This should give them a good understanding of the task and teach them the skills they need to fulfill it. They then have to apply their new abilities by

processing entities during the main phase. All of their actions are recorded and compared to Glanos' original entity data, thus deriving whether an evaluation was correct or wrong. Additionally, all manually added annotations are saved in a separate category. The third phase then gathers information on the experiences and attitudes toward the process in the evaluation phase. It does that with a questionnaire. Its questions are either multiple choice questions, select questions or open questions and focus on usability, motivation and evaluation of the motivation-enhancing techniques. Their specific design is discussed in section 6.2.2.

An interesting issue is how long the evaluation phase can last before participants start abandoning the study. Here, its duration is set to twelve minutes. Experiments with the finished application showed that the feeling of repetitiveness and boredom started around minute seven or eight. On the other hand, people would leave in the middle of the study if they were in this state for too long. Twelve minutes is a reasonable compromise but participants are free to go on evaluating if they like.

With the study setup being defined, the next step is to outline the requirements for the application that the study eventually runs in.

6.2 Requirements and design

6.2.1 Requirements for the application

The main phase in the study concept specifies that the participants should be able to evaluate and add entities in an editor that resembles Glanos' Gold Standard Annotator. This means that an application has to be provided which lets participants interact with the corpus in the same way that the GSA does. Naturally, such an application is subject to certain technical requirements.

First of all, the application used in the study should behave as close to the Gold Standard Annotator as possible. This means it should provide both its two main use cases, which are evaluating given annotations and manually adding new ones that have not been found by the predictor so far. It also entails that the forms of existing feedback have to be transported to the new application. Thus, the gamified version has to include colored markup for example.

Secondly, the study has to be easily usable for the participants. This means that it must be publicly available and easily accessible. Moreover, it has to provide some sort of introduction and training that prepares the participants for the tasks at hand. Only if they understand precisely what they have to do and do not have to go through a lot of trouble to do it they will use the application the way it is designed to be used.

However, this public accessibility entails one more requirement: there must be mechanisms to prevent or detect attack or fraud. It is crucial that the system is built and run in a secure way to ensure that the application cannot be hacked or hijacked. Additionally, cheating in the study data must be easily detectable. The difficulty here is to differentiate between outliers in the data and unusual performance of one participant.

Chapter 7 outlines in detail, how these requirements were met in the application. As stated in the study concept, it records both the number of entities that are processed by a user and whether the evaluations are correct compared to Glanos original data. It also gathers the participants' answers in the questionnaire.

6.2.2 Designing the questionnaire

The problem with recording psychological states like motivation is that individuals are seldom able describe or rate them. Hence, asking people "Are you motivated?" or "Are

you more motivated than this person?" will not be effective. However, they are usually able express whether they agree with specific statements about their state of mind. Similar to the strategy shown in Tuan et al. [**?**] (among others), each oft the questionnaire's items focuses on a single psychological factor such as feelings of success, self-efficacy, or social relatedness. Combining the answers on one motivational affordance technique will then indicate the level of motivation the participant gained from this technique.

The questionnaire is built using three question types. The majority of questions are multiple choice questions. Participants can choose an answer from a five-step Likert skale which varies from *strongly agree* to *strongly disagree*. The advantage of this question type is that its results are easily comparable to answers from other users. The second type of question are select questions which allow choosing one answer from a predefined set of responses. Here again, comparability is kept up while the range of answers can vary from case to case. Finally, open questions allow the participant to put in free text. These answers are seldomly comparable but can give detailed insight into the users' feelings and opinions.

The questionnaire as such is structured into logical units which consist of several questions. Each of these question groups focuses on one specific topic. The questions cover different aspects of the topic and hence attempt to give a comprehensive insight into the participant's attitude toward the topic. In the following, each topic of the questionnaire will be presented shortly.

The first question group concentrates on the application's usability. Its purpose is to understand whether participants had problems using the application and if they understood the task. This information can be useful when analyzing their performance. If a user could not interact with the application properly, they probably will not have a large number of evaluations. Thus, recording the participants' view on the program can help to put the results into the right perspective.

Hereafter, the participants assess their own effort. The question group therefore contains questions on whether the individuals could maintain concentration during the task and whether they think that they evaluated the annotations correctly. It can be expected that users who were subject to gamification will state that they were able to better maintain concentration due to the additional feedback they received from the application. Thus, this assessment tells a lot about the participants' perception of the task.

The next group gathers data concerning the participants' motivation for the task. As mentioned above, it consists of questions that focus on single motives like satisfaction with one own performance. It also asks questions on the three psychological needs competence, social relatedness and autonomy. The questions cover both positive and negative factors. For example, the question on whether the participant had feelings of constraint due to the task assesses if negative effects on autonomy could be avoided. Positive impacts on the need of competence are covered by asking whether the participants have an overview on the efforts they have put in, for instance. The expectation is obviously that the gamified group puts in higher scores for positive effects and same or lower scores for negative effects.

The question group on feedback is also issued to both groups. Although only the group subject to gamification has specific feedback mechanisms installed, the reference group experiences feedback as it currently exists in the Gold Standard Annotator. By seeing the markup change upon entering evaluations and scrolling through the text they are expected to have some sort of feeling for the progress they are making. Here again, the questions are trying cover both positive and negative effects of the feedback the participants experience. Comparing both groups is anticipated to reveal that users who are exposed to feedback will express a more positive attitude than their peers who are not subject to these methods.

The group subject to the motivation-enhancing methods is hereafter questioned concerning their experiences with the game design elements they have interacted with during the main phase. Both the badges and the team challenge have their own question group. This enables an independent analysis of each game design element. Again, the questions

cover a range of positive and negative experiences such as whether the badges or the teamwork were perceived as distracting or helpful. An overall positive evaluation is anticipated in these groups as well.

Even though the comparison group cannot be asked about the gamification methods, they still state their opinion on the potential for motivational affordances in the editor they just worked with. Therefore, they are asked whether they would have liked to work in teams, if they had needed more feedback, and if they would have preferred to specialize on a specific entity type. The intention behind these questions is to find out which of the psychological needs are left unfulfilled. If this data confirms the lack of stimulation of competence and relatedness, it could end up supporting the gamification concept even more.

Both groups have then the opportunity to state their opinion on the application's advantages, drawbacks and future potentials. These open questions are intended to let participants express their feelings freely and thus gather information that has not been covered by the questions above.

Finally, the questionnaire gathers some user information such as age group and occupation group (employee, public servant, student, etc.). It also asks how often participants are working with data management systems and how they deal with routine work. Moreover, they are asked how often they play games and specifically digital games. Thus, this group provides necessary background information on the participants to being able to interpret the results. If an individual does not play games regularly and is rarely occupied with data maintenance task, this person might be less prone to gamification. Hence, their performance might be different from regular players. Thus, having this information will help to find reasonable explanations for different behavior.

The complete questionnaires for both groups can be found in the appendix 9. The tables also include the answers given by the participants. Although the numbers there might already be insightful, they must be evaluated according to scientific standards.

6.3 Evaluation method

With both evaluation data and questionnaire answers being available, it is necessary to define a methodological approach for evaluating these datasets. As most of the data is in numerical values or can be transformed to it, statistical methods are a valid choice for that. To keep this section as concise as possible the specific calculations are not outlined here.

Several statistical values are calculated to gain insights into the users' tendencies and behavior. The first of these values is the arithmetical mean which describes a general tendency of the users' answers to a question or a data item. The mean is often used together with the deviation which indicates how different the values are from the mean value. Combining these two values already indicates a tendency of the study group.

To evaluate whether the result is statistically significant, the p-value is calculated. This includes formulating the null hypothesis and alternate hypothesis. Due to the expectation of a small sample size (i.e. a small number of participants), the threshold for significant values is set to $p = 0.9$. Thus, if the calculated value is bigger than p, the result is seen as statistically significant.

In order to compare mean values of two groups, the two-sample t-test is used. Here again, the null hypothesis can be rejected if the calculated t-value exceeds the critical t-value. Intuitively, that means that there is a significant difference between the two groups which is likely to be caused by something other than coincidence.

Applying these methods to the data recorded during the main phase will not cause any problems, because it is already in numerical values. However, the answers from the questionnaire are not. As most of its items are multiple-choice questions, the corresponding

Likert-scale is transformed to numerical values with *strongly agree* being mapped to 5 and *strongly disagree* being mapped to 1. Thus, a neutral attitude towards a question would be 3. Consequently, every value above indicates positive attitude and every value below suggests negative attitude towards the question. It is then possible to calculate mean values for each person's answers in each question group, which indicates the persons attitude towards the topic. Accumulating these individual mean values within the participant group then shows the groups tendency on the subject which can be compared by running the t-test.

Finally, there is one problem with the Likert-scale items in the questionnaire. Even if the statistical analysis reveals the expected effects, it remains unclear if the scale was adequate concerning the question. Theoretically, it might be possible that the items within the question group are not delineating the topic as anticipated. Thus, the Cronbach's alpha coefficient is used to calculate the internal consistency of the scale and is therefore a measure for reliability. A coefficient's value above 0.7 is generally seen as acceptable for a group of items. It indicates that the items are in correlation with each other to an extend that they belong to the same topic. Hence, the scale is viewed as adequate for the topic.

Having defined the study setup and the tools to record data, the next step is to implement this concept into a usable study application. This application not only has to adhere to the guidelines for implementation laid out above, but it also has to take into account that data that it runs on.

CHAPTER 7

Application design and implementation

The most straightforward way to meet the requirements listed in the last chapter is to design a JavaScript web application, which is later hosted and thus made available to the study participants. It can be designed to behave like the original program while still being easily accessible through a web browser. Participants are trained for the main task with a multimedia tutorial. Moreover, it is possible to secure it to the current standards and run the backend within the secure university network. The details concerning its architecture, the data it uses, and its key concepts are the content of this chapter.

7.1 System architecture and setup

As JavaScript has become one of the most popular programing languages for web development over the years, programmers have come up with various ideas on how to build websites. The architecture which is used to create the application for the study is called the MEAN stack. This acronym describes an architecture which consists of an Angular frontend, a Node.js server built with the Express.js framework, and a MongoDB database for persistent data storage.

There are several reasons for choosing this architecture. First of all, it is a standard form of building modern web applications and therefore well-documented and well-tested. Thus, following this architecture ensures that its frontend and backend will be compatible. It is also a good basis for secure web programming.

Another reason is that the components fulfill the requirements of the application. One major aspect is that the program receives its initial data from a foreign source, which usually entails costly data cleaning and transformation. Glanos provides the original data in the form of JSON documents, which can easily be imported to the document-based MongoDB. Data transformation procedures can hence be kept to a minimum.

Moreover, the MEAN stack's frontend framework, Angular 5, has major benefits to it. It offers non-blocking event handling which enhances the application's performance by continuously being available instead of waiting for responses. Moreover, it provides a modular concept for user interface elements can be extended or grouped easily. By adding libraries like Angular Materials[1] a great variety of standard user interface elements such

[1] For detailed information, see https://material.angular.io/.

as dropboxes, menu bars and dialogs are available. Angular thus enables the programmer to build modern, asynchronous and secure web interfaces in a manageable and testable way. Thus, it is a good choice for building an editor for entity evaluation, which has to dynamically load new text elements on demand and display them with their annotation markup. The details of the annotations' entities are nicely displayed in the dialogs provided by Angular Materials.

As most of the logic is located in the frontend, the backend's main purpose is to receive the requests sent by the frontend and transforming them into database queries. Express provides all necessary functions to run such a backend within the Node environment.

While the architecture is important from a technical point of view, it does not specify what the users actually are confronted with while using the program. Hence, it is time to have a closer look at the data they will have to work with.

7.2 Specifics of the dataset

Getting familiar with the data that is used in the application is important for two reasons. First, knowing the data structures is essential for writing effective and efficient algorithms to process it. Second, it is a necessary precondition for finding adequate motivation-incentivizing methods. Only if the data supports a method like finding a a great variety of different instances, for example, this method can actually be applied. Therefore, the following section describes the data structures used in the original data, the preprocessing which is necessary to fit it to the application's purposes, and the final outcome.

7.2.1 Data structure

Glanos delivered the original dataset in the form of JSON documents which contained either text items or annotations of a certain type. The text as such is a randomized sample of sentences taken from German Wikipedia pages. It includes sentences from articles, citations and editors' comments. This variety makes it a perfect for testing predictors because it contains different forms of language, syntax and grammar. This use as a test corpus also means that most of Glanos' predictors had produced annotations in the past. These annotations are now read from the documents and are thus available for the study.

Before simply importing the entire dataset, though, it is worthwhile to have a look at its internal structure. It is built from objects of the types *CorpusItem* and *PredictorAnnotation*. Each *CorpusItem* holds a single sentence and some necessary metadata like the sentence's position in the corpus. To display the complete text it is thus necessary to sort the *CorpusItems* according to their positions and concatenate their sentences. A *PredictorAnnotation*, on the other hand, contains a reference to a *CorpusItem* and all information of an annotation. This includes its start and end positions in the sentence, the annotation type and the extracted information which is later displayed in the editor when evaluating the annotation.

Even though this modular structure creates some overhead, it was specifically designed by Glanos because it is flexible when running multiple predictors on several corpora. Its main advantage is that it separates the corpus text from the annotations found on it. Therefore, each *CorpusItem* can be referenced by several different *PredictorAnnotations*. These can be edited or deleted without impairing the text as such. It also allows to display a chosen set of annotations by simply loading all their respective *PredictorAnnotations* in combination with the referenced *CorpusItems*. However, this flexibility requires joining *CorpusItems* and *PredictorAnnotations* before being able to display them together.

All in all, the original data can be deemed suitable for the intended purpose of letting volunteers evaluate different types of entities. However, its complex structure and its initial

number of 15 different annotation types call for some selection and simplification before importing it to the new application.

7.2.2 Preprocessing

Even though the dataset is well-structured and well-maintained, it requires some preprocessing to suit the requirements of the application. This is done with a Python script which reads the input files, processes them and then writes them to the MongoDB. Defining these steps in a script additionally ensures that the process is testable and repeatable.

The preprocessing as such consists of three major operations. First of all, it reduces the structure of the dataset so that no more joins between sentence elements and annotations are necessary. Even though the one-to-many relationship between *CorpusItems* and *PredictorAnnotations* makes perfect sense in Glanos' use cases, it is superfluous when there is only one corpus and a selected set of annotations. Joins, on the other hand, are expensive database operations which are to be avoided if they are not necessary. Additionally, the information from the dataset would not be altered by the participants of the study, which means it can be built into a static structure that is only read hereafter. Thus, adding the annotation data to the corpus elements reduces the entire dataset to a fixed single relation.

The second preprocessing operation aims at detecting duplicates in the annotation data and filtering empty or unusable annotations. Due to continuous improvements in the predictors, some of the entered annotations are still in the dataset even though they are deprecated. These have to be removed. Manually added annotations are also sometimes clashing with the ones the predictors put out. To avoid confusion and to reduce the overlaps between several annotations, the manual data was filtered out.

The third step of the preprocessing has to be done manually. After compiling the data for the first time, it is necessary to reduce the set of imported predictor data to a sensible size so that the participants would later not be overwhelmed by the variety of annotations. Keeping in mind that the users of the application would not be experts in linguistics, the chosen annotations have to be easily comprehensible or at least self-explanatory. This disqualified a predictor which annotates profession skills for example, because the definitions for these skills are very specific and cannot be understood intuitively. In addition to that, predictor data with a high percentage of incorrect matches is ruled out as it might compromise the participant's judgment whether an annotation is correct. The *address predictor*, for instance, marks several general location entities as addresses but rarely finds a correct address match. Therefore, users might assume that the incorrect matches are right because they are the most common ones. To avoid such confusion, the *address predictor's* annotations are excluded.

After running through all three steps, the resulting data is a compact structure with no empty or duplicate elements. It contains only a chosen set of annotations whose semantics is easily understandable. Still, it is worth to take a closer look at the outcome to examine its details.

7.2.3 Resulting data

The final dataset contains all of the 10,205 corpus items from the original dataset. 150 of these are reserved for the training corpus in the tutorial phase, the rest is used in the main phase of the study. In total, they contain 9.351 annotations of eight different types. Figure 6.1. shows the distribution of the annotations per type throughout the corpus.

The most obvious feature of this distribution is that the types *location* and *personname* make up more than 85% of all annotations. One possible reason for this imbalance lies in its origin. There is hardly a page in Wikipedia which does not mention names, countries

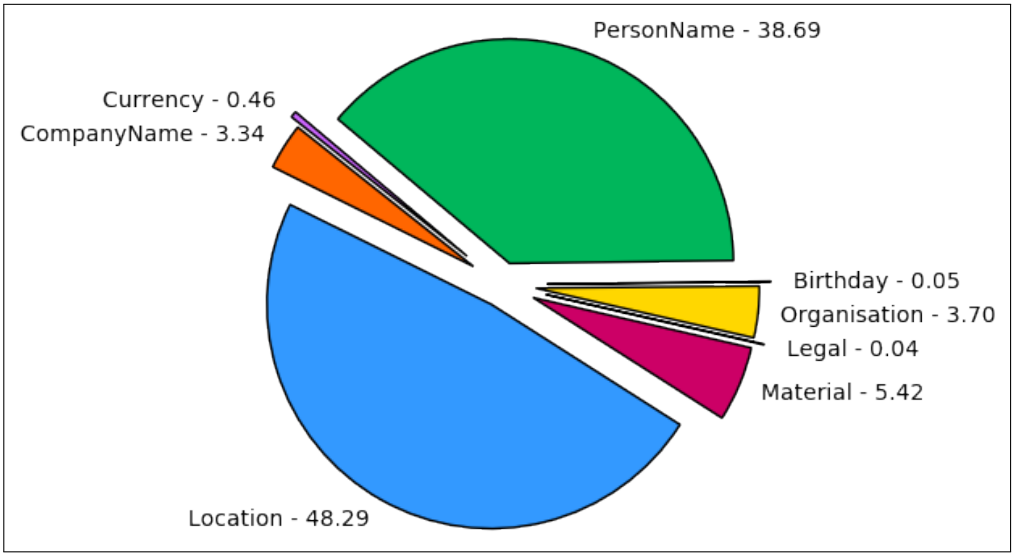

Figure 7.1: Annotation distribution in the dataset used in the study.

and cities, as they usually contain a section on the history of subject presented in the article. Organisations and currencies, however, occur far less frequently. Another explanation is that the predictors for *personnames* and *locations* have been developed very intensively and are hence quite sophisticated. Therefore, they detect greater quantities of annotations because their underlying grammars contain more rules. This also explains why there are so little annotations of the type birthday. Assuming that at at least the articles on prominent persons contain their date of birth, 0.05% seems a very small percentage. When taking into account that the predictor might not be developed to its full extent, this quota actually seems to be reasonable.

It could be argued that the extreme differences in the distribution are not ideal for the study as it might limit the types of gamification to apply. For example gamification methods based on finding a variety of different types might get increasingly challenging because it will take more time to find annotations which occur rarely in the corpus. However, it is worth noting that this is a live dataset which exists in this form at Glanos. Thus, it supports the realistic simulation of the actual working environment and does not impair the validity of the study.

While specifying the basic architecture aimed at defining a scaffold for the application, examining the data was more about learning what the program and ultimately its users have to work with. However, the application is still missing a concept for how the participants are going to interact with the system. Therefore, the next section describes the program's user flow and the different stages the participants go through.

7.3 Stages and user experience design

From a participant's point of view, the application contains four phases. They are welcomed at the start page, followed by the tutorial, where they learn how to use the program. Hereafter, they have to apply their new skills to real data during the main phase. Finally, they fill out a questionnaire to record their experiences with the program. The specifics of each stage are the subject in the following sections.

7.3.1 Start page and participant registration

The first page the participant sees when they call the URL is the start page. It contains some basic information on the study, as for example its purpose, the steps that have to be completed and an estimated duration for the process. It also lists supported browsers and shows the data privacy statement.

All of this seems to be quite ordinary and hardly worth mentioning, but this phase also has another purpose: it registers the participants. As soon as the volunteers click on the button to enter the study, they are assigned a participant ID which identifies them throughout the process. It also requests the backend to insert a new participant data structure into the database. This data structure is used to record all interactions the user omits during the next stages.

Working with participant IDs instead of a signup process has three major advantages. First of all it omits the organizational overhead of creating an account and is thus much more comfortable for the participants. Signing up always requires some initial effort by the user which might even discourage them from participating. Meanwhile, this effort is practically non-existent when all that has to be done is click a button. Secondly, it is easy to implement, as a new ID is generated by the database when inserting the participant's basic data structure. The new ID is then simply returned to the frontend, which uses it as an identification token for all further interactions between the participant and the backend. In addition to that, binding the participant's data to an anonymous ID avoids all types of data privacy issues. As no personal data is stored during any of their interactions with the system, infringement of data privacy is not possible. Nevertheless, each user's interactions can still be identified by the ID so that coherent datasets are created. Keeping all that in mind, assigning participant IDs is an adequate, secure and simple way to solve the issue of registering participants.

Apart from delivering a participant ID and inserting new base user data into the database, the backend also sends a token which assigns the user to either the gamified participant group or the non-gamified participant group. This token is appended as a prefix to the participant ID so that it is not visible as a separate parameter. This ensures that malicious users cannot assign themselves to another group by simply changing the token.

With finishing the registration, the participants are ready to interact with the application. However, they still do not know how to make a valid contribution. The next phase takes care of that.

7.3.2 Basic training with a multimedia tutorial

Before they start evaluating entities, the volunteers have to be trained how to use the program first. A multimedia tutorial walks them through the common use cases and shows them how to properly interact with the application. To make the training as effective as possible, it combines screencast videos and mandatory exercises. This technique has several advantages. One is that the videos are able to convey a great amount of information in a short amount of time, which decreases the overall time required for training. The viewers also receive the information on visual and auditive channels which is far more effective than just reading a set of instructions. Moreover, imitating a process that has been demonstrated before needs far less interpretation than inferring an action from written rules. Additionally, having mandatory exercises forces the participants to put the skill they have just heard about into practice. Thus, the participants acquire the necessary theoretical foundations and practical skills in a compact and easy way.

The tutorial is subdivided in four steps to keep the user's focus and to not overwhelm them. Step one provides a basic introduction and sets a narrative around the task. The video explains that the volunteers are part of a team which has to evaluate entities on a

daily basis. They are also shown a first example for annotated text which is available for them to interact with underneath the video player.

The second step intends to train the participants for making valuable evaluations during the main phase. Here again, users watch a screencast which shows them how to interact with the annotated text. They also learn the rules which define when to accept and when to reject an annotation. After watching the video, they have to evaluate five example annotations correctly before they are able to go to the next step.

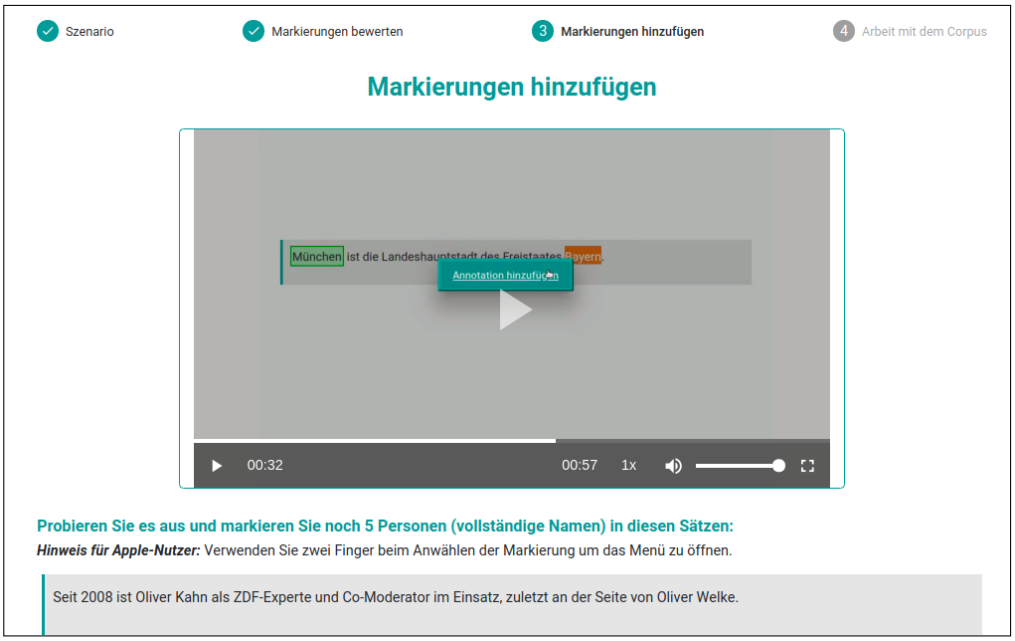

Figure 7.2: Screenshot from the third step of the tutorial. The gray box on the bottom is one sentence from the mandatory exercise.

In step three, the volunteers are taught how to insert manual annotations. The screencast shows the interaction and provides some examples to practice on. Their mandatory task is to annotate five *personname* entities in a given text. Each of their inputs is checked against a set of possible results so that feedback for wrong inputs can be displayed.

The final step exposes the participants to the entity editor for the first time. This editor displays the corpus text and highlights its annotations with colored markup. Its scrollpanel continuously loads new text parts as soon as the users scroll down. Here the users are prompted to evaluate or add fifteen annotations before they are able to go on to the next phase. However, there is a second purpose to this apart from giving the users more chance to exercise. The system records all interaction of the participants in a so-called *InteractionReport*. It also checks the evaluations against expected outcomes and calculates the percentage of correct evaluations. Only if the users evaluated more than 70% correctly, they are allowed to go on, otherwise they have to do another test round until they perform better. This threshold intends to discourage users from submitting wrong evaluations on purpose as they will be faced with another round of test inputs until they either contribute correct evaluations or give up. It also aims at intensifying the training for volunteers who did not quite understand the task. As soon as they reach the threshold they are free to continue to the next phase. Their *InteractionReport* is saved to the database because it will be needed for the motivation-enhancing mechanisms later on.

So far, the participants have been trained in a step-by-step tutorial to interact with the

program's editor properly. Therefore, they are now ready to apply their new skills to the actual data.

7.3.3 Working with the corpus in the main phase

Upon entering the main phase the users are informed that they will work with life data for the next twelve minutes. They are also told that it is possible to intermid the study or to skip the remaining time once they have passed the minimum duration of five minutes. Additionally, they are reminded that correct evaluations are more important than a large quantity.

The options to pause or to skip the main phase were added so that very unmotivated or bored volunteers would still have the opportunity to record their experiences while not having to wait until the working time has run out. By presenting ways to get to the finish line faster, the application tries to motivate users to finish the study even if they do not have much time or are too bored by the task.

Figure 7.3: Screenshot from the main phase (not gamified version).

Depending on which group they are assigned to, the participants now work with either a gamified or an ordinary version of the editor. Here again, their interactions are recorded in an *InteractionReport* object. The working time as such is measured by a timer which is not visible to the participants. Their only time-indicators are the start and end times displayed in the menu bar (see figure 6.3). The decision to not visualize the timer is based on the idea that no further stress momentum should be introduced to this phase. By displaying a countdown clock for example, users might feel pressured to do the task as quickly as possible. As the overall goal is to have the annotations evaluated correctly and not to gain as many evaluations as possible, this sort of pressure is counterproductive.

After running out, the timer triggers another pop-up which informs the participants that the official working time is now over but that they are free to evaluate more annotations if they like. Upon clicking a button, the report with their interactions is saved to the database and they are forwarded to the next phase, where they have the chance to give feedback on the application they just used.

7.3.4 Recording user experience with a questionnaire

In the final phase of the study, the participants' experiences during the main phase are to be recorded. Therefore, they answer a questionnaire, which gathers information in the form of multiple-choice questions, open text questions and questions with predefined answers to select.

Here again, the questionnaire is subdivided in parts to not overwhelm the participant. It is therefore displayed in a five-step stepper as figure 6.4 shows. Each of the steps focuses on a slightly different aspect of the user experience. It starts with a general evaluation of the application as such, its usability and the tasks design. Hereafter, the users are asked to assess their own performance on the main task. They also answer some questions concerning their motivation while working with the editor. The third step asks specific questions on the participants' experiences with the gamification and feedback methods applied to the editor. The questions are adjusted to the respective use case as the volunteers were obviously exposed to different versions of the editor depending on the group they were assigned to. This is also true for the next step which focuses on feedback for improvements of the editor. Finally, some general questions are to answer, like the the volunteers' current employment, their age group and how often they like to play games.

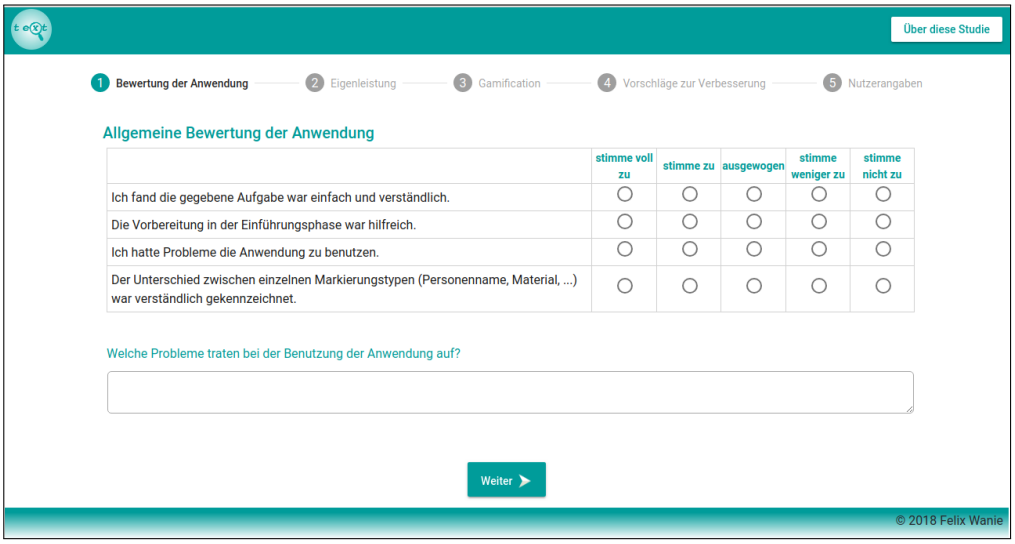

Figure 7.4: Screenshot from the first step of the questionnaire.

To ensure that no user inputs are lost, the application saves them whenever the participants moves from one step to another. Thus, their inputs can be restored in case the internet connection fails or their device shuts down due to empty battery. Here again, the goal is to enhance the volunteers motivation to finish the task by making it as save and comfortable as possible.

Once the participants have answered the last question, they are informed that they have finished the study. Additionally, the *state*-field in their data is set to *'FINISHED'*. Checks on this field throughout the application ensure that participants, who have already finished the study do not reenter it. This is mainly to keep the datasets coherent and to diminish side effects. One side effect might occur, for example, when users start another round of evaluating entities even though they already filled out the questionnaire. To keep the datasets consistent, their experience would have to be recorded again and it would have to be considered how to analyze this multi-facet data. Denying access to finished data is a

simple solution to that. Moreover, it secures the application against session hijacking[2] Theoretically, a hijacker could use a valid participant ID to enter the main phase and alter the participant's inputs, thus making the dataset invalid. This is avoided, however, when the participant ID cannot be reused after the study was finished. Thus, checking and updating the status is essential to the application's security.

The four steps described above fulfill all requirements of a user-centered application as it was proposed in chapter 5. The participants are registered, trained for the tasks at hand and then exposed to the real-life data. Finally, their experiences with the program are recorded. Security and data privacy are upheld throughout the application. With all this in mind, it is now possible to equip this application with gamification and feedback methods.

7.4 Implementing gamification and feedback

Section 5.2.2 already described the functionality and affordances of the motivation-enhancing methods to apply. Now that the application setup has been defined, it is possible to specify the concrete implementation of the game design elements.

All motivation-enhancing elements are displayed on a panel on the right side of the editor (see figure 7.5). Obviously, this panel is only shown for the group that is exposed to gamification and feedback. The elements on this panel receive their data from the interaction report that is written during the process.

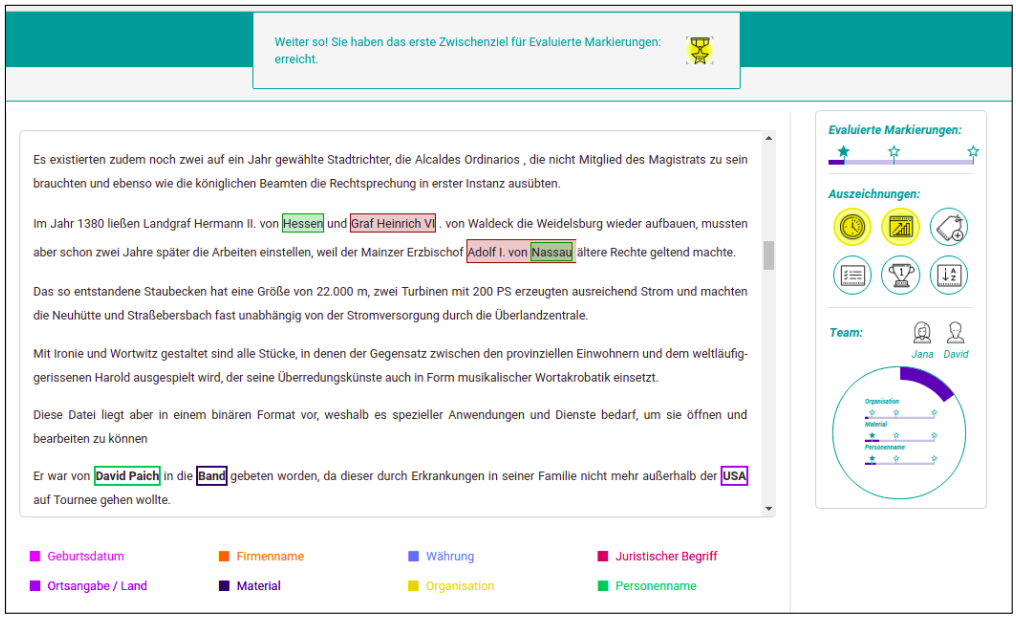

Figure 7.5: Screenshot from the main phase in the gamified application. The game design elements are on the right. A flash message announcing the achievement of an intermediate goal can be seen in the box at the top.

The progress bar showing the number of evaluated entities is the topmost element on the panel. Its target value is 150 elements, which is more than could be evaluated in 12 minutes. Hence, if someone wants to reach the 100% mark, they have to put in additional time. This time might then be an indicator that the motivation-enhancement is actually effective.

[2]Session hijacking denotes all kinds of methods that use sessions in websites to gain unauthorized access to information or services in a computer system (CITE!).

The progress bar's intermediate goals are set to 10% and 40% of the full target value. The two values were not chosen at random but are rooted in the idea that the challenge should increase with every intermediate goal to keep the task interesting. While the first subgoal provides an easy success which reinforces the initial motivation, the second goal already requires the user to put in some work. Each subgoal is visualized by little star that get filled once the objective has been achieved.

Hereafter, the panel displays six badges which are highlighted with a golden background once they have been achieved. With their varying levels of difficulty and their diverse goals, these badges have a great potential to motivate a range of different users. Their objectives are as follows (left to right, top to bottom):

- *More time spent in the editor than last time.* This badge uses the duration from the tutorial's interaction report for comparison with the current time. As users will undoubtedly spend more time in the main phase than they did in the last step of the tutorial, all users can be expected to achieve this goal. It is hence easy to get and more of a example badge to familiarize the study participants with the concept of badges.

- *More entities evaluated than last time.* Here again, the badge compares current values to the data from the tutorial. It is also easy to achieve this badge because it can be assumed that users will evaluate more entities than the fifteen from the tutorial.

- *Ten new annotations manually added.* This badge aims at motivating people to annotate entities that have not been found by the predictor. Its small absolute value makes it fairly easy to get but requires some searching within the corpus to find missing annotations.

- *One entity of each type evaluated in the corpus.* Here, the aim is to motivate the users to find as many different types as possible. Additionally, scrolling through the corpus with the intend to find a variety of entities makes them cover greater amounts of text. This in turn leads them to parts of the corpus which have not been processed so far. Thus, the badge is a good method to simultaneously increase the diversity of matches and expanding the range of text that is processed. Due to the fact that some entities occur very rarely in the test corpus, this badge requires real effort to achieve it.

- *All annotations of one entity type evaluated.* This achievement aims at making users go through the text thoroughly. The system knows how many entities have been found by the predictor and can thus check whether all annotations of one type have already been evaluated. While gaining this badge requires extensive and pedantic work, succeeding to win it will have a major increase in motivation.

- *Scrolled to the end of the corpus.* Here again, the goal is to make the user cover as much of the text as possible. However, it takes a very long time to gain this badge because the participants have to pass through 10,000 sentences before reaching the end.

Beneath the badges, the team challenge is displayed. It consists of three subgoals for the entity types *Organisation*, *Material*, and *Personname* which are visualized as independent progress bars. The target values were set relative to the number of annotations of the type in the corpus. The circular progress bar surrounding the partial elements shows the overall progress concerning the team challenge. To simulate a team interaction, two avatars were added. Hovering over the icons shows their contributions in a tooltip. Unfortunately it was not possible to simulate live contributions by these team members due to the fact that it is unclear which part of the corpus is processed by the participant at the time of such a simulated event. Overlaps with evaluations from the user might create unwanted interferences. On the other hand, if the simulated event occurs in a part of the corpus that is currently not visible to the user it will not have any impact at all because the user does

not observe it. Thus, the simulation of the team is reduced to the avatars and the repeated emphasis on teamwork during the tutorial.

Finally, the users' achievements are highlighted through flash messages as shown in figure 7.5. As proposed in the concept, the messages are formulated in an optimistic and praising tone. To ensure that the participants attribute the message to the correct goal, the task's icon is displayed next to it. This feedback mechanism hence additionally reinforces the motivational stimulus issued by the progress bar, the badges or the team challenge.

This chapter covered the architecture and user flow of the base application and described the data it uses. The program's MEAN-stack architecture addresses all requirements put forth in the last chapter. The original dataset is adjusted to the application's purposes. Additionally, the user experience design ensures that the participants know what is expected of them while the system still gathers all relevant data. Finally, the implementation of the proposed game design elements was presented.

CHAPTER 8

Analysis and evaluation of results

After reviewing the details of the implementation, it is time to have a look at the behavior that users displayed when using the application. Therefore, this chapter presents the results from the study and discusses them. First, it summarizes the steps taken to host the study and recruit participants. It also gives an overview on the autobiographic information that the volunteers provided. Hereafter, the resulting data is analyzed according to the hypotheses presented in section 2.1. Finally, the chapter summarizes the results and evaluates them critically.

8.1 Conducting the study

8.1.1 Hosting and participant recruitment

After finishing the implementation the application was hosted on a university webserver and made available through a public URL. To ensure constant availability, the application was integrated into a Linux service. Thus, it would be restarted every time the server was rebooted. Additionally, the application was run within a separate user that had only limited privileges so that it could not be used to access the server's root level. Unfortunately, this security measure entailed that the communication between the users and the application had to be rerouted within the server because only high-privilege users can access the standard ports. A NGINX instance configured as a reverse-proxy took care of that. Hence, this setup ensured that participants could continuously access the web application while it adhered to common security standards.

With the website being published it was necessary to recruit participants who would work with the program. Therefore, the link was posted in three WhatsApp groups whose members were interested in the project. Additionally, it was sent out to friends and family, and posted on Facebook. To create an initial incentive for participating, four coupons for a local ice-cream parlor were to be raffled off to users who finished the study. The awareness for the study was reinforced repeatedly by sending out the link multiple times and personally reminding people. Over the course of three weeks, this strategy attracted a sufficient amount of participants.

8.1.2 Participant sample

In total, twenty people successfully finished the study. As planned, they were split into two equal groups of ten participants each.

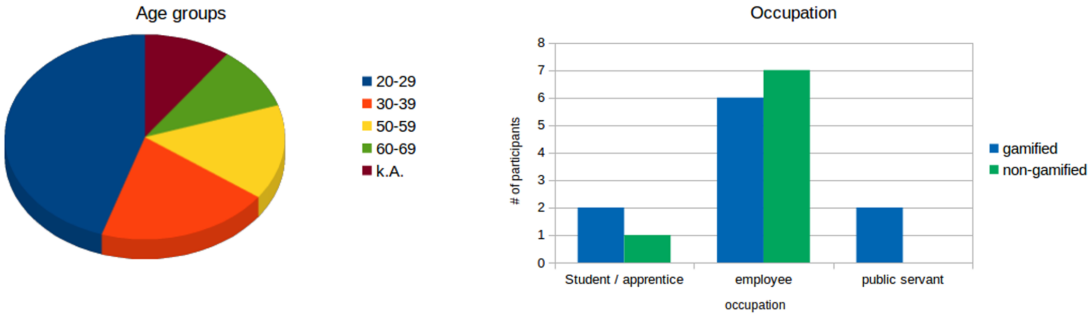

Figure 8.1: Age group distribution (left) and occupation group distribution of the study participants.

Almost half of the participants were between 20 and 29 years old, another 20% were in between 30 and 39. Unfortunately, two participants did not state their age group. Still, most users were under the age of 40 as the left graph in figure 8.1 shows.

Although some students and public servants took part, the majority of participants were employees. Six people stated that data maintenance tasks were a daily part of their job, and seven said that they handled such tasks multiple times per week. Nevertheless, over a third of participants answered that data maintenance was either not part of their job at all or concerned very small amount of their daily occupation. Participants of each answer group were almost equally distributed between the gamified and non-gamified group.

When asked about their attitude towards routine tasks in general, both participant groups displayed a balanced result with a mean value of 3.1 (standard deviation 1.07) in the gamified group and 3.2 (standard deviation 1.22) in the non-gamified group. This means that the majority of volunteers do not have strong positive or negative feelings towards routine tasks. Also, 50% of each group tend to complete routine tasks in one day while the other half likes to process them in parts over multiple days.

Surprisingly, less than 50% of all participants stated that they would play games at all. The numbers are similar for both board games and digital games such as video games or mobile games. Only seven people stated that they played digital games at least once a week. With overall consumption of digital games on the rise[1], the assumption was that least half of the participants played games regularly. Here again, the distribution of players and non-players was very similar for both the gamified and non-gamified group.

None of the participants had severe problems using the application. They stated that the task had been defined in a way that it could be easily understood. They also rated the tutorial as very helpful. While most questionees thought that the difference between the entity types was clearly highlighted, one person stated that he or she thought the definition for the *personname* entity was not precise enough because it left open whether the *first name* field could contain more than one name. However, this complaint only concerns a detail in one entity type and did not cause the user to leave the study. Thus, it is rather a indicator for future improvement than a reason to reconsider the participant's performance.

[1] According to *game*, the German branch association for games, the revenue of games worldwide has been constantly rising over the last five years. Source: `https://www.game.de/marktdaten/umsatzanteile-deutscher-spieleentwicklungen-am-deutschen-games-markt-2017/` (last accessed: 09/26/2018, 11:06)

In summary, the participants form an an appropriate sample for the study. Neither are there any serious imbalances concerning age, occupation, or fundamental attitudes, nor did the users report any technical problems that could impair the analysis of their performance. The only unexpected detail is the small portion of regular players. It seems logical to assume that individuals who play games on a regular basis are more prone to gamification due to their previous experience with game design elements. Thus, having a large amount of non-players in the gamified participant group might soften the effects of the motivation-enhancing methods. As the these methods are intended to unfold their potential on all users, the percentage of players is not a disqualifying factor.

8.2 User study results

In the following, the data is analyzed in order to find proof for the three hypotheses. These claim that there are differences in motivation, number of evaluated entities and quality of the evaluations for participants that were subject to motivation-enhanving methods compared to the group that was not exposed to these techniques. The analysis will draw from both the data recorded during the evaluation phase and the answers to the questionnaire.

8.2.1 Motivation

The first hypothesis claims that users who are exposed to motivation-enhancing techniques while evaluating entities state to be more motivated compared to the users in the control group. In order to prove this claim, it is necessary to examine the participants' self-assessment of their performance and their answers to the question items on motivation.

Both participant groups had a positive attitude towards their own performance. The accumulated mean within the gamified group was 3.62 with a deviation of 0.73. The values for the reference group are similar with a mean-value of 3.44 and a deviation of 0.62. While both values have a tendency to be neutral to positive the difference can be explained by having a closer look at single items. One question the two groups answered differently was on how much fun they had during the task. While the gamified group has a mean of 3.9 and a deviation of 0.94, the non-gamified groups only has a mean of 3.4 and a deviation of 1.2. This suggests that participants exposed to gamification tended to state that they have had more fun during the evaluation. A similar result was recorded for maintaining concentration during the main phase. Here, the group with gamification reports a slightly positive mean value of 3.1. The slightly negative mean of 2.6 in the reference group indicates that on average this groups' concentration tended to decrease during the evaluation. However, since running the t-test did not verify that the results differed significantly, these observations can only be counted as indication. Nevertheless, this positive attitude is a good premise for eliciting motivation. It shows that the users were not averse to the task which would have impaired their need for autonomy.

The items on motivation show a more unambiguous result. With an accumulated mean-value of 3.94 and a deviation of 0.56, the group which was subject to motivation-enhancing techniques clearly reported a better score than the non-gamified group, whose mean is 3.51 (deviation: 0.51). The t-test confirms a significant difference with the t-score of 1.784 being just above the critical t-score of 1.734. Additionally, Cronbach's alpha calculates to 0.72 which indicates that the items are internally consistent. Thus, the motivation of the participants in the group with gamification is indeed higher than the motivation of the participants in the comparison group.

Interestingly, the distributions on a few particular question items was not as distinct as these numbers might suggest. On average, the data shows that participants in the gamified group have a better sense of achievement, a better overview on their accomplishments, and

they are more satisfied with their performance. However, the non-gamified participants show similar scores on the item asking whether they felt like they had made a valuable contribution to the improvement of the text analysis software. The same mean value of 3.4 and a difference in deviation of only 0.29 imply that both groups show a similar attitude towards their contribution. Moreover, both groups indicate that they had a feeling of supporting their colleagues. Although the group with motivational support has a higher mean compared to the non-gamified participants (3.8 vs 3.6, similar deviation), the difference is small considering the fact that some game design elements were introduced to specifically increase motivation based on social relatedness.

These observations raise the question if the difference in motivation was caused by the motivation-enhancing techniques or whether it had another source. To answer it, the participants' evaluations of the methods are examined and compared if possible.

Due to the fact that all participants answered the question group on the motivational potential of the feedback emitted by the editor, this data is probably the most informative. With a Cronbach's alpha coefficient of 0.85, the items asked a concise set of questions which are highly related to the subject. Moreover, the results could not be more clear. With the t-score of 2,639 being well beyond the critical value of 1.734 and a p-value of 0.008, they are definitely statistically significant. Hence, the groups' mean values are so different that it is highly unlikely that the outcome was caused by an unknown third effect. Thus, the improvement of feedback by the game design elements successfully enhanced motivation in the group exposed to it. Comparing the items individually shows that the gamified participants answered almost every one of them with higher average score than the comparison group. The only exception is the question on whether they would have appreciated more feedback on their progress. Naturally, the participants who received limited responses from the system rated the item higher than their gamified colleagues. In addition, they explicitly asked for more feedback and a progress bar for the evaluation process. However, comments from the group with gamification also said that even more feedback should be introduced and that it should be more specific on the task. This indicates that the changes had a positive effect but there might still be room for improvement.

The gamified group also evaluated the motivational affordances of the badges and team challenge. Although the answers cannot be compared to data of the non-gamified group, it is worth examining them to learn about the users' perception of the implemented methods. The badges were generally seen as helpful which is indicated by a mean value of 4.1 within the corresponding question. With the exception of one individual, participants also had a good understanding of how they could achieve a badge. However, while 60% voiced the desire to win all badges, the mean value for actively pursuing this goal is 2.9. This suggests that less than half of them intentionally tried to fulfill the challenges presented by this game design element. One reason for that might be that the individual elements had not been explicitly introduced before to avoid overwhelming the users with information. Instead, the intention was to let them explore the badges on their own which might have needed more time than was available. One participant also criticized that the small number of entities of some types made it impossible to win certain badges and suggested that either more of these entities should be made available or a filtering mechanism should be added. Nevertheless, the overall mean value of 3.63 indicates that the badges have had a positive influence on motivation.

The team-challenge, on the other hand, received only an accumulated mean value of 3.13, which is very close to the neutral 3.0 mark. Three people even stated that the challenge did not interest them much. While the participants show a slight tendency towards the team succeeding in the challenge (mean of 3.3) and were ready to contribute a major part to it (mean of 3.6), the deviations of over 1.0 also show that the group had different opinions on this subject. 60% of the group stated that they actively tried to complete a partial goal of the team challenge, but the overall mean of 3.1 is again close to neutral. Interestingly,

three of the participants felt like they were in competition with their other team members. Even though the guidelines in section 5.2.1 stated that competitiveness should be avoided and the implementation tried to provide a collaborative challenge, this effect still showed. Thus it might have had a negative impact on the motivational affordance of this method in some cases. The most diverse opinions, however, are recorded on the question whether the team challenge was a helpful feedback for the overall progress. With two answers for every option, the mean value calculates to exactly 3.0 but the deviation amounts to 1.4. An explanation for all of these diverse results might be that the sense of team spirit could not be addressed in a way it would be when people actually worked in teams with personal contact. This is backed up by the paricipants' answers to the question whether they felt obligated to their team. Six individuals answered with *do not agree at all* or *do not agree*. Additionally, participants criticized that they did not understand how they could help their team and that they would have expected more explicit directions with distinct goals. Including active social contacts and providing better explanations therefore might improve the motivational potential of this game design element.

The group that had not received motivational support could not be asked about their perception of the gamification methods, of course. However, they rated a set of items concerning the fulfillment of the psychological needs for competence and relatedness. The participants strongly voiced their favor for visualizing the progress which is indicated by a mean value of 4.0 and a deviation of 0.63. They also reported a tendency to having specific goals (mean value of 3.3). However, the deviation of 1.1 suggests that the group members had different opinions on this issue. Nonetheless, the need for competence is clearly neglected by the non-gamified editor. The participants' statements on the need for social relatedness were not as distinct. With a mean value of only 2.7 and a deviation of 1.42, the responses display a mixed attitude. On reason for this could be that the improvement of feedback was the primary issue which made the introduction of teamwork seem less important. Overall, these findings indicate that the participants' need for competence was not addressed adequately by the program and that they regard improvements as necessary.

With the decisive results on feedback and the indications provided by analyzing the motivational affordances of the game design elements, it can be said conclusively that the difference in motivation is caused by the motivation-enhancing techniques applied to the editor. Thus, introducing gamification to the entity evaluation process indeed caused users to state higher scores on motivation compared to the scores of users who have not experienced gamification during the main phase. The first hypothesis can therefore be accepted.

8.2.2 Number of evaluations

The second hypothesis is based on the assumption that increased motivation during the evaluation phase has an impact on the productivity of the participants. Therefore, it postulates that users who are subject to gamification and feedback evaluate larger quantities of entities than users who lack these stimuli. The interaction data which was recorded during the main phase provides some interesting insights on that.

Figure 8.2 shows the number of interactions per participant (P0 to P9) for each group. The chart on the left depicts the values for the users who were subject to gamification while the diagram on the right illustrates the actions issued by the participants in the comparison group. It is noteworthy that the y-scale on the gamified chart ends with the value 700 due to one outlier in the gamified group while the other chart's maximum scale value is 120.

The participant causing this outlier later stated that he tried to win all badges and thus evaluated such an extensive number of entities. For this individual, the incentive strategy had worked better than expected. He evaluated 608 entities in total and spent over 43 minutes in the editor, which is more than thrice the original duration. While this instance supports the claim of the hypothesis, it also biases the results and distorts the comparisons

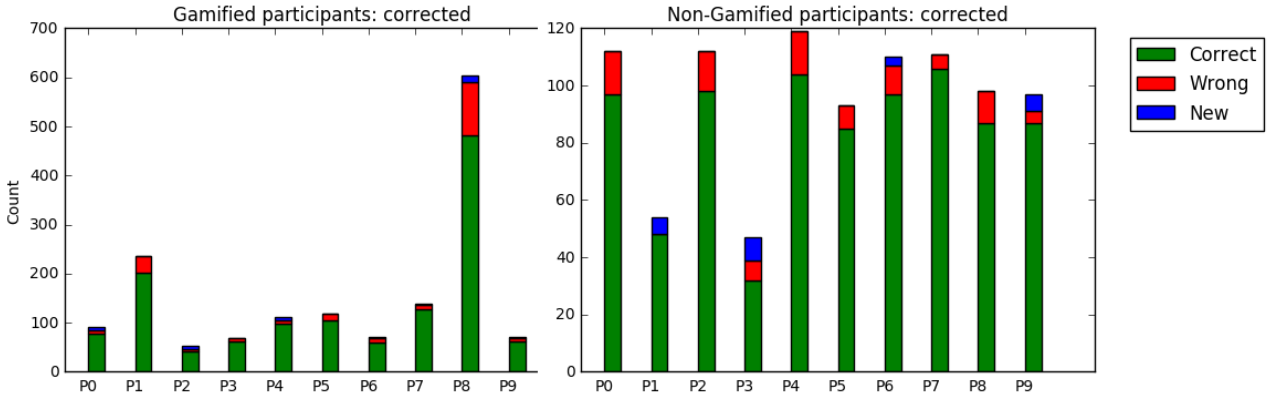

Figure 8.2: Number of interactions per participant within the gamified group (left) and the non-gamified group (right).

to a wide extend. As it is unclear whether this instance is a regular case or a single occurrence, the results are compared once with the outlier included and once without it.

On average, users assigned to the editor without gamification processed 93.9 entities and spent 12 minutes and 20 seconds in the editor. Apart from one user, all users finished right after the twelfe minute period had ended. They then read the final remarks and went on to the questionnaire. One person stayed for over thirteen minutes but he or she later stated in a comment that they had accidentally removed the dialog with the final remarks and had to figure out how to go on. This took some additional time, of course. All of this indicates that the users of this group had no intrinsic motivation to stay and continue processing annotations.

In comparison, the group who was subject to gamification and feedback stayed for more than 16 minutes on average and processed a mean of 157.0 entities. Even if the outlier is excluded they spent about one minute more than the participants of the non-gamified group and evaluated 102.75 entities. However, the data shows a greater deviation of 54.21 (without outlier) compared to the 23.47 of the control group. These values indicate that the number of processed entities varies greatly within the gamified group but significantly less within the non-gamified group. It leads to the conclusion that the participants who were subject to motivation-enhancing techniques tend to evaluate for a longer period of time and process more evaluations on average. However, the non-gamified evaluators are more constant in their performance.

Nevertheless, there are strong indications in favor of the hypothesis because the overall number of evaluations is indeed higher in the group with gamification. Calculating the t-test results in t-scores of 0.485 for excluding the outlier and 1.268 when keeping it in the data. Both values are significantly lower than the critical t-value of 1,734. Thus, the hypothesis cannot be completely accepted because it requires more data to prove its correctness. The current evidence suggests that it is correct though.

8.2.3 Quality of evaluations

Similar to the second hypothesis, the third hypothesis is also based on the expectation that an increase in motivation will have a positive impact on the quality of the resulting evaluations. Apart from examining the interactions recorded during the main phase, it is also insightful to take another look at the participants' self-assessment to contrast expectations and actual values.

Before the discussion of the results is continued, it is necessary to define what quality

means in this context. Glanos' annotation objects contain an evaluation value which was put there by one of the company's computer linguists. During the main phase, an evaluation was deemed correct if a user put in the same value as the Glanos employee did. A discussion on the advantages and drawbacks of this measure is included at the end of this section.

Apart from their positive attitude towards their own performance and a difference in concentration, the participants also believed in the correctness of their evaluations. Overall, 85% of all participants stated that they were certain to have evaluated the majority of entities correctly. Moreover, both groups showed slight confidence on the question item that asked if the participants were able to identify whether an entity was correct or not. However, both groups also had deviations of 1.20 and 1.24 which indicates that the opinions about this question were diverse within the groups. Nonetheless, there is a clear tendency among the participants to believe in the value and correctness of their work.

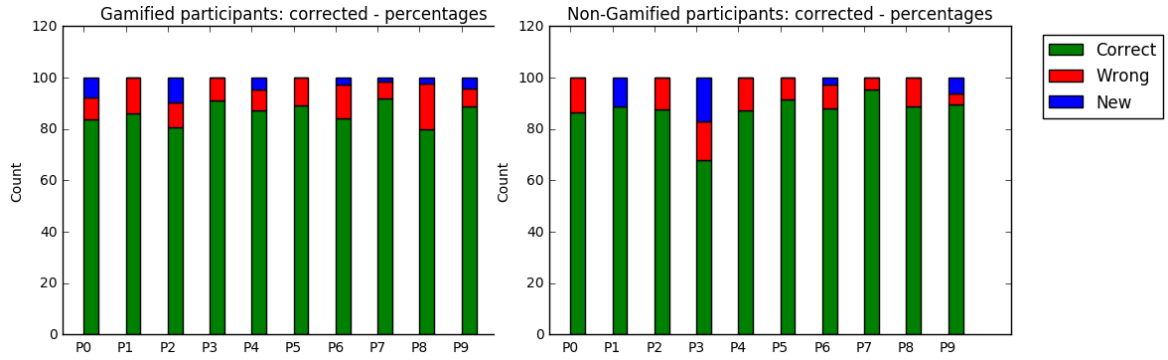

Figure 8.3: Percentage of correct, wrong and new interactions per participant within the gamified group (left) and the non-gamified group (right).

To find proof for the hypothesis, the interactions from the main phase are put in proportion to the total number of evaluations. Thus, it is also possible to see whether the correctness predicted by the participants can be found in the data. Additionally, a comparison of the overall quality is facilitated. The charts in figure 8.3 visualize the percentage of correct, wrong and new annotations of each participant per group. Once again, the gamified group's chart is on the left and the comparison group's is on the right. As the legend indicates, green denotes the correct evaluations while the wrong ones are shown in red and the new ones are colored in blue.

The charts reveal that 19 out of 20 participants evaluated at least 79% of evaluations correctly. Both groups show a mean value of over 85% for correctness. The portion of wrong evaluations is only about 10%. These numbers indicate that the participants indeed worked quite accurately.

Unfortunately, the groups do not show differences in evaluation quality. All values are in very close distance to each other. While the average percentage of correct evaluations within the gamified group is 86.31%, for example, the non-gamified group shows a mean value of 87.20%. Even more peculiar are the percentages of manually added annotations, which are 3.24% within the gamified group and 3.70% within the reference group. The values are almost equal even though the group with motivation enhancement had a badge that could be gained when adding ten manual annotations. It is noteworthy, though, that only four participants of the group without gamification contributed new annotations while seven participants from the gamified group manually added annotations. This data suggests that the game design element activated more users to contribute new annotations but this suspicion cannot be conclusively proven under these circumstances.

There are several explanations for these results. First of all, the sample consisted of volunteers with limited linguistic knowledge who all had the same introduction to the task. As the game design elements were intended to support motivation and not to make up for missing knowledge, they had very limited influence on the users' evaluation quality.

Another issue is the measure of quality. Measuring the correctness by comparing the users' inputs to Glanos' evaluation data has the advantage that the new data is contrasted to a fixed standard. However, these given evaluations were also entered by people who might have made mistakes. Thus, Glanos' evaluations are only little more reliable than the user inputs. An attempt to solve this dilemma is to compare the users' inputs to each other. However, this approach also has some drawbacks. Of the 633 annotations that were evaluated in total, only 287 were evaluated by multiple users. Thus, only 45% of all evaluated annotations can be used for this approach. This distorts the results because all other evaluations have to be accepted as either correct or wrong. Additionally, participants disagree on 61 of those 287 annotations. The most common reason for different evaluations is that some entity types are missing detailed definitions. For example, users disagreed whether *film* can be counted as material in a certain context because they were uncertain whether the term was too specific for a type that usually denotes iron, wood, or fabric. Therefore, these annotations could either be excluded, which distorts the results even further, or a decision has to be made whether they are correct or not. However, contrasting the data from the participants to a fixed decision is exactly what the approach was supposed to avoid. Hence, this method cannot provide any further evidence.

As no indication for increased evaluation quality could be found, the third hypothesis cannot be accepted. This does not mean that it is wrong in general but it cannot be proven under these circumstances with the given data. Therefore, it has to be rejected.

8.3 Summary and critical evaluation of results

After describing the participant sample and ensuring that it did not contain any features that would impair the analysis' results, the study data was analyzed to prove the three hypotheses. These had postulated differences in motivation, quantity of processed entities and quality of the evaluations for users who were subject to motivation-enhancing methods compared to users who did not receive such stimuli during the entity evaluation process. The analysis found that users who are exposed to gamification indeed claim to be more motivated than their peers in the comparison group. Thus, the first hypothesis could be accepted. It also found evidence that participants in the gamified group processed a greater number of entities. Even though these indicators have a strong support in the data, they are not significant enough to be statistically conclusive and require further research. Contrary to that, no evidence for differences in evaluation quality could be found in the data which means that the third hypothesis has to be rejected.

With all this information at hand, the overall research question can be addressed. It had asked whether introducing feedback and gamification to the entity evaluation process in the DataSphere would have a positive effect on its participants' motivation and its results. The answer is that applying gamification definitely has a motivating effect on the participants. According to the research, however, the positive impact on the result of the process is limited. There seems to be a tendency to process greater quantities of annotations but the quality of the evaluations remains the same.

Nevertheless, these results also have some limitations. It is difficult to generalize the effects found in the study due to the small size of the sample and the short exposure to the motivation-enhancing techniques. The impact of the game design elements might be different for another group of users. It also might change if a user is repeatedly subject to these methods over longer periods of time.

As already mentioned above, the results concerning the quality of the evaluations have to be regarded with caution. Due to the issue of finding an objective measure for evaluation quality, the effects caused by the motivation-enhancing methods might actually not be visible.

Finally, it is important to remember that the study was conducted with volunteers and not in a real business environment. By simulating the evaluation process with non-professionals, some adjustments had to be made such as choosing simple entity types. Therefore, it is possible that introducing the game design elements to the Gold Standard Editor elicits different reactions or changes the intensity of the effects. However, the study has shown that gamifying the editor inevitably triggers a positive change.

CHAPTER 9

Conclusion and future work

This thesis examined the effects of gamification and feedback on motivation and performance of users assigned to the entity evaluation process in Glanos' Gold Standard Annotator. Therefore, three hypotheses were formulated which claimed that users who are subject to motivation-enhancing methods (1) claim to have a higher motivation, (2) process more entities and (3) achieve better evaluation quality than users who do not experience these motivational stimuli.

As the gamification techniques were introduced to an existing system, the current state of the editor was analyzed for existing motivational affordances. The analysis revealed that it sends ambiguous messages and lacks any form of procedural feedback. No form of gamification could be found. These findings were confirmed by two of Glanos employees who use the Gold Standard Annotator often.

After mapping these findings to the corresponding psychological needs, a variety of motivation-enhancing methods was proposed. By applying progress bars with intermediate goals and a diverse set of badges the need for competence was addressed. Additionally, the need for social relatedness was supported through a team challenge.

To prove the effectiveness of these game design elements an interactive online study was conducted. It split its participants into two groups. One of these groups was exposed to the motivation-enhancing techniques while the other functioned as a comparison group. The study as such consisted of three parts. First, it trained the participants with a multimedia tutorial. Hereafter, they had to process entities in the evaluation process while their interaction data was recorded. Finally, they filled out a questionnaire which was designed to gather data on their inner feelings, thoughts and attitudes towards the process they had just worked with. The whole procedure ran in a specifically implemented JavaScript application, which was hosted as a website and made publicly available so that participants could easily access it.

After the study was completed, datasets from twenty participants had been collected. This data was analyzed with statistical methods such as arithmetical mean, deviation, p-value, and t-test. The analysis showed that the exposure to the proposed game design elements had indeed a positive effect on motivation. Gamified participants reacted well to feedback and had higher scores in the question items which focused on motivation. They also had a positive attitude towards the badges but were almost indifferent on the team challenge. These effects could probably be changed by including detailed introductions to the game design elements and their goals. The analysis also provided some strong in-

dications that users in the group with gamification processed larger quantities of entities. However, the differences to the comparison group were not large enough to get statistically significant results. Nevertheless, the indications in the data suggest that this hypothesis could be proven with more data and longer exposure to motivation enhancement. Unfortunately, no evidence was found for a positive difference in evaluation quality between the two groups. The reasons for that are probably the simulation of the process with volunteers and the method used to measure quality. Here again, longer exposure time and specific training might impact the results on this hypothesis.

Although the study succeeds in explaining the immediate effects of the chosen game design elements on the participants' motivation, it also entails some interesting questions for future research. One of these is how the effects of motivation-enhancing methods on the participants develop over time. As already mentioned in the critical evaluation, the exposure time to the methods was relatively short. The study's results also indicate that the participants sometimes did not explore the game design elements completely because of they were aware of the time limit. Additionally, users probably become used to the feedback mechanisms which might have a negative impact on the editor's motivational affordance. Therefore, composing a study on the long-term effects of the chosen solution might be insightful.

Another aspect that could benefit from some additional research is the differentiation between the methods. While the study proved that the chosen game design elements increased motivation, it is still unclear which of the elements had the greatest impact. The data from the questionnaire suggests that the feedback mechanisms played a major role, but all of the elements issued feedback in some form. Thus, the study's results refer only to the combination of motivational enhancement techniques. Finding out which of the elements contributed most to the motivational affordance of the system might help to prioritize the elements when implementing them into the real environment. It could also help to understand which psychological needs require increased attention in such editors as the Gold Standard Annotator.

While additionally automatizing data maintenance aims at decreasing the number of tasks in this field, introducing gamification targets the motivational affordances of the remaining tasks which are often perceived as tedious and repetitive by its assignees. Due to the fact that computer systems are developed to help humans with an aspect of their work or their lives, human interaction with the data will always be included in some form. Making this interaction as pleasant as possible thus complements the automation by ensuring that the user is motivated and focused on delivering high-quality results.

Appendix

Interviews

This section contains the two interviews from the case analysis. As both participants were German native speakers, the interviews and transcripts are in German as well.

Interview with Scrum-manager

1. Profil des Interviewteilnehmers

- Scrum-Manager bei Glanos, steuert Entwicklung des Systems
- Sehr gute Kenntnis des Systems, aber nicht aktiv an der Programmierung oder Grammatikentwicklung beteiligt
- Kontakt zu Kunden die den Gold Standard Annotator nutzen, damit auch mit der Außenperspektive vertraut

2. Nutzungsverhalten
Rahmenbedingungen:

i Seit wann nutzt du den Gold Standard Annotator (GSA) schon?

Seit er in aktiver Benutzung ist, ca. 18 Monate

ii Wie oft nutzt du den Gold Standard Annotator (GSA) pro Woche/Monat?

Unterschiedlich, je nach Projekt und Anforderungen. Geschätzt durchschnittlich 2h / Woche

iii Mit welchem Ziel nutzt du den GSA hauptsächlich? (Annotieren? Akzeptieren?)

Akzeptieren der existierenden Annotationen; außerdem finden von passenden Kontexten um nicht gefundene Annotationen zu finden und zu markieren

iv Wie lange arbeitest du mit dem GSA pro Session ungefähr?

Meist ca. 2h, manchmal auch länger am Stück wenn verlangt

Tätigkeitsbeschreibung:

i Erkläre bitte kurz den Ablauf der Tätigkeit. Welche Schritte führst du hintereinander aus?

- Einloggen in die DataSphere, in den GSA gehen, Corpus laden
- Vorgehen pro Annotationstyp (wird unterstützt indem sich das System immer den letzten Annotationstyp merkt und automatisch wieder auswählt):
  - Querbeet scrollen, existierende Annotation eines Typs suchen, verstehen was die Annotation genau ausmacht (Was gehört dazu? Was muss dazu markiert werden?)
  - Herausarbeiten eines Musters, typische Fehler festhalten
  - Filter mit Volltextsuche verwenden um weitere potentielle Kontexte zu analysieren und noch nicht gefundene Annotationen zu finden
  - Suche mit bestimmtem Kontext bringt meist weitere Kontexte hervor, dann rekursiv weitersuchen

ii Welche Auswirkung hat deine Arbeit genau? Wie stellst du die Auswirkungen fest?

Markup ändert sich (Anm. des Interviewers: Unterstreichungen werden durchgezogen);

iii Wie stellst du fest ob die Aufgabe abgeschlossen ist? Wann hörst du auf?

- allgemein ermüdender Prozess; höre auf, wenn mir der Kopf raucht
- großer Corpus: ab einer gewissen Zeit ergeben sich keine weiteren Kontexte mehr oder es wiederholt sich alles in gewisser Weise
- kleiner Corpus: von oben nach unten durch, genauer als bei großen

iv Bearbeiten deine Kollegen parallel die gleiche Aufgabe auf dem gleichen Corpus? Falls ja: Wie koordiniert ihr die Arbeit?

- Der Kollege der die Grammatiken erstellt sagt meist Bescheid dass er noch Annotationen von einem Typ braucht, eventuell auf einem speziellen Corpus;
- sonst kaum weitere Koordination, wäre aber gut

v Machst du diese Aufgabe gerne (im Vergleich zu deinen anderen Aufgaben)?

Wie gesagt, sehr ermüdend; macht jetzt nicht direkt Spaß, es gibt Spannenderes; ist aber notwendig und deshalb versucht man es so gut wie möglich zu machen

3. Bewertung und Verbesserung des Gold Standard Annotator

i Was gefällt dir aktuell gut am Design und der Benutzbarkeit des Gold Standard Annotator?

- auto-accept/ batch-accept hilft sehr
- wildcards in der Suche
- letzter gewählter Annotationstyp wird vorgemerkt
- Markup zeigt deutlich den Unterschied zwischen akzeptierten/zurückgewiesenen Annotationen

ii Nutzt du die Filter? Springst du bewusst an bestimmte Positionen im Corpus?

(vorher erwähnt: ja, aber primär die Volltextsuche, die anderen machen wenig Sinn in diesem Anwendungsfall)

iii Gibt es Funktionen die dir fehlen?

- aktuell gibt es keine Dokumentengrenzen oder es werden keine angezeigt, damit ein einheitlicher Fluss an Sätzen bei denen man versucht eine Struktur zu finden
- Verknüpfung zu Originaltexten wäre gut um besseres Verständnis für die gesuchten Annotationen zu haben
- irgendeine Rückmeldung bezüglich des Fortschritts, man weiß nicht was und wie viel man machen muss um den Prozentteil der animierten Kreisdiagramms zu beeinflussen

iv Nach persönlichem Empfinden: Gibt das System aktuell genügend Rückmeldung (bzgl. Fortschritt, neuen Annotation etc.)?

- Nein, man weiß eigentlich nur selber wie viel man getan hat
- Man ist immer im Zweifel: Hab ich was übersehen? Hab ich versehentlich was weggefiltert? Einer KI wäre da cool, die einem weitere Vorschläge macht basierend auf dem was man gerade annotiert hat, aber das ist wahrscheinlich zu aufwändig

v Sollte der GSA die Koordination zwischen Kollegen mit ähnlichen Aufgaben unterstützen? Würdest du diese Aufgaben lieber im Team erledigen?

- Definitv lieber im Team, a là "ich mache bis dahin, und ab da übernimmst du", dann hätte jeder einen Block zu bearbeiten
- sonst bietet der GSA keine Möglichkeiten zur Koordination, ist aus meiner Sicht auch schwierig einzubauen

vi Würdest du es begrüßen wenn Arbeit/Fortschritt anderer Mitglieder visualisiert würde?

- Ja, wird ja im Moment nur zum Teil abgebildet, mit den "Author"-Tag bei einer Annotation (Anm. d. Interviewers: wenn sie manuell erstellt wurde)
- Irgend ein Chart wäre cool für einen selbst

vii Würdest du insgesamt sagen, dass der GSA ein Editor ist, den Nutzer gerne benutzen? Warum ja / nein?

- Kunden sehen es als wichtig an, deswegen machen sie es
- man hat nicht den Eindruck dass sie es gerne machen
- Grund dafür ist wahrscheinlich dass man nicht weiß ob es richtig ist. Ist die Annotation genau so korrekt oder sollte der Kontext noch dazu? Passt das noch in das jeweilige Feld der Entität oder schon nicht mehr? Oft braucht man da richtig viel grammatisches Hintergrundwissen. Und da fehlen einfach oft Definitionen und Konventionen.
- Es klafft halt eine Lücke zwischen den Anwendern die Annotieren sollen und den Anforderungen der Grammatiker. Deshalb sind die Kunden immer froh wenn die Annotationen schon existieren und sie nur noch Akzeptieren / Rejecten müssen. Darüber lässt sich dann leichter diskutieren.

Interview with computer linguist

1. Profil des Interviewteilnehmers

- Computerlinguist bei Glanos, Teil des Softwareentwicklungsteams für die DataSphere
- Sehr gute Kenntnis des Systems, entwickelt aktiv Grammatiken mit Hilfe des Gold Standard Annotator
- Perspektive primär aus Entwicklungssicht, kein direkter Kundenkontakt

2. Nutzungsverhalten
Rahmenbedingungen:

i Seit wann nutzt du den Gold Standard Annotator (GSA) schon?

Seit er in aktiver Benutzung ist, ca. 18 Monate

ii Wie oft nutzt du den Gold Standard Annotator (GSA) pro Woche/Monat?

Projektabhängig, aber als Unterstützung eigentlich fast jeden Tag

iii Mit welchem Ziel nutzt du den GSA hauptsächlich? (Annotieren? Akzeptieren?)

Sowohl annotieren als auch akzeptieren; Ziel ist hauptsächlich Grammatiken für die Entitätenerkennung zu erstellen oder zu verbessern

iv Wie lange arbeitest du mit dem GSA pro Session ungefähr?

Variiert sehr stark, manchmal den ganzen Tag, manchmal nur ein paar Minuten.

Tätigkeitsbeschreibung:

i Erkläre bitte kurz den Ablauf der Tätigkeit. Welche Schritte führst du hintereinander aus?

- Ideal, wenn die Kunden schon Beispiele vorlegen oder bereits ein Corpus annotiert ist
- Nach dem Öffnen des GSA: schaue mir die existierenden Annotationen an, suche nach weiteren Kontexten
- Filter mit Volltextsuche hilft dabei sehr
- neue Grammatikregeln erstellen anhand der gefundenen Kontexte und Annotator wieder laufen lassen

ii Welche Auswirkung hat deine Arbeit genau? Wie stellst du die Auswirkungen fest?

Das Kreisdiagramm änders sich, wenn man die Filter anklickt sieht man wie viele schon gefunden wurden oder nicht (Anm. d. Interviewers: Es gibt einen Filter, bei dem man einstellen kann, dass z.B. nur alle Annotationen, die noch evaluiert werden müssen, angezeigt werden.)

iii Wie stellst du fest ob die Aufgabe abgeschlossen ist? Wann hörst du auf?

- schwierig, man soll ja meist einen gewissen Prozentsatz an gefundenen Fällen erfüllen. Das Kreisdiagramm hilft da ein bisschen
- Meist eher Gefühlssache, oder wenn einem die Ideen für Kontexte ausgehen in denen die Entität stehen könnte

iv Bearbeiten deine Kollegen parallel die gleiche Aufgabe auf dem gleichen Corpus? Falls ja: Wie koordiniert ihr die Arbeit?

- ja, wir teilen das meist nach Annotationstyp auf
- theoretisch ist es möglich, dass einer eine Annotation macht und ein anderer löscht sie wieder weil kein Timestamp sichtbar ist.

v Machst du diese Aufgabe gerne (im Vergleich zu deinen anderen Aufgaben)?

- Es gehört zu meinem Job.
- Vorher musste ich das immer mit einem Texteditor machen, das war wesentlich unüberschaubarer. Deswegen verwende ich den GSA gerne, macht so mehr Spaß, ist große visuelle Hilfe

3. Bewertung und Verbesserung des Gold Standard Annotator

i Was gefällt dir aktuell gut am Design und der Benutzbarkeit des Gold Standard Annotator?

- farbliches Highlighting für verschiedene Entitätstypen sehr hilfreich
- strukturierte Ansicht der zugehörigen Parameter und Werte, da hat man einen schnelleren Überblick

ii Nutzt du die Filter? Springst du bewusst an bestimmte Positionen im Corpus?

Ja, vor allem Volltextsuche und den Filter um bestätigte und zurückgewiesene Annotationen zu finden

iii Gibt es Funktionen die dir fehlen?

- Volltext könnte noch besser werden, oder man führt eine erweiterte Suche ein, bei der man auf mehreren äquivalenten Begriffen sucht
- es wäre gut wenn man in Echtzeit sehen könnte wenn Annotationen von anderen akzeptiert werden, dann macht man kein Arbeit doppelt

iv Nach persönlichem Empfinden: Gibt das System aktuell genügend Rückmeldung (bzgl. Fortschritt, neuen Annotation etc.)?

- Kreisdiagramm als Feedback schon hilfreich, aber man weiß nicht genau was man machen muss um die Prozentzahl zu erhöhen. Mit ein bisschen Erfahrung kriegt man es raus, aber man kann es dann auch gut manipulieren.
- Markup gut, aber man bräuchte noch mehr Informationen um zu sehen ob eine Entität von jemandem gerade erst gefunden wurde oder ob sie schon alt ist
- genauer Fortschritt im Corpus kann man nur über die Zahlen sehen

v Sollte der GSA die Koordination zwischen Kollegen mit ähnlichen Aufgaben unterstützen? Würdest du diese Aufgaben lieber im Team erledigen?

/

vi Würdest du es begrüßen wenn Arbeit/Fortschritt anderer Mitglieder visualisiert würde?

Wie gesagt, Live-Update wäre gut, z.B. wenn andere eine Annotation akzeptieren oder neu erstellen. ist aber wahrscheinlich sehr aufwändig

vii Würdest du insgesamt sagen, dass der GSA ein Editor ist, den Nutzer gerne benutzen? Warum ja / nein?

- ist schon sehr anstrengend, kann man vielleicht so 1-2 Stunden machen wenn man kein Entwickler ist
- Wichtig ist, dass der GSA so einfach ist, dass Leute ihn ohne große Schwierigkeit benutzen können. Das funktioniert gut soweit ich weiß, die Verwendung ist ziemlich eindeutig.
- Weiß nicht ob die Leute wirklich gerne daran arbeiten, ist natürlich auch nicht so spannend wenn man nur eine bestimmtes Resultat haben will anstatt das, was man selber gemacht hat zu verbessern.

Predictors list

This table contains all predictors that were in the dataset delivered by Glanos. The suffix "DE" indicates that this predictor is language-specific was specifically designed for the German language.

Predictor name	*Entity type*	*used in study*	*Comment*
BirthdayPredictorDE	birthday	yes	
CompanyNamePredictorDE	company	yes	
ContextLocationPredictorDE	location	yes	
CurrencyPredictor	currency	yes	
FullAddressPredictor	address	no	high percentage of wrong matches
HomepagePredictor	homepage	no	no matches in corpus
JobCompanyNamePredictorDE	company	no	very confusing for non-linguists
LegalPredictorDE	legal	yes	
LocationPredictorDE	location	yes	
MaterialPredictor	material	yes	
MeasurementPredictor	measurement	no	summarizes a diverse set of measurements, not intuitively understandable
OragnizationNamePredictor	organization	yes	
PersonNamePredictorDE	personname	yes	
PhonePredictorDE	phone	no	no valid data in the documents
QuantityPredictor	quantity	no	summarizes a diverse set of measurements, not intuitively understandable
SkillExtractorDE	skill	no	specific job skills, not intuitively understandable
StreetPredictorDE	street	no	no valid data in the documents

Questionnaire: questions and results

This section comprises the questionnaire data. Here again, the data is left in its original form, which means text answers are in German. The numbers in the fields under the Likert scale refers to the number of participants that have answered it with this value.

Questionnaire for gamified group

Allgemeine Bewertung der Anwendung:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*
Ich fand die gegebene Aufgabe war einfach und verständlich.	8	1	1	0	0
Die Vorbereitung in der Einführungsphase war hilfreich.	8	2	0	0	0
Ich hatte Probleme die Anwendung zu benutzen.	0	0	1	3	6
Der Unterschied zwischen einzelnen Markierungstypen (Personenname, Material, ...) war verständlich gekennzeichnet.	6	3	1	0	0

Welche Probleme traten bei der Benutzung der Anwendung auf?

- First name war nicht klar definiert, durften da auch mehrere stehen?

Einschätzung der Eigenleistung:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*
Die Aufgabe hat mir Spaß gemacht.	3	4	2	1	0
Meine Konzentration nahm während der Bearbeitung ab.	1	3	1	4	1
Ich bin mir sicher, dass ich die meisten Markierungen richtig zugeordnet habe.	1	7	2	0	0
Ich hatte Schwierigkeiten neue Markierungen zu finden oder hinzuzufügen.	0	2	2	3	3
Manche Markierungen konnte ich nicht eindeutig zuordnen.	1	1	2	4	2

Motivation:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*
Ich hatte während der Arbeit Erfolgserlebnisse.	3	6	1	0	0
Ich bin zufrieden mit meiner Leistung.	4	5	1	0	0
Ich habe eine Überblick darüber, was ich geleistet habe.	3	5	1	1	0
Ich hatte das Gefühl, die Aufgabe nach eigenem Ermessen angehen zu können.	3	3	4	0	0
Ich hatte das Gefühl, die Aufgabenstellung engt mich bei der Bearbeitung ein	0	1	1	5	3
Ich habe das Gefühl, meine Arbeit hat einen wertvollen Beitrag zur Verbesserung der Textanalyse geleistet.	2	4	0	4	0
Ich habe das Gefühl, ich habe meine Kollegen durch meine Arbeit unterstützt.	3	3	3	1	0

Bewertung des Feedbacks:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*	*keine Angabe*
Die Anwendung gab mir angemessene Rückmeldung über meinen Fortschritt.	4	5	0	0	0	1
Die Rückmeldung war für mich gut verständlich.	6	3	0	0	0	1
Es war klar, worauf die Fortschrittsrückmeldung zurückzuführen ist.	4	2	1	2	0	1
Ich fühlte mich durch die Rückmeldung angespornt.	3	6	0	0	0	1
Ich hätte mir mehr Rückmeldung über meinen Fortschritt gewünscht.	1	2	4	1	1	1
Ich empfand die Rückmeldung als störend.	0	0	0	5	4	1

Bewertung der Auszeichnungen:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*
Mir war klar, was ich tun muss, um eine Auszeichnung zu erreichen.	4	2	3	1	0
Ich habe gezielt versucht eine oder mehrere Auszeichnungen zu erreichen.	1	4	0	3	2
Ich hätte gerne alle Auszeichnungen erreicht.	4	2	1	2	1
Ich fand die Auszeichnungen nicht hilfreich.	0	1	0	6	3

Bewertung der Team-Aufgabe:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*
Die Team-Anzeige war eine hilfreiche Rückmeldung.	2	2	2	2	2
Ich habe gezielt versucht, ein Teilziel der Teamaufgabe zu schaffen.	1	5	0	2	2
Ich wollte einen möglichst großen Beitrag zur Teamaufgabe leisten.	2	4	2	2	0
Ich habe mich meinen Teammitgliedern gegenüber verpflichtet gefühlt.	0	2	2	5	1
Mir war wichtig, dass das Team die Aufgabe schafft.	1	4	2	3	0
Ich hatte das Gefühl mit anderen Teammitgliedern zu konkurrieren.	2	1	0	5	2
Die Teamaufgabe hat mich nicht interessiert.	0	3	0	3	4

Was hat Ihnen besonders gut gefallen?

- Das Tutorial in mehreren Einführungsvideos ist natürlich ein Knaller, zumal alles selbst gemacht wurde. ;-)
- die Oberfläche war sehr intuitiv bedienbar
- einfache Bedienbarkeit
- Die Rückmeldung war super, ansonsten wäre es mir zu langweilig geworden.
- Die Einführungsvideos mit den Beispielen.
- Interessante Textausschnitte, gute Belohnungsmethode

Was hat Ihnen am meisten gefehlt? Durch was fühlten Sie sich am meisten gestört?

- Die Mitteilung nach 12 Minuten Bearbeitungszeit hat mich überrascht und ich habe sie vor Schreck weggeklickt, ohne den Text zu lesen. Daher war ich wohl auch deutlich länger als 12 Minuten mit der Aufgabe beschäftigt. :D
- teils seltsame Sätze
- Hätte mir noch mehr Rückmeldung gewünscht und mir war nicht klar inwiefern das Team meine Arbeit beeinflussen sollte.
- Es war nicht immer klar, wofür es positive Rückmeldungen gab.

Haben Sie Verbesserungsvorschläge?

- Vielleicht kann diese Meldung nach der abgelaufenen Zeit genau so eingeblendet werden, wie die Zwischenstands-Mitteilungen. Dann erschrickt man nicht so und klickt wild in der Gegend herum.
- Ein klarer Hinweis: Notieren Sie 5 Orte in der nächsten Minute, um ihrem Team zu helfen.
- Es gab kaum Entitäten für Geburtsdatum und Juristischen Begriffe, es wäre besser wenn auch solche Entitäten gibt, damit mann nicht lange danach sucht

Welcher Altersgruppe gehören Sie an?

- 20 - 29: 4
- 30 - 39: 4
- 50 - 59: 1
- 60 - 69: 1

Zu welcher Berufsgruppe gehören Sie?

- Student/Azubi: 2
- Angestellter: 6
- Beamter: 2

Datenverwaltung und Spielaffinität:

Frage	*täglich*	*mehrmals pro Woche*	*mehrmals pro Monat*	*seltener oder nie*
Wie oft haben Sie Aufgaben im Bereich Datenverwaltung oder Datenpflege in Ihrem Beruf? (Kundendaten eintragen, Bestellungen aufgeben)	3	4	0	3
Wie oft spielen Sie Spiele? (Brettspiele, Gesellschaftsspiele, Videospiele, ...)	1	2	0	7
Wie oft spielen Sie digitale Spiele? (Smartphone, Computer, Konsole)	1	3	1	5

Umgang mit Routineaufgaben:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*
Routineaufgaben wie die Gegebene bearbeite ich gerne.	1	4	1	3	1
Ich bearbeite Routineaufgaben lieber an einem Tag und am Stück.	0	5	2	3	0
Ich bearbeite Routineaufgaben lieber über die Woche verteilt in Teilabschnitten.	1	3	3	2	1

Questionnaire for non-gamified group

Allgemeine Bewertung der Anwendung:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*
Ich fand die gegebene Aufgabe war einfach und verständlich.	8	1	1	0	0
Die Vorbereitung in der Einführungsphase war hilfreich.	9	1	0	0	0
Ich hatte Probleme die Anwendung zu benutzen.	1	1	0	3	5
Der Unterschied zwischen einzelnen Markierungstypen (Personenname, Material, ...) war verständlich gekennzeichnet.	7	2	1	0	0

Welche Probleme traten bei der Benutzung der Anwendung auf?

- Keine nennenswerten Probleme

Einschätzung der Eigenleistung:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*
Die Aufgabe hat mir Spaß gemacht.	2	3	3	1	1
Meine Konzentration nahm während der Bearbeitung ab.	1	5	2	1	1
Ich bin mir sicher, dass ich die meisten Markierungen richtig zugeordnet habe.	2	7	1	0	0
Ich hatte Schwierigkeiten neue Markierungen zu finden oder hinzuzufügen.	0	2	0	5	3
Manche Markierungen konnte ich nicht eindeutig zuordnen.	1	2	3	2	2

Motivation:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*
Ich hatte während der Arbeit Erfolgserlebnisse.	0	4	3	2	1
Ich bin zufrieden mit meiner Leistung.	1	4	4	1	0
Ich habe eine Überblick darüber, was ich geleistet habe.	0	5	1	3	1
Ich hatte das Gefühl, die Aufgabe nach eigenem Ermessen angehen zu können.	2	5	3	0	0
Ich hatte das Gefühl, die Aufgabenstellung engt mich bei der Bearbeitung ein	0	0	2	4	4
Ich habe das Gefühl, meine Arbeit hat einen wertvollen Beitrag zur Verbesserung der Textanalyse geleistet.	0	7	0	3	0
Ich habe das Gefühl, ich habe meine Kollegen durch meine Arbeit unterstützt.	1	6	1	2	0

Bewertung des Feedbacks:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*
Die Anwendung gab mir angemessene Rückmeldung über meinen Fortschritt.	1	4	1	2	2
Die Rückmeldung war für mich gut verständlich.	3	3	0	2	2
Es war klar, worauf die Fortschrittsrückmeldung zurückzuführen ist.	3	1	1	4	1
Ich fühlte mich durch die Rückmeldung angespornt.	1	5	0	3	1
Ich hätte mir mehr Rückmeldung über meinen Fortschritt gewünscht.	3	2	1	4	0
Ich empfand die Rückmeldung als störend.	0	0	1	4	5

Verbesserungspotential:

Frage	*stimme voll zu*	*stimme zu*	*ausge-wogen*	*stimme weniger zu*	*stimme nicht zu*
Ich schätze ich habe die meisten Markierungen bearbeitet.	2	2	5	1	0
Eine Visualisierung des eigenen Fortschritts war nicht notwendig.	0	0	2	6	2
Ich hätte gerne mit anderen zusammengearbeitet und mich abgestimmt.	1	3	1	2	3
Ich hätte mir konkrete Zielangaben für einzelne Markierungstypen gewünscht.	1	4	3	1	1
Ich hätte mich gerne auf einige wenige Markierungstypen spezialisiert.	0	4	1	2	3

Was hat Ihnen besonders gut gefallen?

- sehr gut funktionierende graphische Oberfläche, flüssiges Arbeiten

Haben Sie Verbesserungsvorschläge?

- Kommentarfunktion für abgelehnten Entitäten
- Reihenfolge der Felder in den Klassifizierungsmenüs eher wie man den Text liest (z.B. Vorname - Nachname statt Nachname - Vorname)
- Fortschrittsbalken einfügen

Was hat Ihnen am meisten gefehlt? Durch was fühlten Sie sich am meisten gestört?

- mehr Rückmeldung zum Arbeitsfortschritt

Welcher Altersgruppe gehören Sie an?

- 20 - 29: 5
- 50 - 59: 2
- 60 - 69: 1
- keine Angabe: 2

Zu welcher Berufsgruppe gehören Sie?

- Student / Azubi: 1
- Angestellter: 7
- keine Angabe: 2

Datenverwaltung und Spielaffinität:

Frage	*täglich*	*mehrmals pro Woche*	*wöchentlich*	*mehrmals pro Monat*	*seltener oder nie*
Wie oft haben Sie Aufgaben im Bereich Datenverwaltung oder Datenpflege in Ihrem Beruf? (Kundendaten eintragen, Bestellungen aufgeben)	3	3	0	0	4
Wie oft spielen Sie Spiele? (Brettspiele, Gesellschaftsspiele, Videospiele, ...)	0	1	1	2	6
Wie oft spielen Sie digitale Spiele? (Smartphone, Computer, Konsole)	0	1	2	1	6

Umgang mit Routineaufgaben:

Frage	*stimme voll zu*	*stimme zu*	*ausgewogen*	*stimme weniger zu*	*stimme nicht zu*
Routineaufgaben wie die Gegebene bearbeite ich gerne.	2	1	4	3	0
Ich bearbeite Routineaufgaben lieber an einem Tag und am Stück.	1	4	0	4	1
Ich bearbeite Routineaufgaben lieber über die Woche verteilt in Teilabschnitten.	2	3	0	4	1