02455 - Experiment in Cognitive Science

Final project report

Text formatting and memory recall

Emma Pereira, Marah Marak, Apolline Dersy

s222668, s182946, s212836

4th December 2022

Contents

1 Introduction 1

1.1 Motivation and literature . . . 1

1.2 Our research question . . . 3

1.3 Hypotheses . . . 4

2 Methods planned 5

2.1 Experiment design and setup . . . 5

2.1.1 Creation of the word clouds . . . 8

2.1.2 Data treatment and protection . . . 8

2.1.3 Risks and ethical considerations . . . 9

2.1.4 Informed consent . . . 9

2.2 Data analysis . . . 10

2.2.1 Pre-processing of data . . . 10

2.2.2 Analyses and assumptions . . . 10

2.2.3 Post-hoc analysis . . . 11

3 Results 12

3.1 Tests and results interpretation . . . 13

3.2 Post-hoc results . . . 13

4 Discussion 15

4.1 Interpretation of results . . . 15

4.2 Conclusion . . . 15

4.3 Limitations and our contribution . . . 15

4.4 Possible future work . . . 16

5 Appendix 19

5.1 Methods . . . 19

5.1.1 Informed consent . . . 19

5.1.2 Information sheet . . . 20

5.1.3 Words for wordclouds . . . 21

5.2 Results . . . 24

5.3 All wordclouds . . . 25

1 Introduction

1.1 Motivation and literature

Studies in Cognitive Science have investigated how human brains prioritise and retain information. Specifically, numerous experiments have been carried out on the recall of words and texts. Memorizing some emphasized words right after reading them could be a help for many students and people. In this project, we want to work on this topic in order to find out what is the best font style to use when we have some emphasized words that we want the user to remember or have in mind.

Although not many studies have been found that specifically analyzed the relationship between word formatting and memory recall, we have come across several articles that talk about related issues. In a paper about the influence of colour in memory performance [1], the authors study and highlight ways to enhance memory recollection through the use of different colours, and they conclude on how there appears to be a common way of associating colour such that it has an effect on memory abilities. Colour seems to have a big potential to increase the amount of stimuli that is stored and encoded by subjects.

In another study, the effect of using bold text on search times for information in form fields has been analysed [2]. The authors conclude that writing field identifiers in bold and field values in non-bold, leads to the best performances in terms of information search time and accuracy. This condition was compared to the three other combinations of bold or non-bold field identifiers and field values. Seeing that the text formatting has an influence on search time, in that the field identifiers written in bold were found faster by participants, it seems that bold words particularly stand out in visual stimuli.

Schiano and Watkins carried out six experiments on Short-term recall of pictures and words, where the words to be remembered were the names of objects presented in pictures [3]. They investigated the effects of phonological similarity, word length and concurrent vocalisation on recall. The dependent variable was an estimate of the 'memory span' of participants, which they obtained using the 'up-and-down' method. Participants were presented lists of increasing lengths as long as they did not make mistakes, then decreasing, and increasing again if all their responses were correct. In this way, the authors obtained an estimate of each participant's memory span for a given task.

The authors found effects of the three independent variables on both ordered and free recall. Firstly, lists of phonologically dissimilar words were better recalled than lists of phonologically similar words. "That phonologically similar words in a short-term memory test are more difficult to recall than phonologically dissimilar words is a well-known phenomenon." according to Copeland and Radvansky [4]. Then, the memory span was larger for shorter picture names (words) than longer words. Finally, when subjects were instructed to count aloud from 1 to 3 during visual presentation of the stimuli, memory recall was significantly hindered.

Coltheart and Langdon [5] conducted experiments on recall of words, focusing on short words and fast visual presentation rates. The authors contextualise their study by listing

’frequently replicated findings’. For instance, in alignment with the aforementioned results of Schiano and Watkins [3], they state that a "frequently replicated result is the finding that lists of short words are better recalled than are lists of long words" and also explain that phonological similarity has a negative effect on recall accuracy. In their five experiments, Coltheart and Langdon presented words at rates of 114-243 milliseconds per item. Only one experiment was made with rates of up to 500ms. With fast rates, they observed ’attentional blink’ and ’repetition blindness’, which "are not found at the slower rates typical of short-term memory tasks in which items are shown at the rate of 1 per second.". Another topic addressed by the authors is the rehearsal process. In memory recall tasks, participants repeat the words in ’covert’ (or ’inner’) speech. When the rates are too fast, the authors argue that rehearsal is not possible for subjects, but that ’This difficulty rapidly diminishes over a period of about 0.5 sec’. Short visual presentation times make the task of recall more challenging.

In an article about mental storage capacity, Nelson Cowan brings together a variety of data regarding capacity limits that suggests that there is a smaller capacity limit of three to five chunks of items that is more real than the limit of seven chunks in Short Term Memory (STM) proposed by Miller (1956) [6]. One of the main points of this article is to justify how there is a *magical number* of 4 (plus or minus one) chunks of items that people can form when given an STM task of memorization. The article proposes at least four different ways in which capacity limits might be observed, with one of them being "when there is an information overload that limits chunks to individual stimulus items". This refers to the setup of overloading the processing system of a person when the stimuli are presented to them, so that they are exposed to more information than they can encode before the time limit is over. "This can be accomplished by presenting a large spatial array of stimuli (e.g., Sperling 1960) or by directing attention away from the stimuli at the time of their presentation (Cowan et al. 1999)". This first study (Sperling 1960) followed an experiment where, on each trial, subjects were visually presented an array of characters in a brief flash, which was then followed by a blank screen.

The article concludes how there is a big similarity in the capacity limit observed by using a wide range of procedures, such as the one explained about information overload. However, there is a restricted set of conditions that must be met in order to get these results. The results show how these procedures suggest a mean memory capacity in adults of three to five chunks of items, and the evidence for this limit is considerably more extensive than the higher limit of 7 proposed by Miller.

In their publication about recall of words by bilinguals, Ruth Nott and Lambert [7] report that "free-recall literature has documented fully the fact that Ss recall fewer items from a list of unrelated words than from one comprising items that can be easily grouped into semantic categories". When designing an experiment on memory recall, the words presented to participants must thus be carefully chosen if the effects of semantic similarity are to be limited. Stimuli with words of similar meaning might be easier to recall than other types of stimuli, which could be a confounding variable if not controlled for. Ruth and Nott further studied the difference in recall between ’category’ and ’non-category’ lists by French/English bilingual subjects. The subjects were split into three groups based on whether their strongest language was French or English or whether they were balanced. The results did not show a significant effect of whether the list was in the subject's weaker or stronger language on the number of words recalled. "For the non-category control lists, analysis of variance revealed no significant effects of either degree of bilingualism or language condition.". In the case of lists of words split into semantic categories however, bilingual and stronger language subjects performed significantly better than weaker-language subjects. Interestingly, subjects who were not balanced "recalled significantly more words in their weaker language" in bilingual lists. Although it could seem intuitive that participants find it easier to recall words from their stronger language, the results do not support this idea.

1.2 Our research question

As we have just seen in the previous section, a lot of research has been done in the way that the information presented to subjects can influence their memorization capabilities. Some of these factors explored have been the use of colour, the use of pictures and words of different lengths or the time the information is presented for. However, no specific studies have been found in the influence of word formatting on memory recall.

The research question that we are proposing is **"How does text formatting help with memorisation of important information?"**. With this question, we want to focus on how the way we present words to people influences their memory capabilities. More specifically, we would like to analyse the influence of having emphasised words, and whether this helps or does not help when it comes to remembering them. We are choosing to compare three ways of formatting words: by using bold characters, by using the underlined option or by simply presenting the word on a plain format.

Our research question can be tied back to the the article about mental storage capacity. Although the focus of our experiment is not so much on studying the amount of chunks of items people can remember, the experiment design described in that study can help us choose the right way to present the information to our participants. And, when interpreting the results we obtain, it is also worth taking into account how this possible *magical number* 4 of chunks remembered could influence the way subjects remember words.

The literature previously presented in 1.1 provided a variety of examples of studies, research questions, variables and conditions. These examples were used to refine our research question and choose the topics that would be addressed or not in our experiment. For instance, studies have been made on ordered (also known as serial) recall, which is not considered in our experiment. We also do not study the accuracy of recall, but simply count correct answers and discard the few incorrect entries. Experiments have been carried out both with repeated words and new words presented in each trial. We chose the latter for our study, as this prevents participants from familiarising themselves with the words, which might affect recall. Semantic categorising of words was found to have an effect on memory recall, improving the performance of native speakers. Choosing random words mitigates the risk of overlooking the presence of semantic categories as a confounding factor. The design of our experiment is detailed in section 2.

1.3 Hypotheses

We would like to research our question based on the hypothesis that bold words are easier to memorise compared to underlined or plain words. With this experiment we hope to prove or refute this and maybe understand how, and if, emphasised words are connected to memory retention.

2 Methods planned

In the following section, we explain in detail how the experiment was conducted and what we used to set it up, showing some real examples of what it looked like. Furthermore, we also describe what kind of design we followed and how we treated and prepared the data we used for it, and we share the ethical considerations that were taken into account before starting the experiment.

2.1 Experiment design and setup

The experiment was conducted at DTU by ourselves. Our aim was to recruit between 20 and 30 participants, who are students at the university, in the age range of 18 to 25 years old. In the end, we were able to get 21 participants, which we found through our social network and friends, and also by approaching people at the library who we did not know. In their article about short-term memory recall, Schiano and Watkins recruited 12 participants per experiment and obtained significant results [3]. Therefore, when we reached 21 participants, we believed it would be sufficient for our analysis. The age range was 21-29 (mean: 23.7, standard deviation: 2.2). 13 men and 8 women participated in the experiment. The experiment took place at the DTU library in a booth reserved beforehand, as well as in a classroom, to ensure that the participants could stay focused.

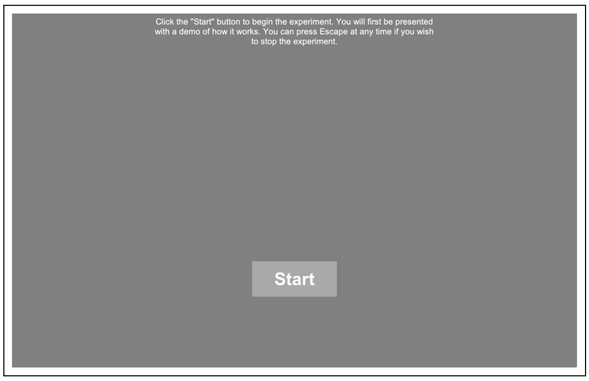
The experiment was prepared using Psychopy, in order to automatize it and make it exactly the same for all participants. Each participant had to finish 15 trials in our experiment, with each of them lasting 15 seconds. Schiano and Watkins in [3] presented 13 trials consisting of blocks in which their four experimental conditions were tested. In our experiment, all conditions are present in a single trial (wordcloud). Therefore, by running 15 trials, each of them presenting our three conditions, we have the same order of magnitude of number of trials. Every trial consisted of the visualization of a word cloud presented on the computer screen with a white background, containing 30 words in total, having them divided into three different formats: plain, bold, or underlined. The words on the screen were written in Arial font, size 16 and they were randomly assigned in terms of the formatting, having the same amount of words of each kind, so 10 of each. Having read about the limit of seven chunks in Short Term Memory, although disputed by Cowan [6], we decided to choose a number larger than seven. Knowing that participants would probably not be able to remember more than seven words, giving them ten words in each condition meant that there was a possibility for them to remember only words from a single category, for example only bold words. The 30 words were displayed in a word cloud format, so that there was no perceived order in which they should be read, contrary to a list or table. The wordclouds were shown for 15 seconds each, so that participants had 0.5s to read each word, which is enough time to read all of them if they wanted. The rate of 1 per second is frequently used in literature about memory recall, as stated by Coltheart and Langdom, "rates typical of short-term memory tasks in which items are shown at the rate of 1 per second" [5]. Our rate is faster than this typical value, while still avoiding the difficulties associated with rates below 0.5s, presented by Coltheart and Langdom. Most importantly, it makes the whole experiment shorter and thus less tiring for participants.

Furthermore, in order to ensure that all participants started off by looking at the same part of the screen, we presented a fixation cross on an empty, white screen for 2 seconds before every word cloud was displayed. The word cloud itself lasted 15 seconds, and after that time it disappeared. The goal of the participant was to remember as many words as they could, so after that time they would see a screen where they could enter the words they recalled by typing with the keyboard. They were not instructed to recall the words in any type of order and were told that spelling mistakes were accepted. To avoid causing any stress or pressure, the time they had to type in the words was unlimited, so they only had to press a "Done" button in order to move onto the next trial.

After 5 trials, the experiment would allow the participant to take a break that could last up to one minute, but they could also choose to continue if they didn't feel like they needed it. To ensure that the participants were familiar with the way they had to type in the words, we added a *trial* screen at the beginning of the experiment, before the first trial, where they got the chance to try to write down any word they wanted on the screen, so that they knew how to do it once the real experiment begun.

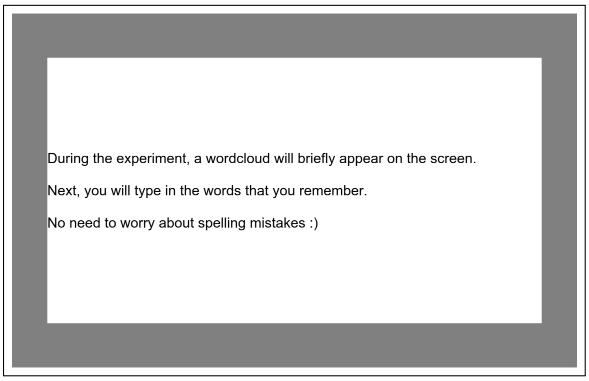
By programming the experiment in Psychopy we were able to automatically record each participant's answers in *.csv* files, which contained their anonymous ID's, their gender, their age, and the remembered words on each trial of the experiment.

The following figures show a short version of the experiment, where we can see the fixation cross, an example of a word cloud presented, the screen where the participant could type in the words, and the *break* screen. The 15 wordclouds of the experiment can be found in the Appendix 5.3.



Click the "Start" button to begin the experiment. You will first be presented with a demo of how it works. You can press Escape at any time if you wish to stop the experiment.

Figure 1: *Start page*



During the experiment, a wordcloud will briefly appear on the screen.

Next, you will type in the words that you remember.

No need to worry about spelling mistakes :)

Figure 2: *Instructions to participant*

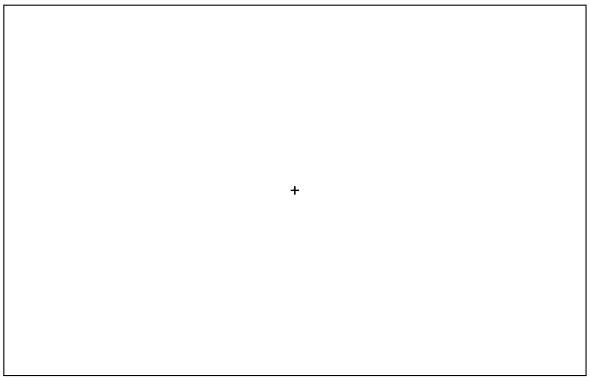


Figure 3: *Fixation cross*

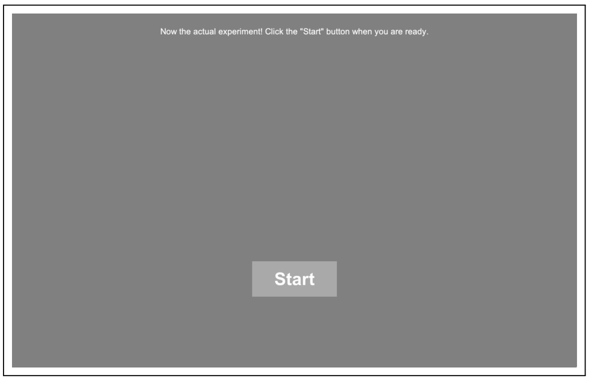

Now the actual experiment! Click the "Start" button when you are ready.

Figure 4: *Break after 5 trials*

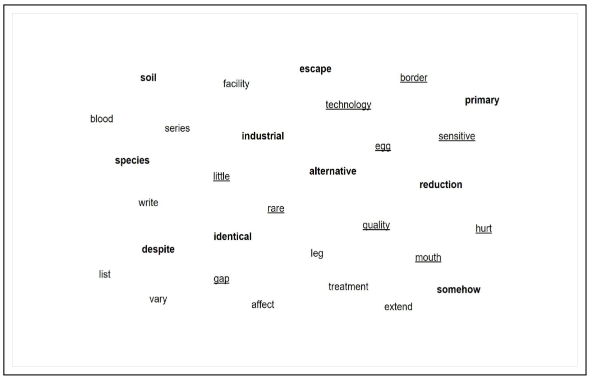
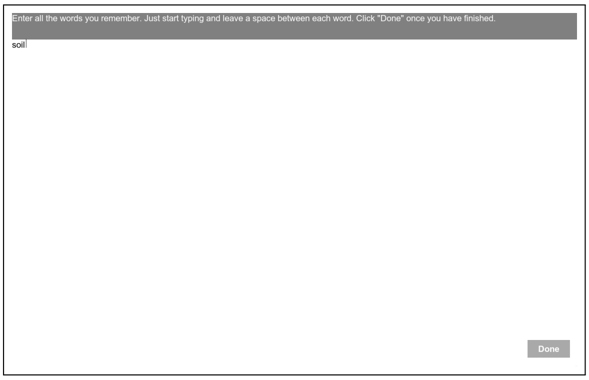


Figure 5: *Example of a wordcloud*


Enter all the words you remember. Just start typing and leave a space between each word. Click "Done" once you have finished.

soil

Figure 6: *Screen to type words*

For the experiment's design, we conducted a within-subjects study. The reason for this is because there is a lower variation between conditions, since all participants would get the same test, and the amount of participants needed for this is usually smaller. Furthermore, we established the following variables:

- Independent variable: the font style, with three conditions (**bold**, <u>underlined</u>, plain).
- Dependent variable: the number of memorised words of each type, where in each trial there are equal number of words presented in each font style (number of bold memorised words, number of underlined memorised words, number of plain memorised words).
- Control variables: the amount of words presented on the screen, the number of total emphasised words and of each type, the average length of the words, the time given to look at the screen, that all students are from DTU, and that they are 18 to 35 years old.
- Random variables: the gender of the participants.

2.1.1 Creation of the word clouds

In this section we will be explaining in more detail how we created the different word clouds, and what we did to the data that we used.

The words presented on the word clouds were obtained from the New General Service List (New GSL) [8]. The first GSL was published in 1953 by Michael West and it was a list of the most frequent words in written English. The New GSL used in this experiment is an updated version. By using words from a *core general vocabulary*, as described by the authors, we mitigated the risk that a participant might not be familiar with a word and not memorise it for that reason. The list initially consisted of the 2494 most frequently used words in the English language, and they were classified by frequency and class.

The idea was to make 15 groups of 30 different words in each, so that we could create the word clouds for the experiments. Each participant would be presented the same word clouds, as they had to do 15 trials each, and no word would appear twice in the whole experiment, meaning they all had to be unique.

The first thing we did to the initial word data set was to remove the *abbreviation* word class, since we did not consider these to be full words worth presenting. The following modification we applied to the remaining list was to remove all the duplicate words (there were words that appeared more than once since they could be of different word classes). These changes ended up giving us a total pool of 2313 words.

We randomly picked 15 times 30 words, without replacement, to create 15 word clouds. After that, we split each of the lists of 30 words into 3 randomly assigned groups: 10 bold words, 10 underlined words, 10 plain words. We also changed all words to be written in lowercase only, as we did not want capitalised letters to potentially influence recall of certain words. Finally, we computed the average word length per font type, as word length is one of our control variables. The average word length was 6.32 for bold words, 6.43 for underlined words and 6.39 for plain words. With a relative difference of under 2% between font types, we consider that we have controlled for the word length. We also performed ANOVA tests to compare the distribution of word length between font types, which will be further explained in 2. However, we have not controlled for the number of syllables.

The next step was to make the actual word clouds. A template with the position of the 30 words was created, such that words would be evenly spread across the screen. The same template was used for every word cloud in the experiment to ensure they had the same layout. Finally, the empty word cloud template was populated with the words in their assigned formatting.

2.1.2 Data treatment and protection

The data was fully anonymised by assigning a random unique identifier to each participant. It should be impossible to trace an identifier or experiment results to a participant. The only data we stored for each participant was their age, gender and the memorized words they wrote down after each trial (in the order they typed them). The data was stored in a shared Teams folder accessible only to group members (Emma, Marah and Apolline) and uploaded to a private Deepnote workspace for analysis. The data was not shared outside the group.

2.1.3 Risks and ethical considerations

When we designed the experiment, there were some ethical considerations we had to keep in mind regarding the data analysis that this project is based upon. Firstly, there is a bias towards younger participants who are between 18 and 25 years old, in particular students that are studying at DTU. The reason for choosing this constraint is that, in general, younger participants who are used to reading and remembering information continuously usually have a stronger memory, and we wanted all participants to have this common condition.

When running the actual experiment, we also had to keep in mind some risks towards the participants. There was the possibility that the given task may cause a certain level of stress on some participants, especially if they intended to perform really well on the experiment, so we tried to avoid them feeling pressured on the outcome of their test. To mitigate this risk, we let them type down the words that they remember themselves, instead of interrogating them. They were given unlimited time to type in their answers to avoid a feeling of time pressure, and they were also offered to take breaks twice during the experiment. The environment was made as friendly as possible, where we offered them coffee and cake, which also served as motivation for recruiting participants.

Another important issue we took into account was that the words we presented were not offensive in any way, and that they could not be misinterpreted. To address this potential risk, the words were picked from an already curated list made by linguists. This also mitigated the risk that the participant might read words that they did not know, which could cause distress as they would struggle to recall them due to a language barrier, specially since many participants were not native English speakers.

Participants who took part in the experiment should not share the words seen on the screen with other participants that were doing it afterwards, as it could have helped them remember some of them and made the results unreliable. Participants should also avoid sharing results with each other, so that there is no comparison or feeling of competition among them.

2.1.4 Informed consent

Participants were given an informed consent form before the experiment begun and once they were aware of what it consisted of. No person was able to participate in the experiment unless they read the information sheet and signed the consent form. Both of these documents can be found in the appendix of the report 5.1.1 and 5.1.2.

2.2 Data analysis

2.2.1 Pre-processing of data

As it was previously mentioned, we were able to record each participant's answers in *.csv* files that Psychopy generated for each of them. The log file recorded the participant's id, the participant's gender and age, as well as their typed-in answers. Psychopy also records additonal data such as the time spent typing for each word cloud and the duration of the break taken by participants.

Once we had gathered all our data, we first needed to do a cleaning of it. The main issues we had to look into were if people had done any spelling mistakes or typed any extra spaces when writing down their answers. Since we needed the words to be correctly written and separated by only one space before analysing the data, we went over each participant's recorded answers and ensured these two things by correcting the files. Then we had to check that all the words that they had typed belonged to the same wordcloud. Indeed, some participants sometimes remembered words from older wordclouds or even wrote down words that had never appeared on the screen. We decided to only keep the correct responses and include the responses with spelling mistakes by correcting them. If a word was written that did not belong to the word cloud, it was deleted.

2.2.2 Analyses and assumptions

After doing this we proceeded to analyse the data on python. We created the necessary functions to read the relevant data from the *.csv files* and we created a data frame with it. Since our dependent variable was the number of words memorised of each font type, we differentiated between them and counted them, which allowed us to get some initial statistics on the results from the experiment.

To further analyse the experiment's results and to check if our initial hypothesis was confirmed, we wanted to use a one-way, repeated measures, Anova test. The reason why we chose this test was because our experiment is within subjects, has one factor, but more than two levels (3 levels specifically).

Before performing the actual test we had to check if the results data passed the ANOVA assumptions. The first assumption says that the data from each condition is normally distributed. We performed *Anderson-Darling* tests to test the hypothesis that our data comes from a normal distribution. At a significance level of 0.05, the critical value for a normal distribution is 0.694. The statistics returned for the bold, underlined and plain words were 0.597, 0.371 and 0.472 respectively. Since all statistics are smaller than the critical value, the hypothesis cannot be rejected. We thus consider that our data is normally distributed.

The second assumption is also confirmed by checking if the three distributions have equal variance. We did this by using Levene's test, which gave us a p-value of $0.23 > 0.05$, meaning that the variances were not significantly different from each other. Finally, the independence of observations is not true in a within-subjects design, which is why we used ANOVA repeated measures.

2.2.3 Post-hoc analysis

Before formulating our main hypothesis for the experiment, and after going through the literature, we also came up with other possible hypothesis that we found interesting to research. However, due to the simplicity of the experiment, and in order to be able to limit the amount of factors to be taken into account, we decided to focus on the effect of word formatting on memory recall.

After analyzing this hypothesis, we had the chance to also explore another factor that we had previously thought of, and we will further analyze it on this section of the Methods chapter. The second research question we are formulating here is whether the length of the words presented to the participants seems to have any effect on their memorization capabilities.

In order to analyze our data in terms of word length, we counted, for each participant, the length of the words that they remembered, differentiating between the three formats of bold, underlined or plain. Then we averaged across the word clouds (trials), taking into account the three formats.

Control for word length

Since we thought that the word length could have an effect on the results, we did a test to see whether the words in each word cloud had the same word length for all the three formats. More accurately, we performed an ANOVA test on the length of the words testing whether the bold, underlined or plain came from the same distribution in each word cloud.

Length of words remembered vs not remembered

Another interesting thing we analyzed is whether the participants were better at remembering short words than long words, as we thought that might be affecting the results. Here we found the words that each participant missed from each word cloud (without taking the formatting into account) and then we computed the average across word clouds (for each participant), finally we did the same for the words participants remembered and compared the two groups with a t-test, as we only had two groups.

3 Results

This section focuses on the results obtained from our experiment and how we interpret them.

The main results obtained after analyzing our data can be seen in the following figure 7, where three Violin plots are displaying the distribution of the words remembered by the participants, which has been averaged across trials (across the 15 wordclouds presented to the participants), for each type of word format: bold, plain and underlined.

Only by looking at this plots we can observe how there is a big difference between the amount of bold words remembered compared to both plain and underlined, seeing as the bold mean value is around 2.5, whereas plain and underlined are both closer to a 1.4 mean. However, we don't see a very obvious separation in results between plain and underlined. This plot gave us an idea of how possible it was that our hypothesis was correct, which we further studied by using statistical testing.

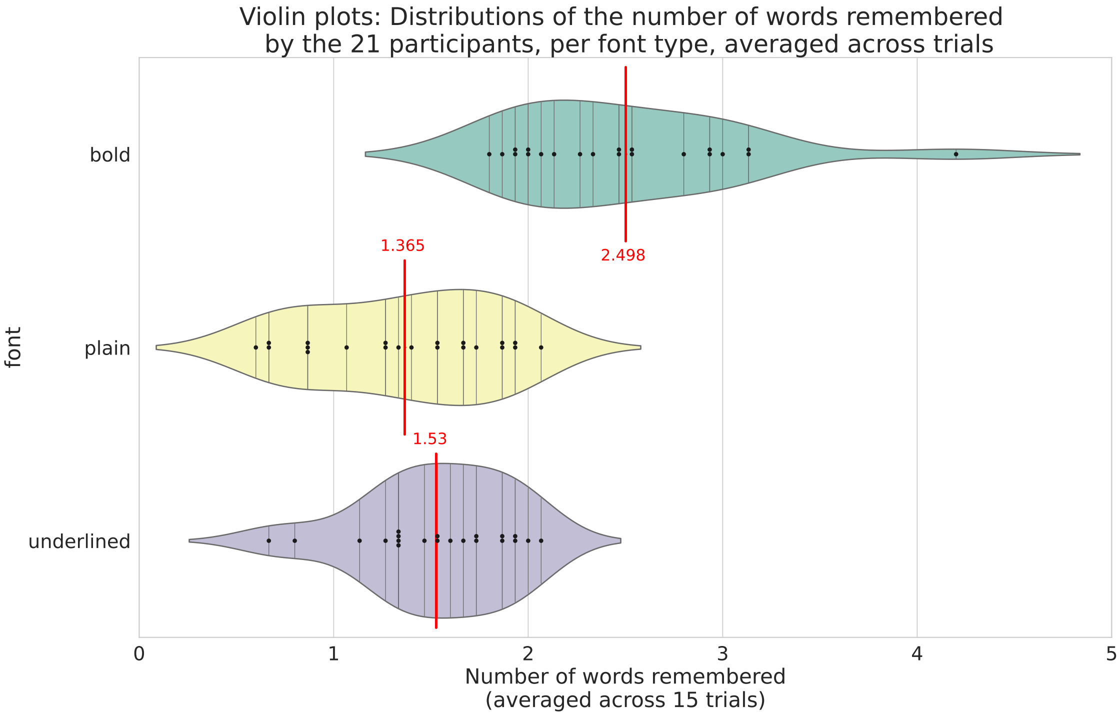


Figure 7: *Violin plots showing the distribution of words remembered by participants, averaged across trials (wordclouds), for the three font types bold, plain, underlined.*

3.1 Tests and results interpretation

As it was described in chapter 2, we wanted to perform an ANOVA test on the three conditions.

Once we had confirmed that the results passed the assumptions, we performed the test to see whether or not we could reject our null hypothesis that says that there is no difference in the means of the three conditions. The results of the ANOVA test are shown in the table below:

	F-value	degrees of freedom	P-value
WordType	36.3927	2	0.001

Getting a significant p-value ($p - value < 0.001$, $\alpha = 0.05$) means the three different conditions are statistically different and do not come from the same distribution. So we reject the null hypothesis.

After rejecting the null hypothesis, we wanted to check which of the three conditions -bold, underlined or plain- had the biggest difference. To do this we used t-tests with Bonferroni correction, which adjusts p-values because of the increased risk of a type I error when making multiple statistical tests.

We started of by doing t-tests between the three combinations of conditions: bold with underlined (BU), bold with plain (BP) and underlined with plain (UP). Once we had the p-values from these t-tests, we were able to correct them using Bonferroni.

The results of the t-test approved that there is a statistically significant difference between bold and underlined conditions ($p - value < 0.001$), and between bold and plain conditions ($p-value < 0.001$), but not between underlined and plain, where the p-value is not significant ($p - value = 0.166$).

3.2 Post-hoc results

Control for word length

Testing whether there was a significant difference in the length of the words between font types in each word cloud resulted in not detecting any difference, where each ANOVA test (we performed 15 as we have 15 word clouds) had p-values greater than $\alpha = 0.05$. We can thus consider that we have controlled for the word length, and that we are not favouring bold words by writing significantly shorter words in bold.

Applying a t-test on the length of words that were remembered and not remembered (regardless of the formatting), resulted in getting a significant difference between them. Figure 8 below shows the histogram of the two conditions.

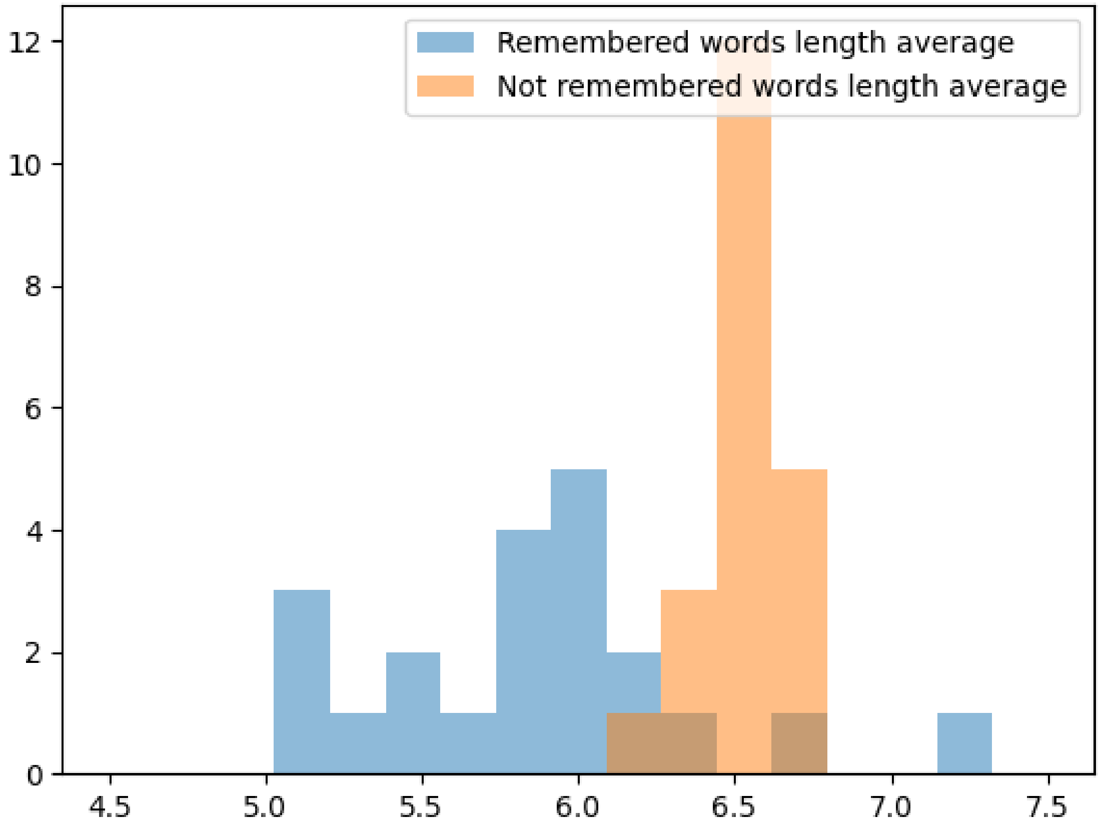


Figure 8: *Histograms showing the distribution of the length of words recalled by participants, along with the distribution of the length of words remaining on the wordcloud (ie. not recalled), for each participant, averaged across wordclouds*

4 Discussion

4.1 Interpretation of results

By looking at the results that we found through the tests described in the previous chapter, we can now reject the null hypothesis and confirm our initial hypothesis that stated that bold words are easier to memorise compared to underlined or plain words. The test results are also aligned with the plot seen earlier in figure 7, regarding the difference between the bold words compared to the two other formats. Both the ANOVA and the mean amount of bold words remembered confirm that at least one of the types of formats has a significantly different mean than the other two. Furthermore, after comparing each combination using corrected t-tests, we are able to find out which of the formats is different than the others, which, as expected from looking at the violin plot, is the bold formatting.

The post-hoc analysis showed that the average length of the words recalled by participants was significantly lower than that of the words which they did not remember. Different explanations are possible. The first one is simply that long words are harder to recall than short words. This is supported by the literature (see section 1.1), but often specifically regarding the length of the word in terms of number of syllables, and not the number of letters. A second explanation could be that, given the large amount of words, participants have to pick a strategy to focus on a limited number of words. In their choices, they might favour shorter words. Another possibility would be that shorter words attract their attention in the word cloud, such that they fixate on them and less on the longer words.

4.2 Conclusion

The experiment and analysis conducted enabled us to confirm the hypothesis that we had formulated, namely that formatting words in bold helps with their recall. From the start of the design of the experiment, we were expecting bold words to be easier to remember than words in other formattings as they visually stand out more on a page. Our hypothesis was confirmed statistically, even though our study has limitations, reported below in 4.3. Discussing the experiment with participants brought further insights on their memorisation techniques, which inspired future work ideas presented in 4.4 below.

4.3 Limitations and our contribution

One of the main limitations that we can observe in our study, which is typical for such a short, time-limited experiment, is the amount of participants that we had. Although we did achieve our initial goal of recruiting between 20 and 30 participants (21, to be precise), it is a clear fact that results would be even more significant if we had managed to have a bigger sample of subjects.

Moreover, although the experimental took place in relatively quiet places, the DTU Library and a classroom, one of the participants told us that they felt the need to put on noise-cancelling headphones in order to better focus on the task. Booking a room in a completely silent area, or providing our participants with noise-cancelling earplugs, might decrease the variance in performance between subjects.

Another limitation that our study faces is the fact that the experiment required people from different backgrounds to all remember words in the same language. Even though some research mentioned in section 1.1 suggests it might not be a big concern, we still believe that there could have been a difference in the difficulty of the task for those participants who were native English speakers versus the rest of them.

This is an important fact to take into consideration when interpreting results, and if we were repeating the experiment with a bigger sample size, we believe it would be worth registering the participant's mother tongue, in order to better understand their answers and how this might influence the data. The ideal setting would probably be to only recruit native English speakers, as there is an important variability in the level of English within non-native speakers.

In terms of our experiment's contribution to science, we believe that it has a positive impact since we managed to perform a simple study obtaining significant results. Furthermore, our experiment answers a question in the area of word memorization that has not been covered in previous research, and that it is worth looking into. Knowing the way formatting words impacts memory recall can be useful to know how to highlight information, specially for students in the age range of 18 to 25 years old, since this is the group of people the experiment was designed for.

4.4 Possible future work

Information memorization is a very broad area that can be influenced by many factors. Although for our experiment we decided to only focus on the type of formatting and its effect on memory recall, after talking to participants and also by looking at our data, we found several other topics or details that we find worth looking into if we were to expand this study.

Participants noticed, when doing the experiment, that they used other techniques to connect words in order to memorize them more easily, that are not related to whether the word was in bold or not. For instance, many people grouped words together when they shared the same root (like *education* and *educate* in Wordcloud 14). Others tried to make sentences with words they observed on the wordcloud, and associated words from a same lexical field, such as *stomach* and *dinner* (wordcloud 3). Making a sentence is an effective way of improving memory recall "lists shown word by word at the rate of about 10 words per second can be comprehended. Such stimuli appear to leave only fleeting memories; however, if they consist of words that form a coherent sentence, up to 12 or more items can be recalled" [5]. One of the participants reported plotting an imaginary diagonal line from the top-left to the bottom-right of the screen, and focusing on remembering the words on that line, as a sort of sentence. Since it is not a 'coherent' sentence, this technique might not be as effective.

Another factor that might have influenced their answers could also be the position of the words on the screen, and the order in which the subjects looked at the different quadrants of the wordcloud (if they went from top-left to bottom-right, or if they started at the center). This is a big area which one could study by also using other methods, such as eye-tracking.

Eye-tracking data could also be used to analyse the scanning process of subjects, specifically whether they look at bold words first and whether they spend more time fixating them. A new research question could then be about how text formatting influences the priotirisation of information.

Many participants repeated the words in their head, which is a process called rehearsal [5]. The phonological effect, shown by Schiano and Watkins [3] could also be studied, by looking at the phonological similarity of words in the wordclouds, or controlling for it when creating the wordclouds.

A limited number of participants declared that they did not feel that they remembered bold words significantly better. Since we did not record the identity of our participants, we cannot verify whether this statement is true for these individuals. Our study, however, has shown that there is a significant effect. Now that we have established that more bold words are recalled than plain or underlined words, we could also refine our study by looking at different levels of boldness. Certain fonts can be written in 'Medium', 'SemiBold' or 'Black', which are variations of a bold formatting, with different weights. Comparing between the effect of ordered levels of 'boldness' could be interesting in a future experiment.

In addition to comparing bold formattings, other types of formatting such as italic, and combinations of bold and italic or bold and underlined, could also be included in future studies, as well as coloured words.

Finally, all participants reported that the task was much harder than it seemed. Hopefully, it should not have caused them any stress. Designing a less strenuous version of the experiment, perhaps with more breaks, could be considered.

References

[1] M. A. Dzulkifli and M. F. Mustafar, “The influence of colour on memory performance: a review.,” *The Malaysian journal of medical sciences : MJMS*, vol. 20(2), 3-9, 2013.

[2] Kurt M. Joseph, Benjamin A. Knott, and Rebecca A. Grier, “The effects of bold text on visual search of form fields,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 46, no. 4, pp. 583–587, 2002.

[3] D. J. Schiano and M. J. Watkins, “Speech-like coding of pictures in short-term memory,” *Memory & Cognition*, vol. 9, no. 1, pp. 110–114, 1981.

[4] D. E. Copeland and G. A. Radvansky, “Phonological similarity in working memory,” *Memory & Cognition*, vol. 29, no. 5, pp. 774–776, 2001.

[5] Veronika Coltheart and Robyn Langdon, “Recall of short word lists presented visually at fast rates: Effects of phonological similarity and word length,” *Memory & Cognition*, vol. 26, no. 2, pp. 330–342, 1998.

[6] Nelson Cowan, “The magical number 4 in short-term memory: A reconsideration of mental storage capacity,” *Behavioral and Brain Sciences*, vol. 24, no. 1, pp. 87–114, 2001.

[7] C. Ruth Nott and W.E. Lambert, “Free recall of bilinguals,” *Journal of Verbal Learning and Verbal Behavior*, vol. 7, pp. 1065–1071, 1968.

[8] Vaclav Brezina and Dana Gablasova, “Is There a Core General Vocabulary? Introducing the New General Service List,” *Applied Linguistics*, vol. 36, no. 1, pp. 1–22, 08 2013.

5 Appendix

5.1 Methods

5.1.1 Informed consent

Informed consent form
[to be filled out by experimenter before the experiment]

Experiment:

Investigators:

[to be filled out by the participant before the experiment]

I confirm that:

- I was satisfactorily informed about the study concerned both verbally and in writing by means of the subject information letter.
- I have had the opportunity to put forward questions regarding the study and that these questions have been answered satisfactorily
- I have carefully considered my participation in the experiment
- I participate of my own free will

I understand that:

- My participation is voluntary, and I have the right to withdraw from the experiment at any time without having to give a reason
- My privacy is protected according to Danish law and European guidelines (GDPR; EU 2016/679)
- My consent will be sought every time I participate in a new experiment

I give my consent to take part in this experiment:

Full name

Date, Place

Signature

5.1.2 Information sheet

Dear participant,
In this letter we will inform you about the purpose of the experiment and the procedures. It is important that you read the letter carefully. If you have any questions, do not hesitate to contact us (Apolline Dersy s212836@dtu.dk, Emma Pereira s222668@dtu.dk, Marah Marak s182946@dtu.dk) for clarification.

Your rights as participant
Your participation in this experiment is voluntarily. This means that you can leave the experiment at any point in time without consequences for yourself, and without having to give a reason. We will ask you for an informed consent to participate after you have been informed about the experiment.

Purpose of the research project
In this research project we are investigating how the human brain chooses which words to memorise in a big group of them.

Data storage and handling
All recorded data (responses to questions, demographic information) will be anonymised, that is, stored not in connection with your name, address, CPR number, or any other information that would allow to identify you. For this, your data will be stored under a handle (key code). Personal information such as your name and email address will not be stored. More specifically, we will be storing your gender, age and the answers to the experiment (the words you type down in the order you do so). Your data will not be shared with any third party outside the group.

Information for the participants
You will have 15 trials. In each trial, you will be presented with a word cloud, and you will have 15 seconds to take a look at them. Your goal is to look at the screen and try to type down as many words as you can remember.

If you have any questions about the experiment, the methods or your safety and rights, do not hesitate to contact us (Apolline Dersy s212836@dtu.dk, Emma Pereira s222668@dtu.dk,Marah Marak s182946@dtu.dk).

With kind regards, Apolline, Emma and Marah.

5.1.3 Words for wordclouds

Table 1: Words randomly selected from the New GSL to be in the wordclouds (1-5)

1	2	3	4	5
expensive	winner	mixture	farmer	live
undertake	weight	absolutely	please	traffic
unless	AROUND	scene	sea	GENERAL
repeat	stand	absolute	matter	volume
knock	apartment	relief	patient	observation
PLACE	above	ROOM	PAPER	EACH
suppose	essentially	discover	investigate	finally
guide	discipline	behaviour	restore	restriction
January	DAY	suffer	tank	dead
election	once	check	wrap	facilitate
exciting	register	wonder	WORLD	FACT
boss	combine	simple	claim	collect
attitude	definition	miss	boot	stop
standard	context	contact	lack	usual
preference	MAY	fly	earlier	regional
mail	desire	stomach	safe	coffee
assume	borrow	likely	impress	NEVER
POSSIBLE	bone	economic	emotion	survive
PROCESS	entry	funny	South	faith
COUNTRY	WHICH	annual	accept	addition
operate	benefit	game	FOUR	recall
guard	oil	electricity	son	decent
perceive	expand	strategy	unit	calculate
approximately	naturally	abuse	establish	BRING
instead	switch	carefully	load	remove
MAN	contemporary	painting	BEFORE	God
BASE	saving	somewhat	fast	seat
website	net	ALONG	plastic	terrible
appoint	weather	participant	meeting	half
merely	majority	dinner	capacity	POINT

Table 2: Words randomly selected from the New GSL to be in the wordclouds (6-10)

6	**7**	**8**	**9**	**10**
attention	trust	WE	quiet	TAKE
approve	wear	college	RECEIVE	FIND
tear	movie	FAMILY	joy	restaurant
regardless	MAKE	sister	administration	fund
everyone	respond	sale	natural	air
consultation	selection	police	agreement	end
perception	institution	FEW	wet	commit
provision	barely	file	committee	solution
wash	incorporate	thank	ANOTHER	platform
relative	sentence	PARTICULAR	brand	clothes
creative	invite	fee	LEARN	declare
reject	sudden	WEEK	emergency	previously
FROM	instance	passenger	discuss	along
scenario	whole	achievement	respect	PER
around	proposal	complete	CAUSE	sell
clean	afraid	motor	THING	SERVICE
youth	constant	fight	employee	master
reflect	entire	fat	MOTHER	seek
quickly	HERE	episode	certificate	university
PAY	obtain	cake	response	DURING
wonderful	analyse	responsible	derive	collection
expose	topic	choice	officer	film
consistent	CONDITION	jump	appreciate	EVEN
delivery	fail	remind	east	disorder
military	disaster	advantage	kitchen	MEET
dry	cry	visit	illustrate	river
flower	lover	climb	sort	creature
victim	THEN	staff	neighbourhood	plane
head	attack	attract	back	element
upper	serious	FIRST	firm	religion

Table 3: Words randomly selected from the New GSL to be in the wordclouds (11-15)

11	**12**	**13**	**14**	**15**
fully	huge	unusual	frame	soil
announce	dark	mode	APPEAR	facility
female	client	partnership	education	industrial
prayer	settle	order	owner	LITTLE
oppose	primarily	dance	wrong	rare
save	before	entirely	sample	hurt
YEAR	philosophy	comprise	alone	vary
everybody	elsewhere	below	MUCH	egg
detailed	SOME	NO	rating	WRITE
description	ALTHOUGH	extent	eligible	primary
shortly	struggle	progress	flow	reduction
indicate	integration	urban	pair	affect
researcher	shop	feature	account	gap
appointment	NEED	cheap	educate	mouth
drink	ACTUALLY	AMONG	December	technology
text	reading	recover	root	somehow
budget	feed	concern	construct	series
improvement	row	spot	lunch	escape
surprise	weekly	easily	folk	treatment
purpose	accuse	CREATE	TYPE	despite
sand	equally	inner	character	species
principle	spending	factor	network	QUALITY
severe	EXPECT	SITUATION	twice	blood
TO	CHANGE	fear	THOUGH	list
AT	independent	pattern	international	sensitive
equal	pool	appeal	theory	alternative
potential	intention	START	understanding	leg
onto	slightly	difficulty	living	extend
happy	soul	SOMEONE	vehicle	identical
maintain	click	DIE	WIN	border

5.2 Results

Table 4: Number of words recalled by participants per font type, averaged across trials

	A	B	C	D	E	F	G
bold	1.87	2.53	2.13	3.13	3.13	2.47	1.80
plain	1.87	0.60	0.87	1.27	1.87	1.40	1.67
underlined	0.67	1.67	1.33	1.87	1.93	1.53	1.47

	H	I	J	K	L	M	N
bold	2.80	2.33	2.93	2.47	2.27	2.00	2.00
plain	1.07	1.53	1.33	0.67	0.87	0.87	1.67
underlined	0.80	1.27	1.53	1.13	1.60	1.87	2.07

	O	P	Q	R	S	T	U
bold	1.93	3.00	2.53	2.93	1.93	4.20	2.07
plain	1.93	0.67	1.27	1.93	1.53	1.73	2.07
underlined	1.33	1.33	1.73	2.00	1.93	1.73	1.33

5.3 All wordclouds

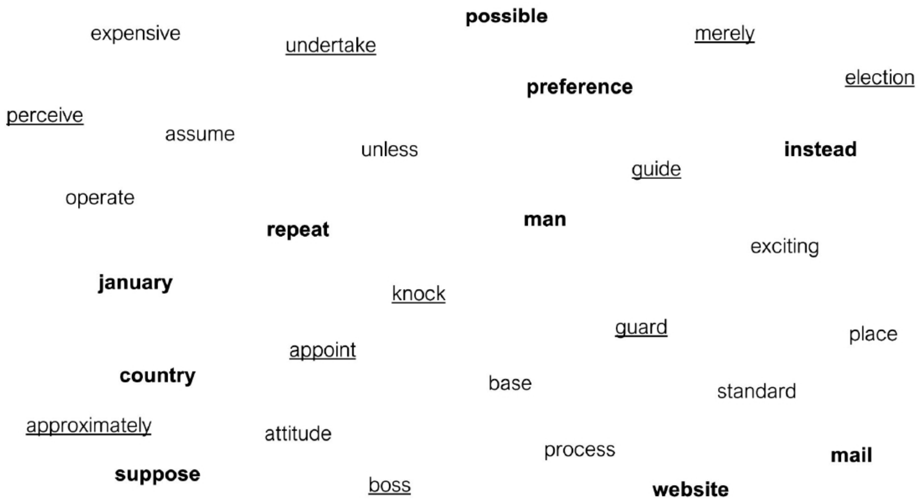


Figure 9: Wordcloud 1

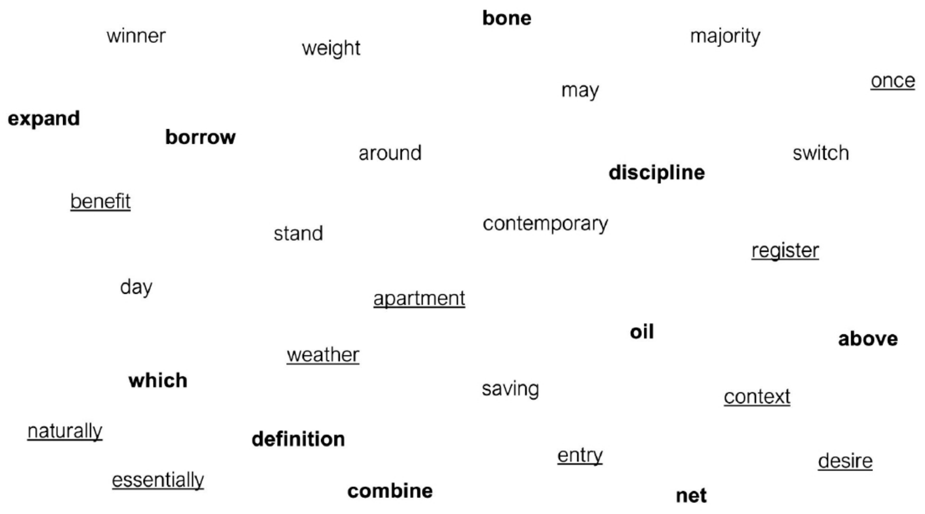


Figure 10: Wordcloud 2

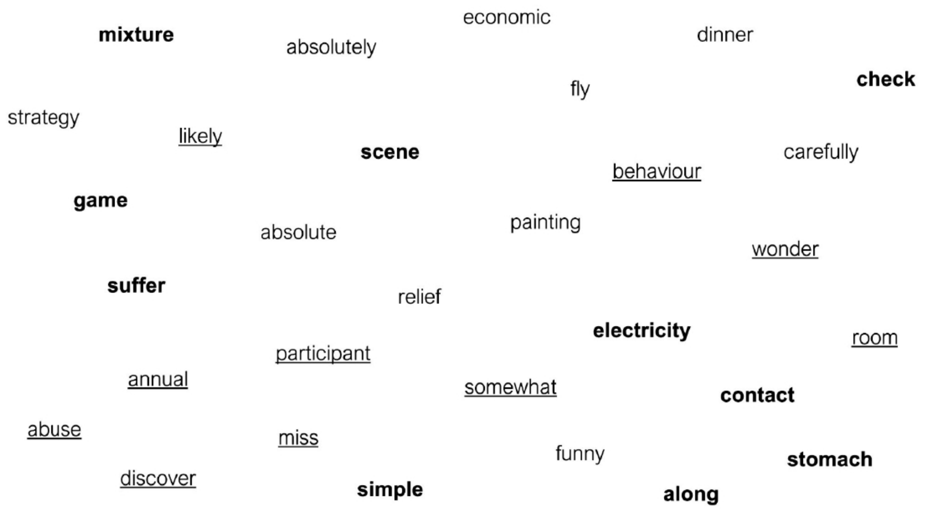


Figure 11: Wordcloud 3

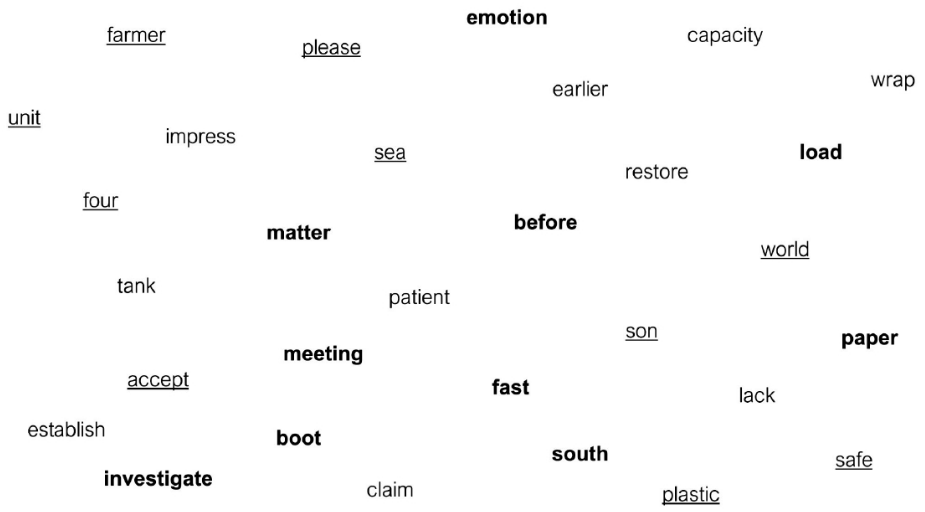


Figure 12: Wordcloud 4

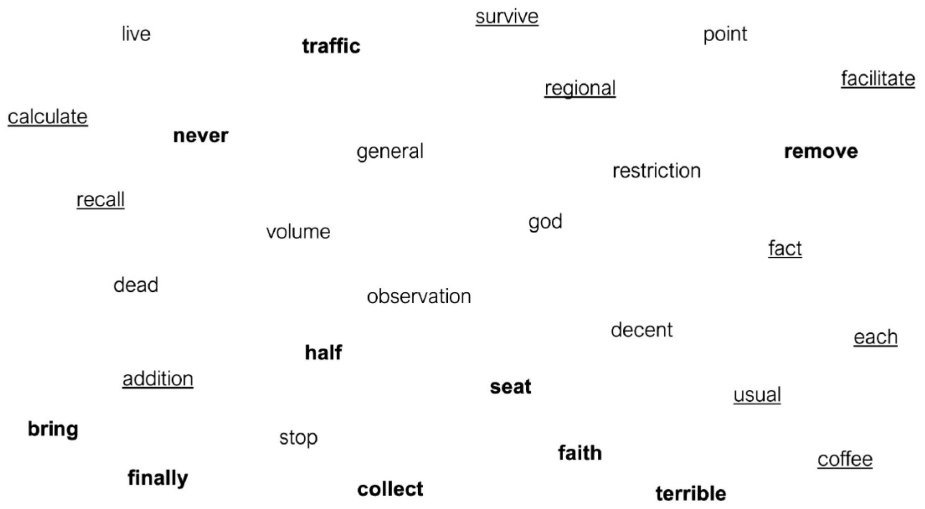


Figure 13: Wordcloud 5

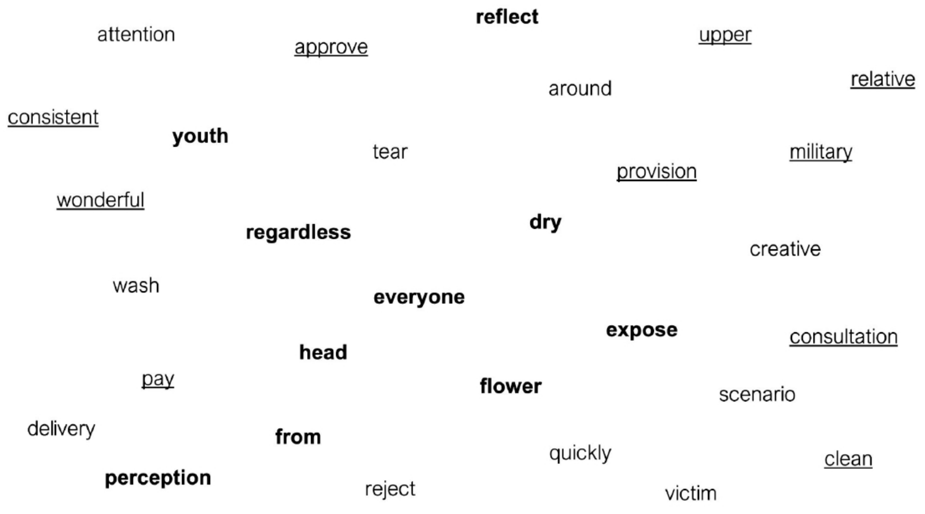


Figure 14: Wordcloud 6

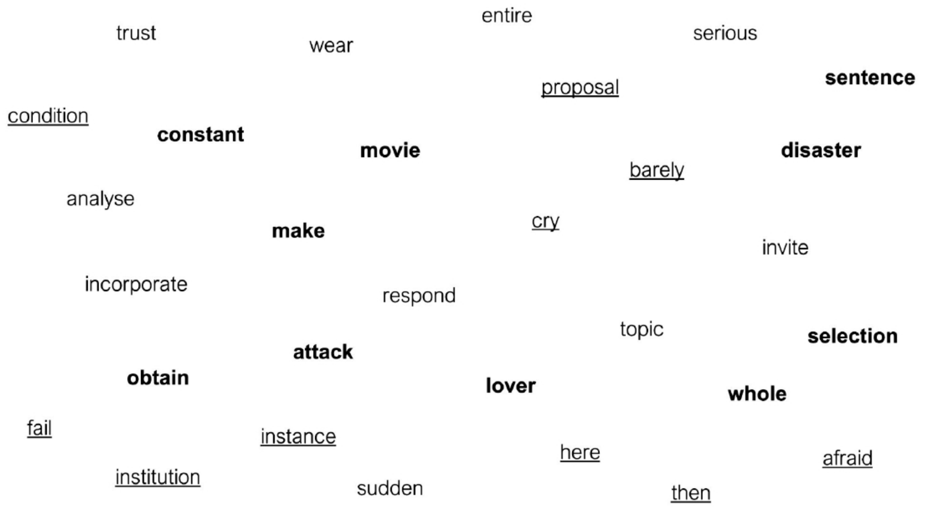


Figure 15: Wordcloud 7

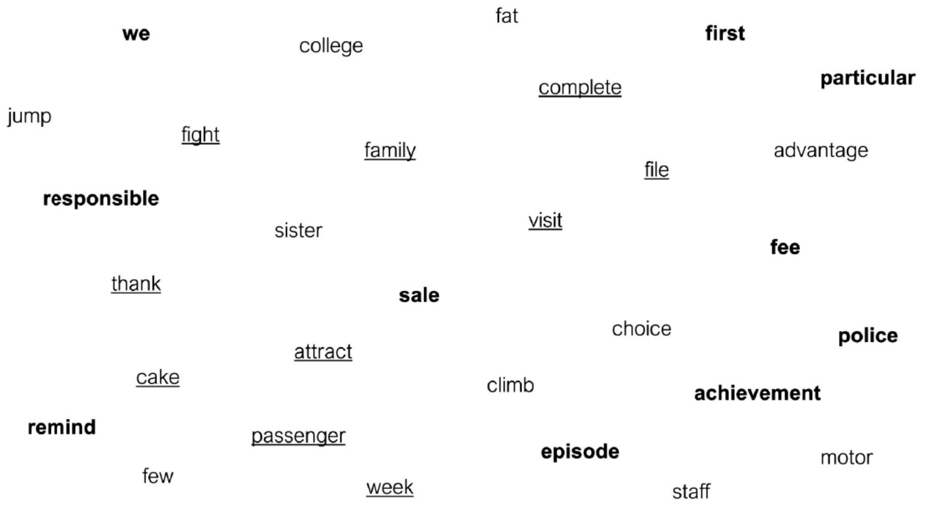


Figure 16: Wordcloud 8

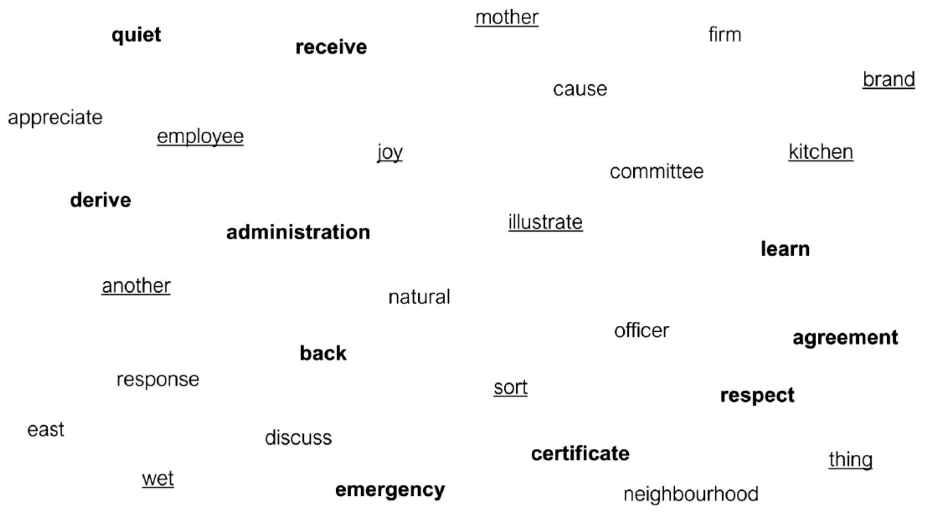


Figure 17: Wordcloud 9

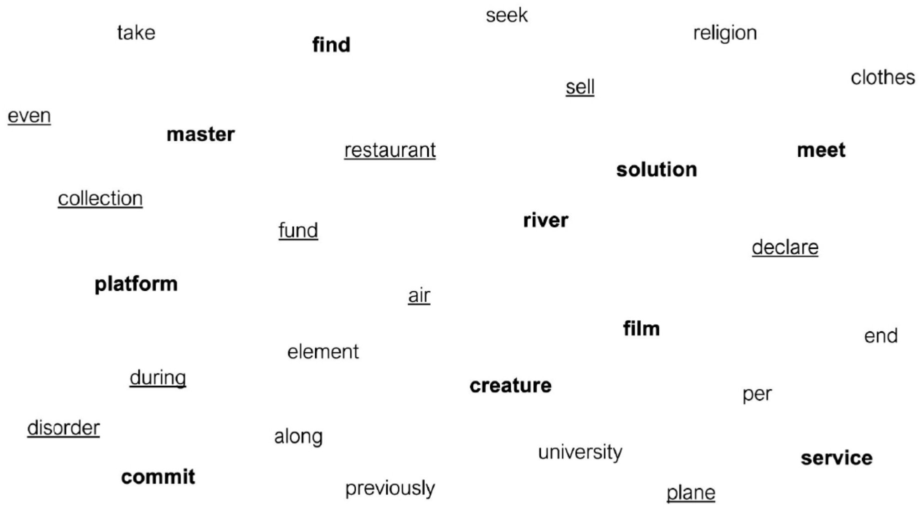


Figure 18: Wordcloud 10

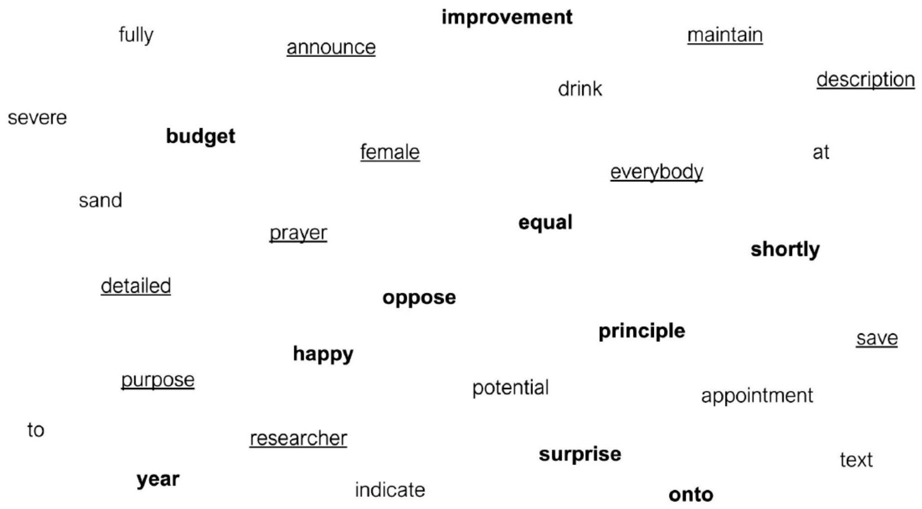


Figure 19: Wordcloud 11

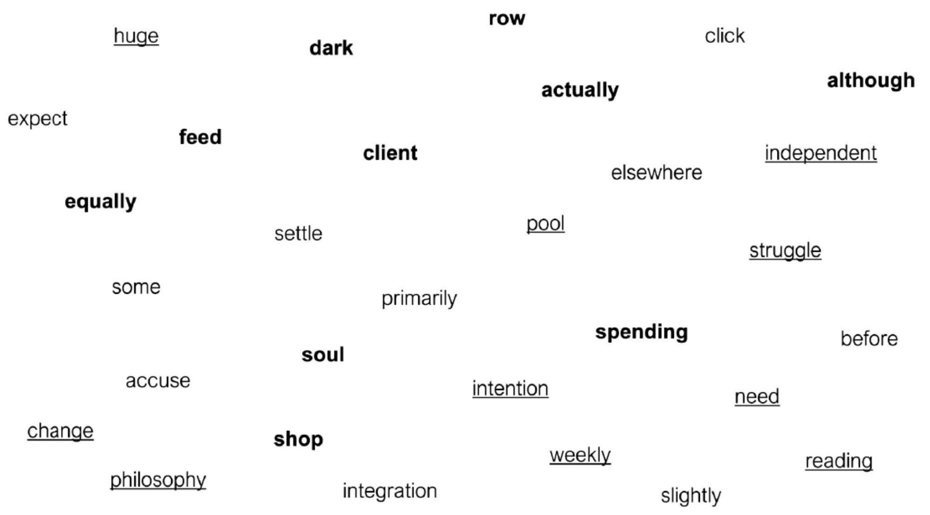


Figure 20: Wordcloud 12

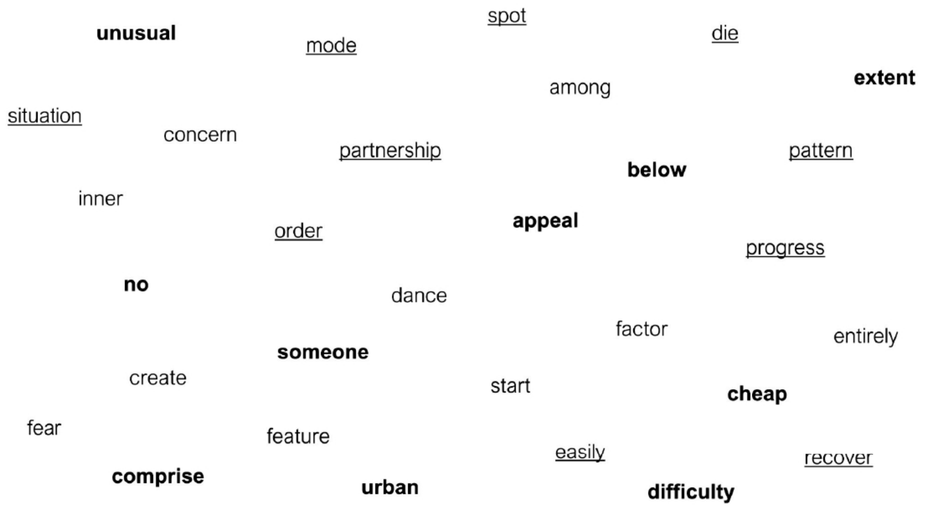


Figure 21: Wordcloud 13

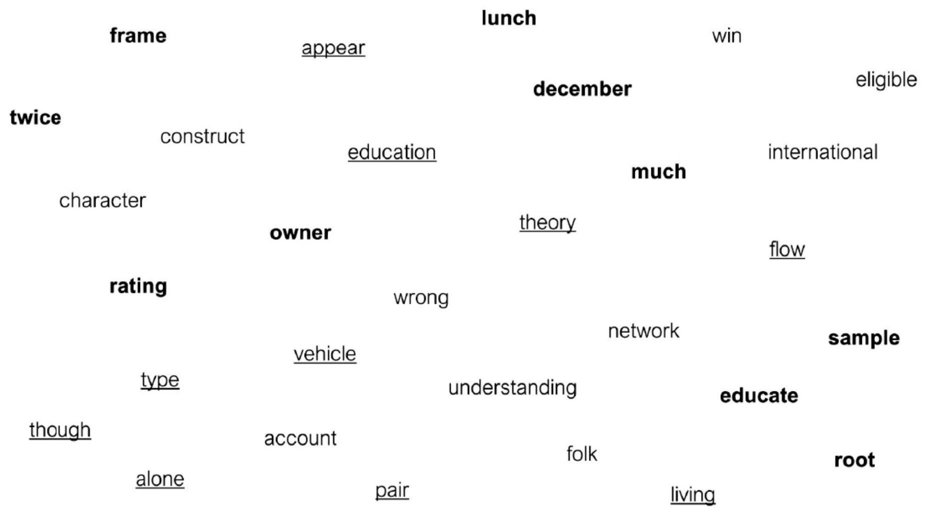


Figure 22: Wordcloud 14

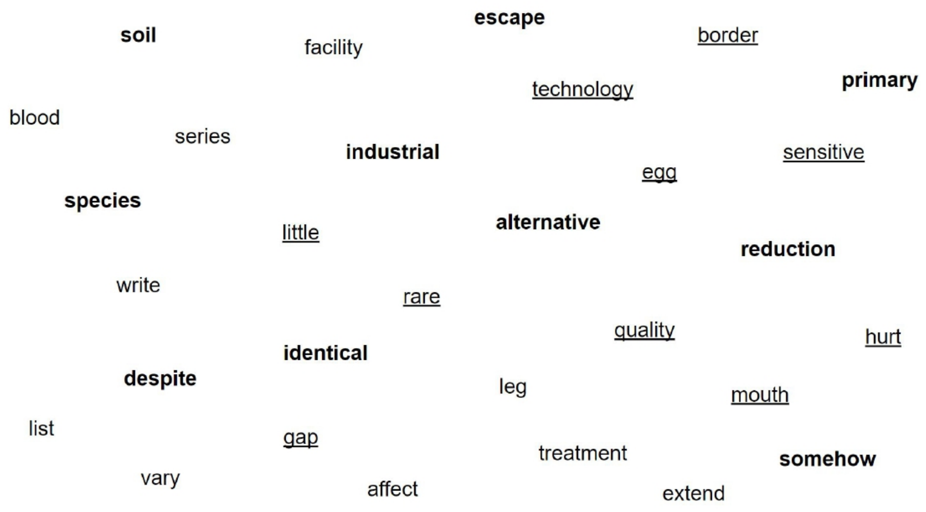


Figure 23: Wordcloud 15