# Designing a single player textual GWAP for validating tacit knowledge elicitation from crowds

**Madalin Broscareanu , Agathe Balayn , Ujwal Gadiraju , Jie Yang**
TU Delft

## Abstract

Machine learning can still make harmful mistakes. A solution would be tacit knowledge.

Machine learning needs this type of knowledge to improve. An example of such knowledge that can help make the system draw better logical conclusions would be: if presented with an open fridge, then it could deduct that the food will go bad. Tacit knowledge or common-sense knowledge (they are synonyms [1]) refers to the type of knowledge which is acquired through experience, the kind only humans can create.

GWAPs (game with a purpose) have shown quite promising results for acquiring such knowledge. Unfortunately, it could still contain errors due to users who only want to harm the game data, etc. and there is no method for validating such knowledge without involving humans somehow. Therefore, from the previously stated problem, our goal has emerged - develop a method for **validating an existing data set** and for later training machine learning models using a GWAP.

There has been work done before using GWAPs to elicit such information, yet they are limited in the sense that their main focus is set on data collection, not validation. Since very few projects looked into it, we decided to investigate a new GWAP which has as main purpose tacit knowledge validation.

The main question which we aim to answer is **"How can we elicit and validate tacit knowledge using a game with the following settings: single-player, textual concepts, goal: associate words with their concepts."**

The game presents hints to the users and they have to guess, as fast as possible and with the least amount of tries as possible, which answer is correct from the 6 options that are provided.

The evaluation of the game will be made using standard metrics such as games played, time spent playing, number of users, etc.

The conclusion is that the GWAP, even with the lack of data, was quite capable of analyzing the quality of the data set and reached a conclusion that is easily confirmed by a mere look over the initial data set.

## 1 Introduction

Machine learning's exponential increase in popularity has led to it being involved in many fields, but mistakes such systems make in unseen situations can still be extremely costly and dangerous. One way of improving its performance and avoiding such mistakes is tacit knowledge. Machine learning does perform well in situations such as video games, that require quick reactions and fixed actions in a well-defined setting, but when it comes to something more abstract that for example depends on multiple past events, then the performance drops[2]. We will specifically be looking into validating an existing data set of such knowledge, which has been previously collected, in order to evaluate its quality and relevance. Validation involves having a provided data set, which is then used to populate a GWAP that people play. With the data collected by the game, we then have to check the quality of the initial data using metrics, filtering, sensible observations and a thorough analysis.

### 1.1 Importance

Other works have looked into the same approach, but they have designed the games specifically for collecting tacit knowledge. Validating that collected data is extremely important because other humans could confirm that there is no incorrect/irrelevant data or spam in it.

This topic is of great importance in the field of machine learning because acquiring common-sense (tacit) knowledge from experts, crowdsourcing, or other similar methods, which involve some sort of compensation (usually monetary), can become extremely costly. On the other hand, collecting data from many people by providing them entertainment as a reward for their contribution to the research is a great alternative

[1] "Tacit knowledge is distinguished from explicit or formal knowledge and the term is sometimes used synonymously with common sense, in the sense of taken-for-granted knowledge" - Oxford Reference

[2] Multiple reasons why tacit knowledge is key, which are more detailed, for machine learning systems can be found here [7]

for eliciting such knowledge, which can be vital for the performance of machine learning systems [11]. "GWAPs can be an effective method to collect data. GWAPs are less expensive in the long term than other approaches for using human power to solve problems, such as Amazon Mechanical Turk." according to Liang Xu and Jon Chamberlain's (2020, p. 19) study.

Introducing the tasks that we want the players to perform in the game itself is of great importance as well. Simply turning the activity into a game by adding some user interface elements to it with no previous research into the design is not the goal. [10]

### 1.2 Relevant work

Different ways of gathering such data have been studied [8], including the GWAP (games with a purpose) approach, and have mentioned quite a few games that were built for such purpose. It is also stated that some other techniques are a viable solution, but the final comparison states that "crowdsourcing systems acquire knowledge from nonspecialists", which is much more convenient nowadays because we can easily reach many people online, who are willing to play the game. Moreover, researchers analyzed results collected from untrained people and they have shown a promising accuracy. [4]

People spend $\approx 9$ billion hours per year playing games online, therefore it can be great results could be derived from redirecting this effort towards common-sense (tacit) knowledge elicitation. A 3D treasure hunt GWAP called GIVE-2 has managed to collect data from over 1825 game sessions in three months. [1]

The game design can have a great impact on performance, data quality, and game experience. Choosing between a textbox where users can input anything and buttons is an extremely important decision that affects both the results and the players. The simple fact that users spend more time typing than clicking a button is crucial for the engagement and for the quality of the output because humans make spelling mistakes. [1] Moreover, this approach is prone to irrelevant data.

Researchers concluded that such games achieve the same quality output data at a much lower investment cost. [6] Therefore, we shall use them for our research as well - analyzing the collected data and rank the most relevant words associated with each concept and discriminating the acquired knowledge.

One of the questions that need more attention would be "To what extent can a GWAP in this setting be used to validate the quality of a data set of common sense knowledge?"

### 1.3 Research questions

1. How has tacit knowledge been collected before, using game and non-game techniques? We have to know what has been done before, how and why in order to provider some new, valuable information regarding this subject.
2. How can we design a game with a single-player setting to validate tacit knowledge? Having a good game design influences everything that follows from it. The collected data, the knowledge extracted from it, and finally the validation itself.
3. How to design the workflow to provide additional information on the relevance ranking of the knowledge for each class? We think that validating an existing data set could also have the potential of providing some extra tacit knowledge to the initial one.
4. How can we design the game experience to keep the players engaged and determined to continue playing the game? The game experience is vital for the quality of the output, since the players are the ones providing the knowledge and a poor design can lead to poor input from the users.
5. To what extent does the proposed design allow to validate the knowledge collected prior to the GWAP? We must take into consideration the fact that our game might not be able to provide an accurate verdict regarding the quality of the initial data set for various reason such as lack of players, lack of games played, not much overlap between answers, etc.

The main contribution of this research is providing a method for validating an existing data set of tacit knowledge. The main focus of many other related works is set on data collection and optimizing the game experience for the users such that they are kept entertained and rewarded for their efforts.

The game has shown, even with a limited number of players, who were mostly people that we know, that it is capable of providing reliable and valuable information regarding the quality of the initial data set. So it can indeed validate, to a certain degree of reliability and accuracy, the given knowledge, especially depending on the seriousness with which it was played and, most importantly, on the number of players and the games they played.

The following section will provide a detailed description of what we want to achieve (see section 2). Then we will take a look at related works that have inspired us to design the game and at our contribution towards the subject at hand (see section 3). The following sections of the paper will take a deeper dive into the subject at hand, with the focus centered on the methodology, where the game idea, workflow, and design are presented (see Section 4). This will be followed by a thorough comparison of different post-processing, filtering, and evaluation methods used to assess the quality of the data that has been previously collected to come up with a suitable one for validating the given data set (see section 5). Furthermore, the data that will be collected shall be objectively analyzed to create some new and more valuable knowledge from the initially provided set (see subsection 5.4) and draw some conclusions about it (see section 7). We will also talk about the ethical aspects of the paper and its reproducibility in section 6.

## 2 Problem description

For us to be able to accurately tell whether the initial data set is qualitative enough and to provide extra valuable tacit knowledge, the game should have a very solid and well thought out design, taking into account it is single-player and

text-based. A relevant example towards defining the "knowledge" we are looking for is: "If you leave the water running in the bathtub, it will eventually flood the house". However, we are especially interested in the kind of knowledge that associates words with concepts, in our case, with rooms, for example "whisk" $\rightarrow$ "kitchen". Validating such knowledge would imply having people play our game, populated with previously collected data that resembles the given example, and checking whether the results we get match the initial data set.

What is most challenging in creating a GWAP is finding a way to keep the players motivated to play the game, generating results that are as unbiased as possible, and collect data that will be useful in the future training of a machine learning system. [9]

After looking at other relevant work and comparing different game setups, we have come up with the main points that our design should focus on:

1. **Text only** - we have decided that our game will be based on text only, so the given hints, concepts, answers, everything will be displayed as text. We thought this would be best, since we don't have to rely on third-party image providers or store them in a database, which can become quite complicated, or store them locally, on the user's phone. Also, the text is much less open to interpretation than images Of course, there is also the downside of not being descriptive enough, such as images are. We have also taken into account that being open to interpretation might be a good asset to have, but we have eventually chosen to use text.
2. **Single player** - this decision was made mostly because it would be much easier to play alone, whenever the user wants, than having to deal with a lack of players. It might be hard enough to find people willing to play alone, therefore adding an extra requirement such as playing with friends would just add extra problems to think about. There is also the lack of competition that was taken into account, which brings out the desire to win, therefore encourages more cooperation from the players.
3. **Player engagement** - the interest of the players is one of the most important and vital aspects of the game design, since their desire to play the game directly influences the final output and its quality. If the game is not exciting enough, then it might encourage people to spam it with unreliable and unusable data. Since there could already be players who will do this, we do not need another motivating factor for them to neglect the game.
4. **Fast, spontaneous, easy to use** - if the game is kept simple, then it becomes easy to use, easy to collect data, and easy to process. Furthermore, by designing it in such a way it has fast gameplay, the human flaw of overthinking will not defile the data. Furthermore, it will also increase player engagement by forcing the player to constantly adapt and act fast.
5. **Quality checks and filtering** - since the crowdsourcing market has become infested with people who try to gain as much as possible for their benefit by cheating or working in a very relaxed manner that shows lack of interest and sloppiness [5], we must make sure that the collected data is not affected by such unethical behavior, therefore serious quality measures must be put in place. Moreover, thorough filtering shall be applied to further remove irrelevant data.
6. **Easy to filter and post-process** - we have to make sure the structure of the collected data is easy to post-process in order to apply solid filtering methods, therefore we took into account storing each answer separately. Each row consists of game session id, player id, shown hints, guesses the player has made until now, the timestamp when the answer was stored, and so on. This way we can easily see valuable and highly important information such as time between answers, time spent on a game, and so on.
7. **Cross-check** - one of the final steps of validation would be checking the initial popularity list against the final one. This is also something we have to take into account when designing the database where the data will be stored. We want to avoid any problems that might come up when post-processing the data and creating the ranked lists. If enough concepts in the output have in their top 3 most relevant words some or, even better, the same words as the initial list, then we could easily conclude that the given data set is of high quality.

Having all of these in mind, the complete design, detailed game design choices, and scoring method are in section 4

## 3 Related Work

The initial game idea was to collect tacit knowledge by showing people a concept for which they would have had to input as many words related to that concept as possible, as fast as possible. To avoid the "cold-start" problem that was introduced by this approach, we had to find a solution and started looking into related works. Unfortunately, the few ones that suited our description (text only, single-player) were using a "golden standard" or some kind of initial data set to work with.

Some of the related works that we have taken inspiration from while creating the game design are:

### 3.1 Phrase detectives

Players are given pieces of text that have been previously annotated by experts and now they have to do the same thing. In the end, their answers are scored based on how well they did. [2]

Unfortunately, this approach did not suit our intentions and design, because it involved a "golden standard". So they already had a data set that was previously annotated. Since we couldn't find that much work investigating the subject of validation, we thought having an initial data set that will further be inspected, with the help of information provided by other humans, would be a viable option.

### 3.2 1001 Paraphrases

Players are shown an expression at the top of the screen and they have to paraphrase it. If they guess one of the paraphrases already produced by another player, they get some points. If not, their guess is saved to the database and they can guess again for a lower amount of points. [3]

We liked this scoring system because users had to "guess" with the least number of mistakes. It motivates the player to answer correctly from the first tries to get more points. It encourages people to take the game more seriously and not spam it. This inspired us to use the same approach of having players guess in the least possible tries.

This approach requires, again, something to check the answers against, so we were suggested again to go towards validation and not data collection.

### 3.3 Sentiment Quiz

In this game, players have to vote, on a scale from 1 to 5, the intensity of the sentiment that they feel when shown a word selected from some of the 2008 US Presidential election documents. These votes are compared against each other with other people who have also voted and points are given depending on whether they agreed or not. [9]

Even though it didn't fit our criteria, we did find something quite smart and interesting about it. The use of buttons instead of allowing users to input whatever they desired. This fixed the issue of spelling mistakes and randomness. It is much faster and easy, and less error-prone. Better game experience, better data.

### 3.4 PhraTris

This game was developed by Giuseppe Attardi at the University of Pisa using a platform for creating GWAPs. It is based on the classic game of Tetris [3]. Instead of having to arrange colored blocks to win points, the users had to arrange parts of sentences, that were provided as separate blocks, such a way the sentence made sense. [1] It is also similar to the approach of the mobile app that helps users learn new languages called Duolingo [4].

This was a great idea, but we couldn't find a way to adapt our design to it. It had competition, the famous style of a world-renowned game such as Tetris, and provided the users with an urge to just make those words fit. It seemed quite addictive and it also highlighted the idea of using pre-made buttons with labels on them instead of giving users free will.

### 3.5 Shortcomings of existing games and designs

All the games that have been presented before had as their main focus data collection. Since most of them were using an already existing collection of data of some sort in one form or another, we decided to move our focus towards validating a provided data set of tacit knowledge. Moreover, most of the GWAPs either have a multiplayer setting or use images.

[3]Tetris (game) https://en.wikipedia.org/wiki/Tetris

[4]https://www.duolingo.com/

## 4 GWAP design

Our proposition for the GWAP is developing a Flutter Android game that presents words (hints) to the user and they have to guess the correct concept associated with these hints. **The game shall be populated with data from a given data set.** This data has been previously collected by someone else with a different game and needs to be validated using our game proposition. The game will only validate the "object belongs to room" kind of knowledge.

### 4.1 Game workflow

The players are given hints and a few buttons to pick from (concept options - e.g. "kitchen", "living room", "bedroom", "dining room" and "I have no clue") and have a limited amount of time (e.g. 90 seconds) to guess the concept that the words belong to (e.g. "table"). If they make a wrong guess (points are deducted) or wait more than 10 seconds they are shown a new hint (e.g. "chair"; no points are deducted). If they make another wrong guess, a new hint comes in (e.g. "fridge") and they would eventually guess the correct one, which is "kitchen". If they spend too much time or all the hints have been displayed and they still guess wrong, the game will end (see Figure 5). Hints are shown randomly for each game played to increase player engagement and improve the overall game experience.

The goal is to guess the correct concept as fast as possible in the least amount of tries. The score they can achieve for guessing a concept decreases with time and with every wrong guess, therefore they are motivated to think fast and answer correctly from the very first tries to achieve a high score and rank up in the league system presented below.

To keep the game even more interesting, there is an "I have no clue option" (see Figure 6) which introduces the strategy of minimizing the losses and forces the player to dynamically think about the next move: Wait 10 seconds for the next free hint? Try to guess faster and risk it? Quit early to avoid losing more points and being demoted to a lower rank? From the 560 recorded answers (only consider answers of games above 5 seconds to avoid spam), only 168 answers were given when there were 3 seconds or less until the next free hint would have been provided. Therefore, $\approx 70\%$ of users have opted for using the strategy to better wait than guess and risk losing points. This is an extra assurance that our game was taken seriously by most players and reinforces the idea that it is capable of making assumptions about the quality of the initial data set.

A way to ensure the player's interest in the game is not lost is the use of a ranking system, friends, and cosmetics combined (e.g. skins, badges, icons, etc. see Figure 7). For example, a frame around the user's profile picture or name is given as a reward for the "league" they are currently in (see Figure 5). The friend list contributes to this even more, because players will want to compete against each other and beat their high scores, as well as earning better-looking cosmetics (see Figure 7).

### 4.2 Game design choices

This particular topic of the project was subject to the most debates since it highly affects the game experience and there-

fore the quality.

We have chosen this game design because it keeps the game fast, entertaining and encourages players to constantly adapt and change their game strategy. We have set a time and guessing limit to the game such that players can't take advantage of it and think too much about the answer or, even worse, spam the game by pressing the same wrong option multiple times until the game ends, if it would ever end.

For ease of use, the buttons have been placed such a way that tapping roughly in the same area - which is also near the bottom of the screen since people might not have such big hands to reach the top screen corners of the very popular slim and tall smartphones with a single hand - will get the user through a long session of playing round after round after round, until they finally decide it's time to get back to the home screen.

After a wrong guess, the button is still clickable and doesn't become greyed out. This way, the game becomes a bit harder since the user has to remember past wrong guesses. Furthermore, it adds some quality to the final data since we can better differentiate between spam and valuable knowledge by looking at the sequence of the guesses and whether they constantly repeat or not.

Moreover, randomizing the hints is not only a way of making the game more interesting and unpredictable but also assures some quality of the data. If we can observe from the results that multiple people guessed the correct answer only after they have seen a specific hint, then that would prove the importance and relevance of the word for its concept. This is especially the case if the hints are shown in completely random order and they were also different. We tried to design the "hint path" (what hint is displayed when) as unpredictable as possible to make some more assumptions and perform extra data quality checks.

We have also taken into consideration offering a time bonus, so guessing faster would reward the user with more points. Unfortunately, this introduces the problem of randomly clicking on every button until you get the correct one. Therefore, the bonus has to be a moderate amount (not too big to avoid encouraging random guessing and not too small to still motivate the players to guess fast) and the penalties for guessing wrong should be extremely harsh. Another reason to have high penalties for wrong answers would be that it encourages people to think about minimizing losses and choose the "I have no clue" option.

In the long run, the strategy of "fast random guessing" is not a viable one, since the players will lose many points for the mistakes they make and it will take much more time to get promoted than it would take playing the game the way it should be played (see Section 4.4).

To further improve the data quality, every 20 rounds, there is a special bonus round (see Figure 9) where users can earn some free extra points. They have to play a mini-game which is similar to the game Taboo [5]. The users are provided all the other existing words in the data set, except for the ones associated with the concept shown at the top. They have to select a few of them that can be associated with. This way we can identify some already existing relationships between words and concepts which were not initially associated in the data set. If many users associate the same word(s), which didn't initially belong to the concept (e.g. "Conference room" and "people"), then we could assume that there is a connection between the two.

This round requires the players to select at least one option from the list, but no more than 3, to avoid the unpleasant situation when people might try to select as many options as possible to be awarded more points. The amount of points they receive is fixed, no matter how many options were selected, for the same reason.

For ease of use, the players have a search bar to easily look through all available options, or they could just scroll the list which is alphabetically ordered. Moreover, there is a "deselect all" button to not have to find again all the selected options and manually deselect them if the users change their minds. If they wish to not put the effort into finding an option to pick, they can just skip the bonus round by pressing the button at the bottom of the screen.

### 4.3 Scoring method

The goal of the scoring is ranking the players for motivation. First of all, certain thresholds need to be overcome to rank up in the ranking system previously described.

Being promoted from one league to another requires having a total score greater than:

- bronze → silver: 20.000 points
- silver → gold: 100.000 points
- gold → platinum: 200.000 points
- platinum → challenger: 500.000 points

Moving on to the actual game, the players are harshly penalized for wrong guesses. The amount is calculated as follows:

**-100 - (penaltyFactor * timestamp)**

where the ***penaltyFactor = 0.83***, and ***timestamp*** is the number of seconds elapsed since the start of the game. So as the game progresses, mistakes are more costly, therefore keeping it dynamic.

For guessing fast and correct, the user is rewarded with a time bonus, which is calculated as follows:

**(timerMaxSeconds - timeSpent) * timeBonusFactor**

where ***timerMaxSeconds = 90*** (since the game only lasts 1 minute and 30 seconds), ***timeBonusfactor = 3.2***, and ***timeSpent*** is the number in seconds that the user spent playing. The faster they guess, the more points they are awarded. If they don't answer correctly, they will not receive any bonus for the time, so only lose points for all the wrong tries.

The bonus round offers a fixed amount of 200 points if the users decide to play it. This way they are rewarded for their effort of scrolling through the list and searching for a good option to select.

We have chosen this scoring method and strategy in order to avoid all the spamming, irrelevance problems previously described, while maintaining an enjoyable and attractive game experience for the players.

[5] Taboo (game) - https://en.wikipedia.org/wiki/Taboo_(game)

### 4.4 Handling spammers

We have inspected 220 game sessions from a user who has played 278 games in about 20 minutes and analyzed the score evolution for only the games when the user managed to randomly guess the correct answer fast. This way, we could prove that guessing fast is not a good strategy. Although, this was proven given a condition.

We have only taken into account games that took under 4 seconds to play and only those when the final answer was correct. The player managed to guess correctly almost 80% of the games and achieve a total score of almost 60.000 points, which is quite impressive. As you can see in Figure 10, the trendline shows a high linear increase.

We have also looked into the score evolution of the same game sessions, but without adding the time bonus points. So guessing fast is not an advantage anymore and doesn't bring any benefits. The total score is -2533, so the user not only lost points overall but also wasted 20 minutes of precious game time that could have been used to honestly gain points.

If we take into account all the games of this player, then we reach a total of 57.732 points in 278 games, in approximately 20 minutes of game time. The player was punished with a loss of around -2268 for the times that the right answer was not randomly found. If the bonus time is removed and we also take into account all the games, then we reach a total of approximately $-2533 - 2268 = -4801$ points.

Unfortunately, these results show us that the punishment for guessing wrong might be too low and that the time bonus for guessing fast is too high. But the problem is not so easily solvable, because losing more points for wrong answers and awarding fewer points for guessing fast will not encourage players to think fast anymore and will also neglect the regular ones who try to honestly win points. Spammers could take advantage of the time bonus, but then normal players will have much more to lose and their game experience will be ruined.

A better and much easier solution to solve this would be to filter out the games that last under 3 - 4 seconds since those will most likely be spam, especially if more than one guess was made.

Therefore, the graph presented above shows that **guessing fast and random is not a viable strategy, if there is no time bonus applied**, but, unfortunately, we want to stimulate players to think fast and keep the game engaging and entertaining.

### 4.5 Processing game data into knowledge

The processing of the collected data focuses mainly on first filtering out the spam and irrelevant data, and then on determining the quality of the initial data set. Some filtering can be easily applied by using a combination of timestamps, the number of tries, and the total time spent playing. Thanks to the way the data is stored, we can see the progress that the players made since we store information with every answer they give.

To further filter out irrelevant data, we ignored the games where the player has spent very little time on that specific round and/or only answered "I have no clue". This would be a clear indication of the fact that there is no valuable data whatsoever provided by that round.

After generating the previous data, we have analyzed the answers that the users provided. As an intermediate result, we have selected only the games in which the users correctly guessed the concept and spent more than 5 seconds playing. We have also looked into the number of tries for each round, to further filter out the data (see Figure 1).

Figure 1: Relevant games

| | last_hint | tries | user_id | concept | game_session | time_spent |
|---|---|---|---|---|---|---|
| 1 | window | 3 | 62 | dorm room | 21 | 19 |
| 2 | people | 1 | 62 | classroom | 22 | 23 |
| 3 | curtains | 1 | 62 | bedroom | 23 | 7 |
| 4 | bin | 2 | 62 | dorm room | 25 | 6 |
| 5 | car | 1 | 62 | dorm room | 26 | 12 |
| 6 | chair | 4 | 62 | dining room | 27 | 22 |
| 7 | door | 1 | 62 | conference room | 29 | 17 |

### 4.6 Output analysis

Regarding the quality of the initial data set, it is determined as follows: the initial data set has rows of each object, the concept they were associated with, and their popularity. Popularity denotes the times players associated for example "bedroom" with "window" - see Figure 2. Then we rank the answers provided by the users for each concept based on the number of times they guessed correctly when seeing the same hint. For example, if many users guessed the concept "kitchen" only when "sink" was presented as a hint, then we could safely assume "sink" has a big relevance towards the concept "kitchen". Now if the initial data set has "sink" in the top words associated with "kitchen", then it has a certain degree of quality in it.

We have also taken into account the case when the entire path of hints contributed to the user guessing the correct concept. For example, maybe the player is shown "oven" as well before "sink". Should we also give credit to "oven" for contributing to the identification process? We think that if "oven" was truly relevant towards "kitchen" or more relevant than "sink", then the player would have guessed earlier. Furthermore, the same concept doesn't have the same hints presented in the same order, since it is randomized, therefore a high number of players that guess the same concept when shown the same hit shows a strong relevance relationship between the hint and the concept.

Looking at the last 3 shown hints was also a possibility. But to be certain, to some extent, that the last 3 hints are all truly relevant means having multiple users guessing the concept after seeing the same 3 concepts, maybe in the same order as well. This would only introduce problems given the proposed design because the order is random and the number of games played would have to be much, much greater to result in a reliable overlap.

We also gain quite some insight regarding the players' behavior thanks to the timestamps registered with each guess they make. We can easily filter out spam, as seen in section 4.4, check the time spent playing every round, guesses made, the time interval between guesses, number of guesses, etc.

The game elements introduced have shown quite promising results and were proven to be well thought out. The motivation of the spamming player presented earlier was to gain as

many points as possible, in a short time to rank up and reach the gold league. Luckily, he got discouraged when finding out about the high thresholds set in between leagues, therefore quit early and did not bother trying to get to 100.000 points anymore. The only game elements that are in dire need of improvement are presented in the time bonus vs penalties issue.

# 5 Evaluation

We evaluate how well the game allows us to validate knowledge by taking one data set with both correct and incorrect/irrelevant knowledge (e.g. "car" $\rightarrow$ "dorm room", or "wall" $\rightarrow$ "hospital") and giving the game to people who play it. We then collect the outputs of the game and evaluate them in comparison to the original data set.

## 5.1 Initial data set

The initial data set had a different structure and had to be processed to be usable in populating the game with it. It was split into multiple tables, with a lot of data that was irrelevant towards our game, because it has been previously collected using another completely different method. After removing the useless columns and combining the data, we have ended up with something similar to Figure 2. The actual data were correlations between scenes (images of rooms) and objects (people had to label what they saw in the given images) with both correct and incorrect relationships."Popularity" is calculated by counting the number of users who labeled the same object in the same category of images.

Figure 2: Words associated with the concepts, ranked on popularity

| | concept_name | obj | popularity ↓ 1 |
|---|---|---|---|
| 1 | bedroom | bed | 18 |
| 2 | bedroom | window | 18 |
| 3 | bedroom | pillow | 14 |
| 4 | dining room | chair | 9 |
| 5 | kitchen | cabinet | 9 |
| 6 | kitchen | drawer | 8 |
| 7 | kitchen | counter | 8 |
| 8 | bedroom | wall | 8 |
| 9 | bedroom | lamp | 7 |
| 10 | bedroom | chair | 6 |
| 11 | bedroom | table | 6 |

## 5.2 Game execution

To make a good estimation of the quality that the initial data set has, we need to make sure that our game is also entitled to critique the data. Therefore, we have looked at a few relevant metrics that could help us estimate the accuracy with which the game could estimate quality:

- users - we have gathered a total of 20 different account
- games played - we have collected data from 618 different game session
- average time spent per game - 5.37 seconds/round
- relevant games - after thorough filtering was applied, we took into consideration 130 games

We think that even though there is not much overlap between what the players have answered, as you can see in Figure 3 in the Results section, and taking into account that the initial data set had only 10 different concepts with an average of 23.1 hints/concept, the game can say, to a certain degree of accuracy, something about the quality of the given data set.

## 5.3 Results

Taking into account that there were only 10 concepts provided, we thought there would be enough overlapping data to get more users to guess the same concept when the same hint was shown. From the amount of games that we were safely able to consider relevant, as described previously, this is what the top most relevant words towards a concept looks like according to what our players have said (see Figure 3)

Figure 3: Most relevant words for each concept, ranked by popularity

| | concept | last_hint | users |
|---|---|---|---|
| 1 | classroom | people | 5 |
| 2 | kindergarden classroom | puppy | 5 |
| 3 | dining room | curtains | 3 |
| 4 | conference room | chair | 3 |
| 5 | kindergarden classroom | toy | 3 |
| 6 | dining room | table | 3 |
| 7 | conference room | plant | 3 |

We have gathered all the relevant answers and grouped them based on the last hint shown to the player and managed to obtain 33 words that had more than 2 users confirm their relevance towards the concept. Since those that have less than 2 are not very reliable, we thought about not taking them into account. But there is a total of 115 words that users confirmed are relevant towards some concept, therefore the lower-ranked ones, with a popularity of 1, might still be usable and relevant.

The highest popularity achieved was 5, but we only took into account games that took $\geq 6$, which could have been a quite harsh filtering criterion. If we lower the standards to time $\geq 4$, then we get a total of 137, not only 115, and maximum popularity of 9.

After checking one list (see Figure 2) against the other (see Figure 3), we have reached the conclusion that the initial data set was of slightly poor quality. As you can see, the most popular words are associated with bedroom, dining room, or kitchen (see Figure 2). These concepts are not even in the top results of the output, so their words were confirmed by less than 3 users (see Figure 3)

To further confirm these results, we have taken another look over the provided relationships between the words and the concepts. There have definitely been wrongly associations, for example: "object" $\rightarrow$ "living room", "car" $\rightarrow$ "dorm room", "room" $\rightarrow$ "bedroom".

Furthermore, relevant words such as: "kettle" $\rightarrow$ "kitchen", "bed frame" $\rightarrow$ "bedroom".

had very low popularity of only 1. Therefore, we can easily say, just by having a look at the data, that some of the relationships are absolutely outrageous and incorrect. Out of

around 230 relationships, we have identified, according to our knowledge regarding the concept of room, $\approx$ 66 words that had no relevance towards their associated concept, or were just wrong.

One of the threats to the validity of the evaluation would be human bias and error. Since not everything is black and white, there are some gray areas as well, such as "railing" $\rightarrow$ "bedroom", and we don't know if this is something common, relevant, or even plausible since it is not common for a bedroom to have stairs and therefore a railing attached to them.

### 5.4 Bonus round results

The bonus round only managed to gather 17 answers from 7 different users. Unfortunately, the additional information about other possible relationships between a word and a concept, which was not present in the initial data set, that we could gather **is not very extensive, therefore neither reliable**. The upside of this result is that the list is very short and can easily be manually interpreted by a human (see Figure 4). There is little to no overlap between the identified relationships and some can be questioned such as "advertisement" $\rightarrow$ "kindergarden classroom" or "shoes" $\rightarrow$ "living room".

Figure 4: Additional knowledge collected with the bonus round

| user_id | concept | answer |
|---|---|---|
| 109 | bedroom | [beds, cat] |
| 110 | childs room | [books, bookshelf, children] |
| 109 | childs room | [children, room, toy] |
| 62 | classroom | [bookshelf, desk] |
| 109 | dorm room | [bed frame, beds] |
| 56 | dorm room | [air conditioner, beds, books] |
| 62 | hospital room | [fire extinguisher] |
| 111 | kindergarden classroom | [adult, artwork, bags] |
| 111 | kindergarden classroom | [advertisement, air, air conditioner |
| 62 | kitchen | [fish, pumpkin] |
| 110 | kitchen | [tablecloth] |
| 109 | living room | [tv] |
| 72 | living room | [shoes] |

## 6 Responsible research

### 6.1 Reproducibility

One could conduct a similar experiment/validation using the code, initial data set and some other extracted and processed data made available on request on GitLab. It contains everything needed to reproduce the experiment, including some of the results that we have presented in the previous sections.

The research that we have done can be reproduced with a certain degree of accuracy, but since we have based our design on randomness, there will undoubtedly be a difference between our results and another similar research. We have to take into consideration the fact that each round was made to be different for every player, so there is an extremely small chance that the same player will get the same hints, in the same order for the same concept. There is also the issue of not having the same players and human uniqueness.

### 6.2 Ethical issues

The only ethical issue that is applicable in this case would be the security of the username/password combination, which has to be taken into account. The stored accounts are not publicly available, as well as the collected data. The only exception is some chunks of data that have been extracted from the answers that players provided to analyze the game, results, and draw conclusions based on them.

## 7 Conclusions and future work

### 7.1 Conclusions

We have reached the conclusion that our approach of validation using a GWAP has shown satisfactory results, the design of the game is quite problematic and still needs to be perfected, the extra knowledge provided seems correct, but need to be checked by an actual human since it is not very reliable with few responses collected, the game engagement is good and the players felt entertained and the proposed design is capable of validating knowledge (the more games played, the better).

### 7.2 Future work

Something that could be improved about our game design, as one of the users has suggested after deploying it, is that we should skip the entire "Game summary screen" (see Figure 8) where we offer an overview of the score. If we ask the players after every round whether they want to play again or not, then they might think that they don't want to anymore. On the other hand, if we skip this part and just keep them playing round after round, then they would have to stop the process themselves, which would involve more implications from their side. And since they are focused on playing the game, not on making such decisions, there might be a chance that the playing rate will increase.

Another way to go about this would be to introduce a timer somewhere on the results screen which announces the user that in e.g. 5 - 10 seconds that a new round will start, this way they will not have to press the "Play again" button and still can see an overview of the results. Also, there will be a way to cancel the timer and just wait a little longer to have a look at the results.

The scoring method needs to be perfected since it was designed to keep the spammers away, but people could still take advantage of it and the time bonus previously described in subsection 4.3. The outcome would not be harmed, since fast games below 2 - 3 seconds are instantly discarded when it comes to processing the collected data and turning it into knowledge. In combination with the ranking system which requires a lot of points to get some sort of benefit from spamming, it discourages people to play that strategy. But, finding a good, viable solution to prevent the creation of bots, without harming the overall, normal game experience is quite tricky. Punishing harsher for wrong answers would lead to people reaching scores below 0 extremely fast and it would make it almost impossible to get higher up in the ranks.

If the scoring system is kept the same, resetting the leagues for everyone after each week would solve the problem to

some extent. But it also introduces more frustration from players who actually struggle to play the game.

The threats to the validity of our experiment are the fact that we only tested one data set, with a specific type of knowledge and a small number of people that we know. Therefore, future work should also expand the evaluation aspect of this research.

Therefore, this project introduces a lot of room for new ideas and improvements that could perfect the game experience and the quality of the output.

## References


[1] John Chamberlain, Karën Fort, Ugo Kruschwitz, Mathieu Lafourcade, and Massimo Poesio. Using Games to Create Language Resources: Successes and Limitations of the Approach. In Gurevych, Iryna, Kim, and Jungi, editors, *Theory and Applications of Natural Language Processing*, page 42. Springer, January 2013.

[2] Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, pages 42–49, 01 2008.

[3] Timothy Chklovski. 1001 paraphrases: Incenting responsible contributions in collecting paraphrases from volunteers. In *AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors*, pages 16–20, 01 2005.

[4] Otto Chrons and Sami Sundell. Digitalkoot: Making old archives accessible using crowdsourcing. In *Human Computation*, pages 20–25, 2011.

[5] C Eickhoff, CG Harris, AP de Vries, and P Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In W Hersh, J Callan, Y Maarek, and M Sanderson, editors, *Proceedings of the 35th International ACM SIGIR Conference on Research  Development in Information Retrieval*, pages 871–880, United States, 2012. Association for Computing Machinery (ACM). SIGIR 2012, Portland, OR, USA ; Conference date: 12-08-2012 Through 16-08-2012.

[6] Carsten Eickhoff, Christopher G. Harris, Arjen P. De Vries, and Padmini Srinivasan. Quality through flow and immersion. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR 12*, 2012.

[7] Mit Ide. What machine learning can and cannot do, Aug 2018.

[8] Erik T. Mueller. Acknowledgments to the second edition. *Commonsense Reasoning*, pages 339–362, 2015.

[9] Arno Scharl, Marta Sabou, Stefan Gindl, Walter Rafelsberger, and Albert Weichselbraun. Leveraging the wisdom of the crowds for the acquisition of multilingual language resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 379–383, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[10] Luis von Ahn and Laura Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, August 2008.

[11] Liang Xu and Jon Chamberlain. Cipher: A prototype game-with-a-purpose for detecting errors in text. In *Workshop on Games and Natural Language Processing*, pages 17–25, Marseille, France, May 2020. European Language Resources Association.

# A Figures

Figure 5: Home

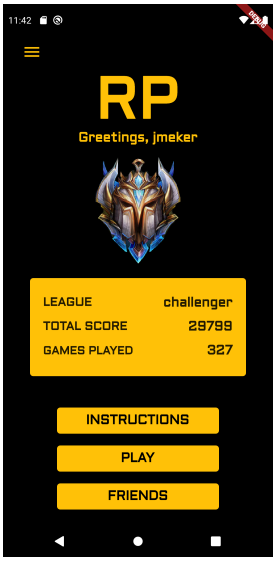



Figure 6: Game

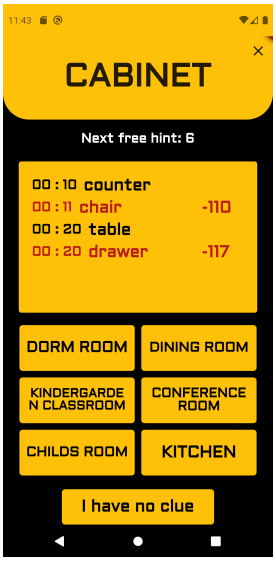
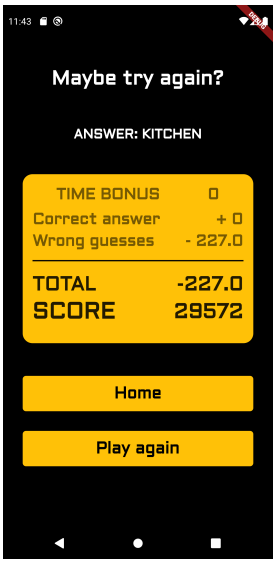



Figure 7: Friends


Figure 8: Game summary




NOTE: The scores, leagues and other player stats presented in Figures 5, and 7 are manually set in the data base and are random, just for the purpose of display.

Figure 9: Bonus round

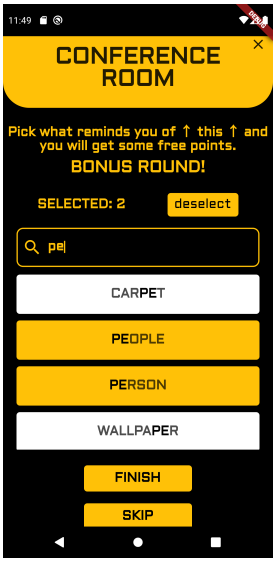



Figure 10: Spammer score evolution (with time bonus)

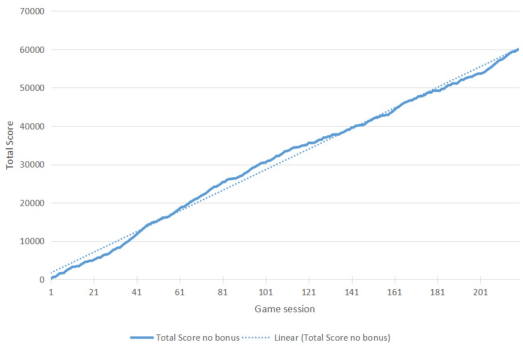



Figure 11: Spammer score evolution (no time bonus)