



UMEÅ UNIVERSITY

An abstract, painterly background image with swirling colors of teal, blue, and orange, resembling a landscape or a close-up of a mineral surface.

Proceedings of
Umeå's 26th Student Conference in Computing
Science

USCCS 2023

Suna Bensch (editor)

UMINF 23.01
ISSN-0348-0542

Preface

The Umeå Student Conference in Computing Science (USCCS) is organized annually as part of a course given by the Computing Science department at Umeå University. The objective of the course is to give the students a practical introduction to independent research, scientific writing, and oral presentation.

A student who participates in the course first selects a topic and a research question that they are interested in. If the topic is accepted, the student outlines a paper and composes an annotated bibliography to give a survey of the research topic. The main work consists of conducting the actual research that answers the question asked, and convincingly and clearly reporting the results in a scientific paper. Another major part of the course is multiple internal peer review meetings in which groups of students read each others' papers and give feedback to the author. This process gives valuable training in both giving and receiving criticism in a constructive manner. Altogether, the students learn to formulate and develop their own ideas in a scientific manner, in a process involving internal peer reviewing of each other's work and under supervision of the teachers, and incremental development and refinement of a scientific paper.

Each scientific paper is submitted to USCCS through an on-line submission system, and receives two reviews. Based on the review, the editors of the conference proceedings issue a decision of preliminary acceptance of the paper to each author. If, after final revision, a paper is accepted, the student is given the opportunity to present the work at the conference. The review process and the conference format aims at mimicking realistic settings for publishing and participation at scientific conferences.

The conference is the highlight of the course, and this year the conference received 10 submissions (out of a possible 15), which were carefully reviewed by the reviewers listed on the following page.

We are very grateful to the reviewers who did an excellent job despite the very tight time frame and busy schedule. As a result of the reviewing process, 8 submissions were accepted for presentation at the conference. We would like to thank and congratulate all authors for their hard work and excellent final results that are presented during the conference.

We wish all participants of USCCS interesting exchange of ideas and stimulating discussions during the conference.

Umeå, 10 January 2023

Suna Bensch

Organizing Committee

Suna Bensch

With special thanks to the reviewers

Suna Bensch

Henrik Björklund

Johanna Björklund

Frank Drewes

Jerry Eriksson

Thomas Hellström

Lili Jiang

Timotheus Kampik

Ayush Kumar Varshney

Sudipta Paul

Ola Ringdahl

Mantas Simkus

Table of Contents

Comparison of usability with a focus on efficiency between iOS and Android system icons	1
<i>Tilda Engberg</i>	
Efficiency and interaction design within smart TV text input methods . . .	11
<i>Johanna Lindoff</i>	
Exploring How User Desirability Is Affected By Emojis in Search Results	23
<i>Jakob Marklund</i>	
Sigal: Instantiation for SCAN	35
<i>Willeke Martens</i>	
Reducing Gender Biases with Semi-Supervised Topic Modelling	49
<i>Salome Müller</i>	
Minimizing Lost Updates Under Read-Atomic Isolation With Lazy Transactions	69
<i>Lucas Paes</i>	
Privacy-preserving mechanisms for graph databases: a preliminary study .	83
<i>Duarte Silva</i>	
The Uncanny Valley Effect in Zoomorphic Robots: Univariate analysis . . .	97
<i>Jiangeng Sun</i>	
Author Index	111

Comparison of usability with a focus on efficiency between iOS and Android system icons

Tilda Engberg

Department of Computing Science
Umeå University, Sweden
`id19teg@cs.umu.se`

Abstract. Icons are used daily to facilitate understanding at various times and situations. This study focuses on the usability of icons on iOS and Android smartphones. There are previous studies that have compared these two systems against each other, but they have then included the entire Graphical User Interface (GUI). There are also studies that have compared application icons between these systems but not the system icons themselves. This lead to the research question of this paper, *how does the usability with a focus on the efficiency of system icons differ between Apple iOS and Android systems in smartphones?* The approach to answering the research question was a quantitative study with user tests. The test was constructed with help of a created test application with the intent of finding requested icons as quickly as possible. The participants were grouped into four groups depending on which system they were familiar with and also which icons they tested. The result of the test showed a significant difference between testing a familiar and unfamiliar system. On the other hand, it could not be proven that there is a difference in efficiency between iOS and Android system icons.

1 Introduction

Smartphones are a part of our daily life and new applications are released continuously and have to satisfy the user's needs. One of the needs is usability. One factor to consider is then the choice of suitable icons. Usability describes how easy it is to use a system and can be measured in five components constructed by Nielsen [1]. These are satisfaction, efficiency, memorability, learnability, and error, though this study will only focus on efficiency to limit the scope. Nilsen defines efficiency as how quickly users can perform a task when they have learned the design. Different types of systems and applications use different types of icons to improve usability within the five components. The two leading operating systems [2] in the smartphone industry are iOS and Android, which have their differences. One of the differences is the system icons, which are the icons built into the operating system. Android has guidelines that follow its Material Design ¹ and iOS has a collection of its own San Francisco symbols ² (SF symbols) as system icons.

¹ <https://fonts.google.com/icons>, Material Design, accessed 2022-09-20.

² <https://developer.apple.com/sf-symbols/>, SF Symbols, accessed 2022-09-20.

The aim of this paper is to investigate the following research question: *How does the usability with a focus on the efficiency of system icons differ between Apple iOS and Android systems in smartphones?*

Languages have different rules and expressions, and icons have their own languages for expression and explanation. It would be easier if icons and symbols followed the same guidelines to increase smartphone usability independent of the operating system. There exist significant studies about icons in the area of smartphones and icons in other contexts. Furthermore, there are studies that compare the whole system of iOS and Android. This study only focuses on the system icons and not the context of the GUI. The icons are tested independently from the interface.

2 Earlier work

In 2018 [3] an experimental study investigated icon characteristics on mobile systems between age groups. Ghayas et al. found that the user experience increased the more familiar the users were with the icons. The study also concluded that the recognition of icons could be improved by grouping them together in context.

Icon characteristics in the context of complexity, concreteness, and distinctiveness described in a study from 2000 [4] have achieved results from 5 different experiments. The results of the experiments that measured complexity with response time on finding icons showed that simple icons (less detailed) are more effective when the goal is to find an icon fast. Therefore, McDougall et al. suggest that interface design should prioritize simple icons. Moreover, when the goal is understanding it could be advantageous to use detailed icons.

Besides the system icons, there are application icons that differ depending on the operating system. In a study from 2015 [5] the user preference and recognition of application icons were investigated. From app stores on smartphones, different icons were collected and tested on users. Moreover, accuracy, recognition time, and subjective opinion were measured and collected from the tests. As a result, the study showed that detailed icons with many elements were preferred by the users. On the other hand, the recognition time was improved by simplified icon design.

3 Method

To answer the research question a quantitative study was formed with a focus on testing the efficiency with four test groups consisted of:

1. iOS users tested iOS icons
2. iOS users tested Android icons
3. Android users tested Android icons
4. Android users tested iOS icons

By having groups that tested on an unfamiliar system, the test could measure if there was any difference between the icons they were familiar with and not. The participants of this study were mainly collected from students at Umeå University with a total sample size of 20 participants with 5 per group. The specified target group was chosen to be young adults (16-29 years) and they had to use either an iOS or Android smartphone.

3.1 Application and material

A survey was created to gather knowledge about what system the participants were familiar with and other factors that could have affected the result. Another purpose of the survey was to verify that the participants belonged to the target group before the test. The following questions were asked in the survey:

1. Age?
 - (a) 16-19 years
 - (b) 20-29 years
2. Which smartphone system are you most familiar with?
 - (a) Apple iOS
 - (b) Android
3. How much time do you spend on your phone on an average day?
 - (a) 1-2 hours
 - (b) 2-4 hours
 - (c) 4-6 hours
 - (d) Over 6 hours

A test application was created in the prototype tool Figma to test the efficiency of the icons. The set of icons that were collected was picked from Google Material Design and iOS SF icons. The most frequently used icons in smartphones were picked out by sorting them by popularity, see Figure 1 and 2. Many Android and iOS icons are identical to each other, some of them were not included in the test because it was not interesting to test icons against each other if they look the same.



Fig. 1: Set of chosen Android icons.



Fig. 2: Set of chosen iOS icons.

3.2 Procedure

The participants were informed that the test is anonymous, and the results could not be connected to them. After the survey, the participants were informed about how the test is done. During the test, a computer screen displayed either the iOS or Android icons depending on which test group the participant belonged to. In total, the participants were asked to find the same set of 11 icons with English names. In addition, all icons were randomized before each new task to prevent the users from recognizing where the icons are placed. The Android and iOS icons were randomized in the same order for the same icon to ensure that the placement would not affect the result. The test groups did not have information about what type of icons they tested, and the icons were shown out of context on a blank page, see Figure 3. When the requested icon was clicked on, the icon shifted color to green and a button appears to continue to the next icon, see Figure 4. If participants clicked on the wrong icon, it changed color to red and they could continue until the correct icon was found. Furthermore, the test measured how long it took for the participants to find the requested icons. To measure the time it took for the users to find the icons a screen recorder was used to minimize errors in timing.

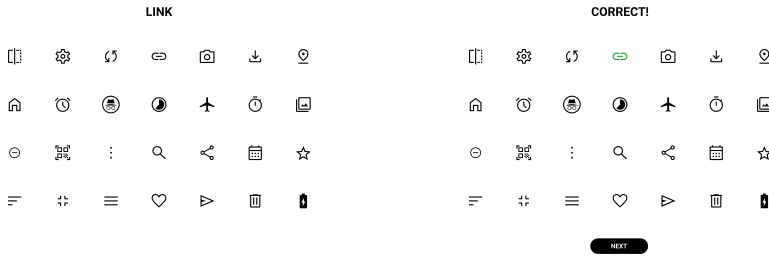


Fig. 3: Test application before click.

Fig. 4: Test application after click.

4 Results

From the survey answers in Table 1, the majority were 20-29 years old and had a screen time between 2-6 hours per day. By grouping the participants by screen time and their mean time to find each icon there was a maximum difference of 2,4 seconds between the two age groups.

A compilation from the icon tests has been made by calculating the average time in seconds it took to find an icon for each participant, see Table 2. To be able to statistically test if there is a difference between the efficiency in sections 4.1 and 4.2 the data from Table 2 has to be normality distributed. To check the data for normality a Lilliefors test in Matlab was done. First, all the data from Table 2 was tested and it showed that the data is normally distributed. Later in section 4.2 the data also had to be normally distributed. Then a Lilliefors test

Table 1: Results from the survey.

	Result
Age (years)	16 - 19: 10% 20 - 29: 80%
Smartphone system	iOS: 50% Android: 50%
Screen time / day (hours)	1 - 2: 10% 2 - 4: 45 % 4 - 6: 40% 6: 5%

Table 2: Mean time of finding icon for each participant in seconds.

	Familiar system	Unfamiliar system
iOS users	5,69	9,95
	1,91	6,80
	7,18	8,30
	3,46	19,31
	5,20	4,75
Android users	5,41	13,51
	7,68	11,37
	4,11	7,97
	8,81	11,91
	7,55	7,31

was done again with the data of the familiar users testing on familiar icons. The test showed that the data is normally distributed.

4.1 Difference of efficiency between testing familiar and unfamiliar icons

To verify that the test results were not influenced by which system the different participants were used to, a two-sample f-test and t-test were calculated. The f-test was done to see if the two groups had unequal variances. Depending on the result from the f-test, a suitable t-test could be chosen. The data used for the calculation is the average time it took to find an icon for each participant, see Table 2. Test participants were grouped into two groups based on whether they were testing icons they were familiar with or unfamiliar with. The result of the f-test resulted in a p-value of 0,061 which fails to reject the null hypothesis that says the groups have equal variances. From the result of the f-test, the two groups' variance is considered equal. Therefore, the t-test with two-sample assuming equal variances was chosen and the hypothesis can be seen below.

$$H_0 : m_1 - m_2 = 0$$

$$H_1 : m_1 - m_2 \neq 0$$

Where m is the average value in seconds it takes to find each icon. Group one (m_1) is the group that tested the icons they were familiar with, and group two (m_2) tested the icons they were not familiar with. The null hypothesis describes that there is no difference between the two groups. The alternative hypothesis describes that the groups of familiar and unfamiliar systems differ in efficiency.

Table 3: T-test of difference between testing familiar and unfamiliar icons in seconds.

	Familiar system	Unfamiliar system
Mean	5,70	10,12
Observations	10	10
StDev	3,32	
P-value	0,008	

In Table 3, the p-value from the t-test was calculated to be 0,008 which is less than the significance level of 0.05. With this result, the null hypothesis could be rejected and this difference is considered to be statistically significant. Therefore, it is shown that it is a difference between testing a familiar and an unfamiliar system in efficiency.

4.2 Difference of efficiency between iOS and Android icons

To find out which type of system icons has the highest performance in efficiency another two-sample f-test and t-test were constructed. As mentioned earlier, efficiency is measured by how fast a task can be performed when the user has learned the design. Therefore, the test groups of iOS user testing iOS icons and Android users testing Android icons was the two test groups. If the earlier t-test showed that it does not matter which system the user is familiar with, the other test groups could have been included in this comparison. The data used in the tests was the mean time it took to find an icon for each participant. The result of the f-test was a p-value of 0,8947 which fails to reject the null hypothesis that says the groups have equal variance. Therefore a t-test with two-sample assuming equal variances was chosen and the hypothesis for the t-test can be seen below.

$$H_0 : m_i - m_a = 0$$

$$H_1 : m_i - m_a \neq 0$$

In the hypotheses, m is the average time it takes to find each icon. Group one (m_i) consisted of iOS users testing iOS icons and the second one (m_a) consisted of Android users testing Android icons. The null hypothesis describes that it is no significant difference between the icons and the alternative hypothesis describes that iOS and Android icons differ in efficiency.

From the t-test data from Table 4, there is a difference between the mean value with 2 seconds, which supports the alternative hypothesis. On the other

Table 4: Two-sample t-test of difference between testing iOS and Android icons in seconds.

	iOS testing iOS	Android testing Android
Mean	4,69	6,71
Observations	5	5
StDev	1,98	
P-value	0,144	

hand, the p-value of the test is 0,144 which is greater than 0,05. The null hypothesis fails to be rejected and it is not significant to conclude that there is a difference between the icons within efficiency.

The result is also interesting in a view of inspecting the average time for each icon which can be seen in Figure 5. From the results, the icons that differ the most in time between the groups are *Link*, *Share*, *Send SMS*, and *Do not disturb*. These icons do not only differ in time, but they also differ in appearance, see Figure 6. Both *Link*, *Share*, and *Do not disturb* showed that Android users had problems with finding and recognizing. In fact, the only icons that iOS users had less efficiency than Android users were *Send SMS*, *Photo library*, and *Timer*.

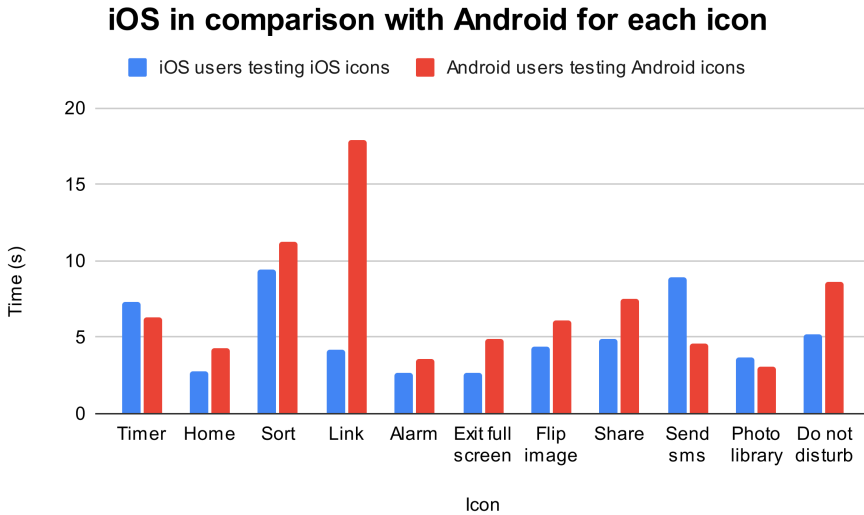


Fig. 5: Participants testing familiar system icons.

5 Discussion

This study tested the icons out of context from the GUI. If the icons were tested in the context of the whole GUI the participant's screen time could have









	Link	Share	Send sms	Do not disturb
Android				
iOS				

Fig. 6: Icons that had the most different results between the groups.

affected the result. With the whole interface, the participants could have been more familiar with the icons. On the other hand, it would have been difficult to test and measure because the participants could have different versions of the operating system. The purpose of this paper was to test the usability of different icons and take them out of context to make a clearer result because icons should explain themselves.

The result showed that there is a difference between testing a familiar and unfamiliar system and therefore half of the participants could not participate in the comparison between the iOS and Android icons. Consequently, there were only 10 participants that could contribute to answering the research question. There is no minimum of participants (observations) in a t-test but it is always advantageous to have more participants to get a more accurate result. On the other hand, the result could have been split into one for experienced users and one for new users. With two comparisons the other 10 participants could have contributed to the result.

Though the result could not prove that there is a difference between iOS and Android icons there is an indication that iOS icons could have better efficiency. The icons in Figure 6 where the icons with the most time difference in results are also icons that differ in appearance. iOS users had better results in 3 of the 4 icons. Moreover, by looking at the diagram in Figure 5 iOS has better efficiency in 8 out of 11 icons. Though this result it is not possible to draw any conclusions.

5.1 Sample of participants and icons

The participants were selected at Umeå University because of the time limit and to gather enough participants. In combination that the icons being named in English this could have affected the result when some of the participants asked what the icons are called in Swedish. Therefore, the icons could have been requested in Swedish to ensure that the participant's English competencies did not affect the result.

Another factor that could have affected the result was the sample of icons that was chosen. Material Design where the Android icons were collected from their website and there is a sort function on popularity. It was simple to choose the icons with the highest popularity and compare them with the corresponding iOS icons. On the other hand, iOS has colored icons in some cases. If colored icons were used instead of black and white, iOS users may have improved in time. To be consistent and make the test as equal as possible only black and white icons were used.

5.2 Suggested future work

In this paper, the difference between iOS and Android system icons with a focus on efficiency has been studied. The result of the difference is interesting although it is not statistically proven. However, in the future, there are opportunities to develop this study. The areas that could be taken into account to develop and modify are the following:

- More participants
- Another age group
- Another sample of icons
- Construct the test in the participant's native language

The above-mentioned areas could improve the result of the comparison, especially with more participants.

5.3 Conclusion

The purpose of this paper was to investigate the difference between iOS and Android system icons with a focus on efficiency. The chosen method was to perform user tests on an application to compare the different icons. The result of the tests was that there is a difference in efficiency if the user is familiar with the icons. Consequently, the participants that tested on unfamiliar icons could not contribute to answering the research question. The difference in efficiency between iOS and Android system icons only included the participants that tested the icons from the system that they were familiar with. The result showed that it is no significant difference between the system icons with a focus on efficiency. Though it could not be proven statistically there is an indication that iOS could have better efficiency by inspecting the mean time for each icon, see Figure 5. However, the indication and the data from the tests could not answer the research question of how the system icons differ from each other in effectiveness.

References

- [1] bt Mohd, N.A., Zaaba, Z.F.: A review of usability and security evaluation model of ecommerce website. *Procedia Computer Science* **161** (2019) 1199–1205
- [2] Mahalakshmi, M.K., Kavitha, K., et al.: A comparative study on customers satisfaction towards android operating system and iphone operating system in moblie phone. *Annals of the Romanian Society for Cell Biology* (2021) 12337–12344
- [3] Ghayas, S., Al-Hajri, S.A., Sulaiman, S.: Experimental study: The effects of mobile phone icons characteristics on users' age groups. *Journal of Computer Science* **14**(8) (2018) 457–478
- [4] McDougall, S.J., De Bruijn, O., Curry, M.B.: Exploring the effects of icon characteristics on user performance: the role of icon concreteness, complexity, and distinctiveness. *Journal of Experimental Psychology: Applied* **6**(4) (2000) 291

- [5] Chen, C.C.: User recognition and preference of app icon stylization design on the smartphone. In: HCI International 2015 - Posters' Extended Abstracts. (2015) 9–15

Efficiency and interaction design within smart TV text input methods

Johanna Lindoff

Department of Computing Science
Umeå University, Sweden
joli0630@ad.umu.se

Abstract. The use of smart TVs has become more common during the 2010s. With a large market of different smart TVs comes a large variance in TV text input methods. In this study, the efficiency of two methods is tested and discussed based on theories within interaction design. One of the methods is the combination of a virtual keyboard with the QWERTY layout and a regular remote control. The other method is a virtual keyboard with an alphabetical layout, structured on a single row. The associated remote control to this method has touch navigation instead of regular click navigation. The same group of 13 test subjects typed the same titles with both methods. By conducting a statistical analysis of the input times, the study claims that the method with the QWERTY virtual keyboard and regular remote control is more effective than the other. However, it is difficult to determine whether it is the keyboard, the remote control, or a combination of both that is responsible for the result.

1 Introduction

During the 2010s, the use of linear TV was reduced more and more [1]. Today in 2022, we can choose what to watch, where, and when. A popular alternative for this is a smart TV with built-in streaming services. With a range of different smart TVs comes a range of different systems with different interfaces and functionalities. These different systems have their own ways of offering text input functionality for their users. Some systems offer voice control, while others do not. Regardless, most systems offer an on-screen keyboard to navigate on using a remote control. These keyboards, in turn, can differ from system to system. Some keyboards are designed as the regular QWERTY keyboard which can be found on most computer keyboards. Other keyboards are in alphabetical order, structured in a single row or in a grid.

Apple tvOS and Android TV are two operating systems that are available on the smart TV market in 2022. Apple offers an on-screen keyboard that is designed in alphabetic order and structured in a single row. The on-screen keyboard belonging to Android is structured as the QWERTY keyboard. These systems offer different remote controls with different navigation technology. Android's remote control has four navigation buttons representing up, down, right,

and left. However, Apple offers a remote control with a touchpad that eliminates the need for users to perform a physical click to navigate. Instead, the user can swipe¹ up, down, right, and left.

The goal of this paper is to answer the research question *Comparing Apple tvOS and Android TV, which text input method is the most effective among young adults, and how can the result be related to theories in interaction design?*

2 Earlier work

In 2011, the possibility of interacting with television applications was quite new. Perrinet et al. [6] wrote an article that year about IDTV applications and text input methods. Four text input methods were analyzed, and three of the methods were virtual keyboards. More specific, the different virtual keyboards were QWERTY, see Figure 1, alphabetic grid, see Figure 2, and genetic grid, see Figure 3. The genetic virtual keyboard uses a genetic algorithm so that the most used letters are placed in the middle. The fourth method is the multitap mechanism that can be found on the remote control, see Figure 4. The test was conducted on Spanish test subjects, therefore, the keyboards are adapted to the Spanish language.

q	w	e	r	t	y	u	i	o	p
a	s	d	f	g	h	j	k	l	ñ
z	x	c	v	b	n	m			←

Fig. 1. QWERTY virtual keyboard.

a	b	c	d	e	f
g	h	i	j	k	l
m	n	ñ	o	p	q
r	s	t	u	v	w
x	y	z			←

Fig. 2. Alphabetic grid virtual keyboard.

The goal of the study was to determine the efficiency of the methods, mainly by measuring entry speeds and error rates. The study showed that the best method for simple texts was in descending order multitap, genetic, alphabetic, and QWERTY. Neither Apple nor Android offers a genetic virtual keyboard or a remote control with multitap functionality. Apple, on the other hand, has a

¹ A quick movement done with the thumb



Fig. 3. Genetic grid virtual keyboard.



Fig. 4. Multitap remote mechanism.

remote control with a touchpad that enables swipe gestures, whose efficiency has not been studied yet.

According to a study done in 2017 [2], the use of touch interfaces and gestures for search is heavily motivated by the low effort required from the users to provide input. Furthermore, in 2019 Fernandes Samuel et al.[4] conducted a study on young people and their relation to the swipe gesture. 34 students from the Department of Communication and Art of the University of Aveiro participated in the test. They investigated whether the students chose to use the swipe gesture or not when solving given tasks on a mobile phone, for example deleting an email on the Gmail mobile app. The study points out that students who are used to the Android operating system tend not to take advantage of the swipe logic. On the contrary, students who are used to the iOS operating system used the swipe logic to complete the tasks. Overall, the study shows that the tasks were solved fastest when the swipe gesture was used.

3 Interaction design theories

3.1 Fitt’s law

Paul Fitts [3] was a psychologist that stated that the movement time to a target is a function of the distance to the target divided by the size of the target. According to [3], it has been proven that *Fitt’s law* is the most successful and robust model of human movement. The law is expressed by the following function.

$$T = a + b \log_2\left(\frac{A}{W} + 1\right)$$

In the function, a and b are constants that vary depending on the type of pointer (e.g., finger, mouse). A is the distance from the starting point to the center of the target, and W is the width of the target. This function says that the bigger the distance to the target, the longer it will take for the pointer to move towards it, and the larger the target, the shorter the movement time to it.

Research in 2004 [7] showed that *Fitt's law* could be applicable to touch-screens. This information has been used to optimize the layout of on-screen keyboards. MacKenzie and Zhang [5] designed an on-screen keyboard for mobile phones called OPTI based on this interaction law. The goal of the design was that the targets should be large enough for users to accurately select them, that the targets would have ample spacing between them, and be placed so they can be easily acquired.

3.2 Jakob's law

Jakob's law [8], or *Jakob's law of internet user experience* states that users spend more time on other pages than on your page. This means that users create prejudices about how websites should work. They observe common patterns on the web and develop expectations for similar pages. This law can be applied to other user interfaces. Users create expectations about how smart TV user interfaces should work based on their previous experiences.

4 Method

4.1 Input methods

This paper aims to determine which of Android TV and Apple tvOS has the most efficient text input method. In order to resolve this, the methods have been tested and compared. The two methods are specified in Table 1 and both methods are adapted to the Swedish language since the native language of the test subjects is Swedish.

Table 1. Input methods

	Input method 1	Input method 2
OS	Android TV	Apple tvOS
Smart TV	Xiaomi MI Box 4k model MDZ-16-AB	Apple tvOS HD 32GB 4th generation
Keyboard design	QWERTY, see Figure 5	Alphabetical structured on one single row, see Figure 6
Remote navigation technology	Click, see Figure 7	Touch (Siri Remote 1st generation), see Figure 8



Fig. 5. Mi Box on-screen keyboard.

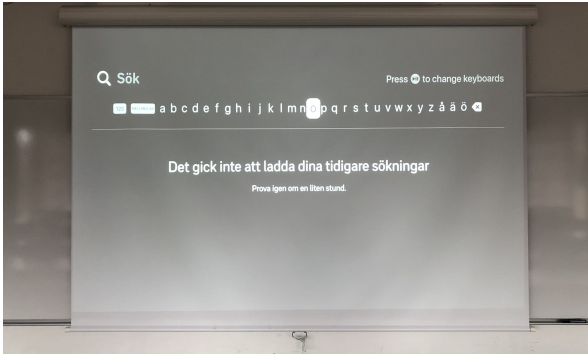


Fig. 6. Apple's on-screen keyboard.



Fig. 7. Mi Box remote control.



Fig. 8. Apple's Siri Remote (1st generation).

4.2 Procedure

To analyze the efficiency of the two methods, a test has been conducted to gather quantitative data. The same 13 people tested the Android method one day and then the Apple method a couple of days later. By testing the two methods on the same 13 subjects, the experimental variability could be reduced. If the test had been carried out on two different groups there would have been a risk that one group would contain a few who are extra fast or a few that are extra slow. The test subjects were 20-30 years old since the research question focused on young adults.

Before the test, the six following titles were gathered on a slideshow:

1. Bonusfamiljen
2. The playlist
3. House of the dragon
4. Känn ingen sorg
5. Top gun
6. Ur spår

The slideshow consisted of 6 white slides and 6 slides with each one of the titles. The white slides were interspersed with the text slides, where the first slide was white. The slideshow was shown on a computer placed on a table. The table was in front of the projector screen, which the test subjects entered text on, shown in Figure 5 and Figure 6. The test subject sat on a chair facing the computer screen and projector screen. An iPad was also placed behind the test subject, only recording the two screens during the tests.

During the test, the test subject was given the remote control and was given information about how the test would be carried out. One title was said aloud one at a time while the current title was shown on the computer screen. The time from when the title appeared on the screen until the test subject had written the complete title was measured. This was made easy by looking at the iPad recording after the test. Before the test subject was presented with each title, all test subjects started on the same letter on the keyboard. That was done because everyone should have the same conditions before each test. When using Apple, the test subjects started with the letter "n". When using Android, the test subjects started with the letter "h". That was done because these letters are placed in the middle of each keyboard. When the test subjects had written all the titles, they were asked the following questions:

1. How old are you?
2. How often do you use a text input method like these ones on a smart TV?
 - (a) Every day
 - (b) 2-3 times a week
 - (c) 4 times a month
 - (d) Less than 4 times a month
3. Which system are you most used to, Apple tvOS or Android TV?

5 Results

The profile of the test subjects is shown in Table 2.

Table 2. Test subjects

	Apple or Android	Smart TV search habits	Age
Person 1	Android	j4 times a month	24
Person 2	Apple	j4 times a month	22
Person 3	Android	Every day	20
Person 4	Apple	j4 times a month	22
Person 5	Apple	4 times a month	23
Person 6	Android	2-3 days a week	21
Person 7	Android	2-3 days a week	24
Person 8	Android	j4 times a month	20
Person 9	Apple	j4 times a month	28
Person 10	Android	j4 times a month	24
Person 11	Apple	Every day	24
Person 12	Apple	Every day	24
Person 13	Android	Every day	23

The test conducted with the Android method produced the results shown in Table 3. The test conducted with the Apple method produced the results shown in Table 4. The mean time it took to type a specific title appears at the bottom of each title column. The different mean times are compared to each other in Figure 9 and Table 5.

Table 3. Results for the Android method

	Title 1	Title 2	Title 3	Title 4	Title 5	Title 6
Person 1	22 s	22 s	30 s	32 s	9 s	8 s
Person 2	29 s	28 s	38 s	28 s	10 s	13 s
Person 3	20 s	19 s	27 s	25 s	9 s	10 s
Person 4	21 s	19 s	30 s	22 s	9 s	10 s
Person 5	16 s	16 s	25 s	24 s	7 s	9 s
Person 6	21 s	21 s	30 s	19 s	13 s	9 s
Person 7	19 s	18 s	35 s	19 s	10 s	12 s
Person 8	19 s	24 s	29 s	20 s	10 s	10 s
Person 9	18 s	18 s	27 s	23 s	9 s	14 s
Person 10	19 s	20 s	29 s	27 s	19 s	9 s
Person 11	24 s	24 s	46 s	31 s	18 s	18 s
Person 12	24 s	24 s	42 s	25 s	12 s	12 s
Person 13	22 s	22 s	29 s	21 s	10 s	11 s
Mean value	21,1 s	21,2 s	32,1 s	24,3 s	11,6 s	11,6 s

Table 4. Results for the Apple method

	Title 1	Title 2	Title 3	Title 4	Title 5	Title 6
Person 1	32 s	27 s	32 s	24 s	11 s	22 s
Person 2	34 s	33 s	59 s	30 s	12 s	22 s
Person 3	30 s	27 s	46 s	42 s	12 s	13 s
Person 4	28 s	25 s	40 s	32 s	13 s	16 s
Person 5	22 s	20 s	38 s	28 s	14 s	13 s
Person 6	60 s	27 s	42 s	24 s	15 s	26 s
Person 7	24 s	21 s	41 s	43 s	15 s	17 s
Person 8	33 s	25 s	39 s	26 s	13 s	13 s
Person 9	32 s	38 s	47 s	24 s	20 s	24 s
Person 10	36 s	40 s	62 s	40 s	13 s	15 s
Person 11	21 s	20 s	48 s	25 s	11 s	14 s
Person 12	26 s	29 s	43 s	30 s	20 s	16 s
Person 13	27 s	22 s	42 s	26 s	14 s	20 s
Mean value	31,2 s	27,2 s	44,5 s	30,3 s	14,1 s	17,8 s

Table 5. Mean values and their differences

	Title 1	Title 2	Title 3	Title 4	Title 5	Title 6
Mean value Android method (m_1)	21,1 s	21,2 s	32,1 s	24,3 s	11,6 s	11,6 s
Mean value Apple method (m_2)	31,2 s	27,2 s	44,5 s	30,3 s	14,1 s	17,8 s
$ m_1 - m_2 $	10,1 s	5 s	12,4 s	6 s	2,5 s	6,2 s

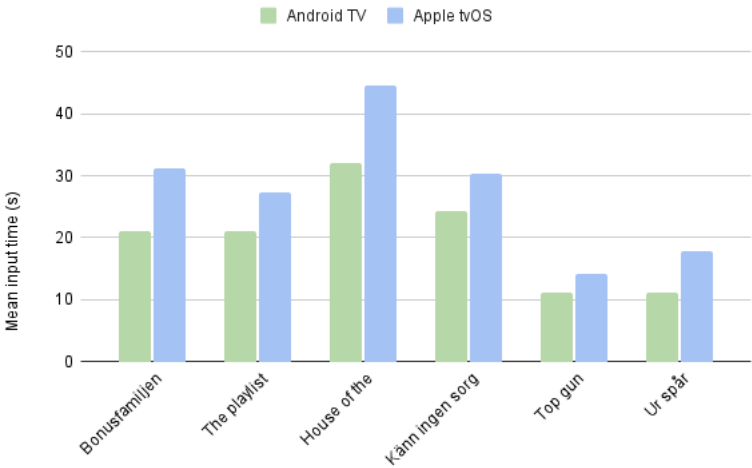


Fig. 9. The different mean values

Table 5 shows that the mean values differ from each method. The Android method shows lower numbers than the Apple method. Also, from the recordings that were done during the tests, it appeared that the error rate was more frequent when using the Apple method. Despite this, it is not possible to say that the Android method is more efficient than the Apple method without statistical analysis. Since the subjects for the two methods are the same, a paired statistical t-test can be conducted to evaluate the results. The hypothesis of interest is the following.

$$H_0 : m_1 - m_2 = 0$$

$$H_1 : m_1 - m_2 < 0$$

The variable m_1 is the mean input values from the Android method. The variable m_2 is the mean input values from the Apple method. The null hypothesis says that there is no difference between the input methods. The alternative hypothesis says that the Android method gave more efficient results than the Apple method. In order to determine that the Android method is more efficient than the Apple method, the paired t-test has to reject the null hypothesis.

A paired t-test can only be used if the differences between the mean values, shown on the last row in Table 5, are normally distributed. This can be checked by using the function *lillietest* in Matlab, a numeric computing platform. The expression $h = \text{lillietest}(\text{differences})$ in Matlab, where *differences* is the data on the last row in Table 5, gave the result $h = 0$, which says that the differences are normally distributed. A paired t-test can then be conducted with the data. This can be done in Matlab with the expression $h = \text{ttest}(\text{android}, \text{apple}, \text{"tail"}, \text{"left"})$. The *android* expression is the data shown on the second row in Table 5. The *apple* expression is the data shown on the third row in Table 5. The expression gave the result $h = 1$ in Matlab. The test thus rejects the null hypothesis which claimed that the efficiency of the two methods is the same. Therefore, the results show that the Android method is more effective than the Apple method.

6 Discussion

The goal of this paper was to answer the research question *Comparing Apple tvOS and Android TV, which text input method is the most effective among young adults, and how can the result be related to theories in interaction design?* The statistical analysis performed in the result section rejects that the text input methods gave equal input time. This study can therefore claim that the Android method is more effective than the Apple method among young adults.

By looking at the result data in Table 5, it appears that the Apple method gave higher input times than the Android method, which aligns with the statistical result. Despite this, earlier work claimed that the swipe gesture, which the Apple remote takes advantage of, is effective. However, it cannot be decided if it was the remote control technology, the keyboard layout, or a combination of the two that was ineffective.

6.1 Interaction law theories

According to *Jakob's law*, users build expectations about how applications should work based on previous experiences. Six of the thirteen test subjects that participated in the study claimed they were more used to Apple tvOS. With this knowledge, one could have believed that the test results would have been more equal between the different methods according to *Jakob's law*. When the test subjects that were used to Apple tvOS had tested the Apple method, some of them mentioned that they usually perform their search in a different way on their Apple tvOS. This is also shown in Table 2. Many of the Apple users only use the remote input method four times a month or less. Some of them said they use the voice search functionality while some search through their iPhone that is connected to their Apple tvOS. That might explain why the test results weren't more even, even though almost half were used to Apple tvOS.

Apple's keyboard design cannot be found on many other smart TVs, therefore it requires more resources for the user to understand how to best use the keyboard according to *Jakob's law*, especially for those who are more used to Android. This can be a reason why this keyboard gave higher input results than the other method. The keyboard belonging to Android has the QWERTY layout which many people recognize since it is common on computers. This can be a reason why it gave better results. However, it is interesting that previous studies have shown that the QWERTY keyboard is the slowest method when typing short pieces of text on a TV using a remote control.

As earlier mentioned, *Fitt's law* says that the movement time to a target is a function of the distance to the target divided by the size of the target. The shorter the distance and the bigger the target, the better design. When looking at the keyboard layout belonging to Apple, it appears that the longest possible distance is longer than the longest possible distance on the Android keyboard. Consequently, the layout of Apple's keyboard is not very user-friendly. On the other hand, it could be argued that the swipe gesture is a more efficient navigation technology than the clicking technology. The constants a and b represents the pointer (e.g., finger, mouse) in the function describing *Fitt's law*. However, it has not been possible to find the accurate constants that represent the touch gesture. The correct constants would maybe give a great value when describing the Apple's keyboard layout with the function of *Fitt's law*, proving that the swipe gesture has a positive impact.

As mentioned in the result, the Apple method gave higher error rates than the other method. It seemed more difficult to stop the cursor on the right letter when using the swipe gesture. Since the swipe gesture is said to be efficient, it could be argued that it is the remote control design by Apple that is poor, though it seemed hard to be accurate using this remote.

6.2 The test design

This study focused on younger adults. This target group was chosen since the research question has been of interest for a longer time. However, the test subjects were known from earlier. What impact this had on the study cannot be

determined, but certainly, it was a convenience measure. The study would have higher quality if the test subjects were randomly selected. Also, the study could have given more reliable results if more people participated, especially if more titles had been involved in testing the different methods. Six titles were used to compare the two methods. This is a very low number. Six different titles gave a result, but maybe not a fair one. In future studies, a larger amount of titles should be tested.

This study did not consider if the different input methods had a learning period, if the periods differed from each other, and what impact it had on the results. The learning period is the time period where the input times decrease significantly as the inexperienced test person has more experience with the method. There is a chance that the test subjects that were unfamiliar with an input method became faster at using the method as they typed more titles. One method may have required a longer learning period than the other method. As this study used six titles, there is a risk that the persons that were not used to a method, never reached the point where they felt comfortable with the technique. Each method should be tested in multiple sessions in future studies. As a result, the learning period becomes clear and the study could give more reliable results.

Despite this, the fact that typing titles like this on a TV does not correspond to reality. When a user wants to search for something on a TV, the user usually only writes the first letters of the title, before the item the user is searching for appears in the search results.

References

- [1] Alva Alriksson Lind. Den nya generationen tv-konsumenter : En undersökning om hur svenska producenter av barnprogram ser på övergången från linjär-tv till strömningstjänster, 2021. Bachelor's Thesis, Dalarna University, School of Culture and Society.
- [2] Juan Felipe Beltran, Ziqi Huang, Azza Abouzied, and Arnab Nandi. Don't just swipe left, tell me why: Enhancing gesture-based feedback with reason bins. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI 17, pages 469–480, New York, NY, USA, 2017. Association for Computing Machinery.
- [3] Xiaojun Bi, Yang Li, and Shumin Zhai. Fitts law: Modeling finger touch with fitts' law. CHI '13, pages 1363–1372, New York, NY, USA, 2013. Association for Computing Machinery.
- [4] Samuel Fernandes, Rui Rodrigues, and Lidia Oliveira. *Young Users and the Swipe Logic in Smartphones*. Meltemi Editore, 03 2019.
- [5] I. Scott MacKenzie and Shawn X. Zhang. The design and evaluation of a high-performance soft keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 25–31, New York, NY, USA, 1999. Association for Computing Machinery.
- [6] Jonathan Perrinet, Xabiel G. Paneda, Sergio Cabrero, David Melendi, Roberto Garcia, and Victor Garcia. Evaluation of virtual keyboards for

- interactive digital television applications. *International Journal of Human—Computer Interaction*, 27(8):703–728, 2011.
- [7] Barry A. Po, Brian D. Fisher, and Kellogg S. Booth. Mouse and touchscreen selection in the upper and lower visual fields. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 359–366, New York, NY, USA, 2004. Association for Computing Machinery.
- [8] Steven Schirra, Shraddhaa Narasimha, Sasha Volkov, and Justin Owens. Understanding user mental models through app sketches from memory. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery.

Exploring How User Desirability Is Affected By Emojis in Search Results

Jakob Marklund

Department of Computing Science
Umeå University, Sweden
`id19jmd@cs.umu.se`

Abstract. Emojis are a growing phenomenon that is used in various ways to enhance communication in computer-mediated communication. This study examines how user desirability is affected by emojis in search results. A desirability study ($N = 20$) is conducted where participants have to describe pictures of search results by selecting three reaction words. The participants are divided into two populations, one is exposed to search results with emojis, and the other only without emojis. The findings of our work show that the ratio between the positive and negative reaction words shows no significant statistical difference between populations. Furthermore, only 35% of all participants believed that emojis in search results might be a good idea.

1 Introduction

Emojis are becoming a more significant part of how we express ourselves in computer-mediated communication (CMC) [1]. The role of emojis in text-based CMC is to enhance the interaction between the sender and receiver by increasing expressiveness. Many brands are trying to create more substantial social and emotional ties by including emojis in their newsletters and social media accounts.

However, there is still debate on whether the user sees the usage of emojis by brands as a good thing. Some studies [2] have shown that emojis can enact a sense of incompetence, whereas other studies [3] found a correlation between emoji use and increased purchases in specific areas. Furthermore, there are indications¹ that people tend to focus more on the visual aspects when an emoji is present in email subject lines.

This study answers the question *How is user desirability affected by emojis in search results?* One population is only subjected to search results with emojis and the other population is only subjected to search results without emojis. Each participant undergoes the same experiment, with the exception that they only see search results that correspond to their population. The experiment assesses their overall reaction to a specific search result and examines the basis of their reaction.

¹ <https://www.nngroup.com/articles/emojis-email/>, Emojis in Email Subject Lines: Advantage or Impediment?, accessed 2023-01-03

2 Background

2.1 Emoji

Emojis are a relatively new phenomenon that originated from smiley and emoticons in the early 21st century [1]. These graphical symbols have a unique name and code (Unicode) that depicts different expressions, activities, animals, gestures, feelings, and other objects. The usage of emojis is ever-increasing, and as of October 2022, there are 3,633 emojis in the Unicode standard, and 92% of the world's online population uses emojis².

Emojis can improve the recipients' reactions in CMC by adding more contextual and emotional meaning [4]. Likewise, the sender can use emojis as a paralinguistic cue to further express their feelings and state their intention.

However, the interpretation of emojis can be highly ambiguous and personal. Some factors that influence the interpretation of emojis are technical differences and cultural background [5]. The emoji design is also different across different platforms [1], which further aggravates the ambiguity.

2.2 Search Engine Optimization

When shopping, around 60% [6] start their search process by visiting a search engine online. Search engine optimization SEO lets brands make their search results more user-friendly and thus generate more user interaction with their target customer. Google also plays a role in determining the appearance of these results. In fact, Google modifies approximately 60%³ of all title tags. Google may choose not to display certain emojis in search results if they believe the emoji could be disruptive or misleading⁴. Google tries instead, to use an equivalent word that conveys the same message.

2.3 Desirability

There are multiple ways to test how visual design affects user experience. This paper focuses on desirability.

Desirable can be defined as, 'Worth having or seeking, as by being useful, advantageous, or pleasing' or you can replace desirable with words, such as those that follow, to stimulate ideas - fun, engaging, natural, enjoyable, valuable, essential, suitable, worthwhile, beneficial, likeable, in demand, amusing, and appealing. [7, p. 57].

² <https://home.unicode.org/emoji/emoji-frequency/>, The Most Frequently Used Emoji of 2021, accessed 2022-11-24

³ <https://zyppy.com/seo/title-tags/google-title-rewrite-study/>, Google Rewrites 61% of Page Title Tags [SEO Study], accessed 2022-11-24

⁴ <https://youtu.be/a5J73nYDU8E?t=1983>, English Google SEO office-hours from January 28, 2022, accessed 2022-11-24

Since desirability can be somewhat elusive, it can take much work to test and evaluate. Usually, post-test questionnaires and semi-structured interviews are used to gather data. However, much time is spent transforming the test data to useful conclusions [8].

Microsoft usability engineers noticed these limitations and proposed a way [7] to get at the intangible quality of desirability directly. One of their solutions was to employ a set of reaction cards. During the test, participants are asked to choose cards that summarize their feeling about the experience that a specific artifact gave them. Each card contains a descriptive word or phrase, and the whole set is divided into positive and negative words. Microsoft’s usability engineers observed that people want to be friendly and prefer more positive words. Therefore they choose to keep the ratio between positive and negative words to 60% positive and 40% negative words.

A later study [8] on the process of using reaction cards found that the reaction cards give meaningful insight into how users perceive the experience of a specific artifact. Furthermore, they believe that reaction cards should not be the only way of gaining feedback. Instead, it should be used in conjunction with other satisfaction survey instruments.

3 Method

This desirability study uses reaction words to measure how user desirability is affected by emojis in search results. The participants are divided into two populations that are only exposed to search results with emojis or without emojis. The reaction words assess the participants’ overall reaction to a search result. However, to gain better feedback, reaction words should be used in conjunction with other satisfaction survey instruments, as suggested by [8]. Therefore, the participants also get to answer follow-up questions during the experiment. This allows the participants to explain the reason behind the choice of the reaction words. After the experiment, the participants are exposed to the other populations’ search results. During this time they must answer the post-experiment question, which is described later in this paper

The definition of a stimulus in this experiment is a picture of a single search result from Google, see Figure 1, 2 and 3. Some stimuli, like Figure 1 contain other visual elements like a rating and a picture. Whereas others only contain text, see Figure 2 and 3. Figure 2 and Figure 3 illustrates the difference between the no-emoji and emoji version of the same stimuli that each population got to see. Table 1 contains a translation of the text in Figure 1, 2 and 3. Table 2 describes all 13 search results used in this experiment and defines the topic, brand, and which emoji were inserted in the emoji version.

3.1 Participants

A total of 20 participants (Swedish residents, 50% women) aged between 19 and 30 ($M=23.8$, $SD=2.73$) voluntarily participated in this study.



Fig. 1. Example of a stimulus with a picture and an inserted emoji.



Fig. 2. Stimulus without emoji.

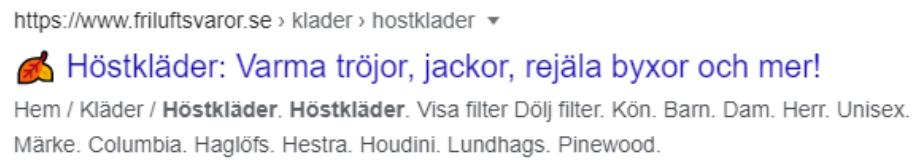


Fig. 3. Stimulus with an emoji inserted.

Table 1. English translation of Figure 1, 2 and 3.

Figure	Title	Description
1	Guide: Best pizza & pizzerias in Umeå 2022 - Utmärkta	Pizzeria Roberto Umeå, Ristorante Ruccola Umeå, Westside Restaurang & Pizzeria i Umeå, Mimo Pizzeria Umeå, Kebabnekajse Pizzeria Umeå, You might also like.
2 & 3	Autumn clothes: Warm sweaters, jackets, reliable pants and more!	Home / Clothes / Autumn clothes. Autumn clothes. Show filter Hide filter. Gender. Kids. Women. Men. Unisex. Brand. Columbia. Haglöfs. Hestra. Houdini. Lundhags. Pinewood.

Table 2. Each search results topic, brand, and emoji were inserted in the emoji version.

Topic	Brand	Emoji
Baking	Arla.se	
Pizza	Utmärkta.se	
Apartment	Bostaden.umea.se	
Sports	Iksu.se	
Nature	Umea.se	
Events	Visitumea.se	
Clothes	Friluftsvärld.se	
Technology	Kjell.com	
Repair	Fixiphone.se	
Soccer shoes	Sportamore.com	
Hotels	Hotels.com	
Skiing	Hemavan.nu	
Train	Sj.se	

3.2 Stimuli and Reaction Words

Each stimulus is a picture of a single search result. The search results are text search results and are chosen by the authors. The topics in Table 2 are chosen to include a wide variety of topics. Therefore, no participant can only get topics that they are either disproportionately positive or negative towards.

There are 13 stimuli without emojis, see Figure 2. A corresponding emoji version was created for each non-emoji stimulus, see Figure 3. The version with the emoji was created by inserting an emoji with the use of Google Chrome's developer tool. The choice and placement of emojis are inspired by a readability guideline⁵ for emojis.

The picture of the stimuli is taken from Google's Swedish search engine. The Swedish version is used since the experiment is created for participants with Swedish as their primary language. Google's search engine was used because it dominates the search engine market with 86-89%⁶ market share worldwide.

A total of 22 reaction words are chosen from Microsoft's usability toolkit [7] and translated to Swedish for this experiment. These words can be seen in Table 3. The order of the reaction words was scrambled before the experiment. The proportion of negative and positive reaction words is kept to the same proportion as suggested in [7]. Therefore around two-thirds of the reaction words are positive, and one-third are negative. The 22 reaction words are chosen to mimic the natural reactions one might get from the experiment.

⁵ <https://readabilityguidelines.co.uk/images/emojis/>, Readability guidelines - emojis, accessed 2022-11-27

⁶ <https://kinsta.com/search-engine-market-share/>, Search Engine Market Share: Who's Leading the Race In 2022, accessed 2022-11-27

Table 3. The 22 reaction words that were used in the experiment.

Positive		Negative	
Tilltalande	<i>Appealing</i>	Irriterande	<i>Annoying</i>
Övertygande	<i>Compelling</i>	Otydligt	<i>Unrefined</i>
Spännande	<i>Exciting</i>	Billigt	<i>Cheap</i>
Vänlig	<i>Friendly</i>	Förvirrande	<i>Confusing</i>
Engagerande	<i>Engaging</i>	Ineffektivt	<i>Ineffective</i>
Inspirerande	<i>Inspiring</i>	Distraherande	<i>Distracting</i>
Tydlig	<i>Clear</i>	Opersonligt	<i>Impersonal</i>
Motiverande	<i>Motivating</i>	Tråkigt	<i>Boring</i>
Professionell	<i>Professional</i>		
Relevant	<i>Relevant</i>		
Kreativt	<i>Creative</i>		
Kul	<i>Fun</i>		
Hjälpsam	<i>Helpful</i>		
Stimulerande	<i>Stimulating</i>		

3.3 Procedure and Measures

Twelve experiments were conducted on campus at Umeå university, and the remaining eight were conducted via Zoom. The participants were reminded of their rights to withdraw at any time and that the data would only be used for academic purposes. The participants were also informed of the test process and had a chance to ask questions. They were encouraged to “think aloud” during the experiment. The participants performed the experiments individually and were not informed that emojis were the subject of the research. Therefore the samples are assumed to be independent.

The moderator used a computer to display the 22 reaction words on half of the display and the stimuli on the other. The reaction words were presented as a mixed list, and the stimuli were presented one by one. The stimuli and reaction words were presented in the same order to all participants.

In the experiment, the participants would have to describe each stimulus by selecting three reaction words [7]. They were encouraged to choose words that they thought best fit the stimulus based on their reaction to the design of the stimulus. They were not allowed to choose the same reaction word more than once for each stimulus. During the experiment, the moderator continuously asked the participants follow-up questions to understand why they chose the different reaction words.

Lastly, when the participants had completed the experiment, they got to see the other version of the stimuli. This means that if they were a part of the no-emoji population they now got to see the stimuli that the emoji population saw. After some consideration, the participants had to answer the post-experiment question *do you believe that emojis in search results are a good idea?*.

4 Results

4.1 Reaction Words

The relative frequency of reaction words for both samples is shown in Figure 4. From this figure, we can see that on average, the two most chosen reaction words for both populations are *clear* and *helpful*.

The reaction words that had the most substantial difference between both populations were *Friendly*, *Exciting*, *Annoying*, and *Appealing*. The percentage difference is presented in Table 4.

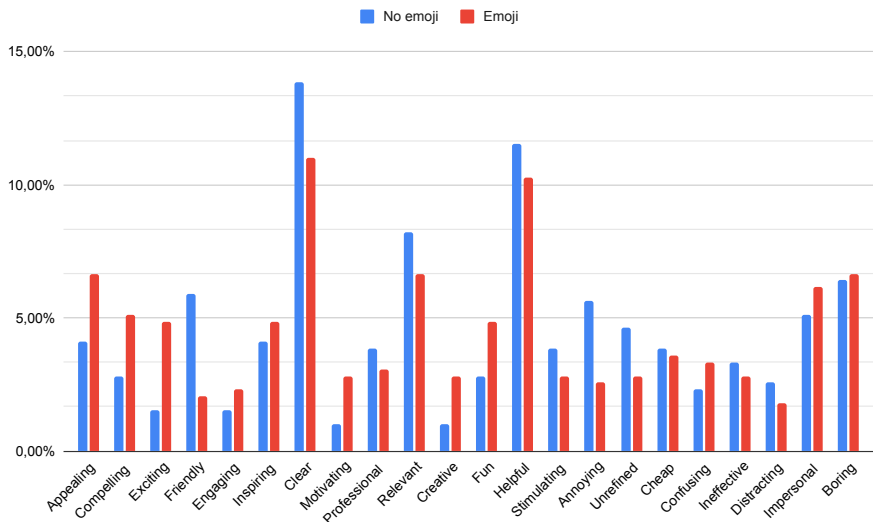


Fig. 4. Relative frequency of reaction words for emoji. Sorted left to right by positive to negative reaction words.

Table 4. The reaction words that had the most substantial difference between both populations.

	Friendly	Exciting	Annoying	Appealing
Type	Positive	Positive	Negative	Positive
Emoji	2.05%	4.87%	2.56%	6.67%
No emoji	5.90%	1.54%	5.64%	4.10%
Higher occurrence	No Emoji	Emoji	No Emoji	Emoji

On average, the participants selected more positive reaction words for stimuli with emojis than without emojis. The average number of words for stimuli without emojis is 1.98 positive words and 1.02 negative words. In contrast, stimuli with emojis received 2.11 positive words and 0.89 negative words on average.

The average values are based on the ratio between three reaction words that are either positive or negative and can be found in Table 5.

Table 5. The mean value (average amount of positive words based on the ratio between three reaction words that are either positive or negative), standard deviation, and variance for both populations.

	N	Mean	StDev	Variance
Emoji	10	2.1077	0.1952	0.0381
No emoji	10	1.9846	0.1568	0.0246

To test if these results are statistically significant, we first performed an Anderson-Darling test to check for normality. This test is defined as: H_0 : the data is not different from normal and H_A : the data is different from normal. With both populations' P-values above 0.05 (emoji 0.446, no emoji 0.579), we can not reject H_0 . We can therefore assume that both populations follow a normal distribution.

We already assume that our samples are independent and that both populations are assumed to follow a normal distribution. Therefore we can examine if both populations' mean values are equal with a heteroscedastic two-sample T-test. The hypotheses are defined as: $H_0 : \mu_{no\ emoji} = \mu_{emoji}$ and $H_A : \mu_{no\ emoji} \neq \mu_{emoji}$. With a two-tail distribution and a significance level of 0.05, we can not reject the P-value 0.16 from the T-test. This indicates that there is no statistical difference between the mean value of stimuli with emojis compared to those without emojis.

4.2 Follow-up Questions

The population that only saw emoji stimuli thought that search results with pictures were a positive addition. A few participants from the emoji population also answered that pictures could improve the search results when asked what would make it *fun* after answering *boring*. Some participants liked when the title started with an emoji, while others considered emojis disturbing regardless of location. Many participants reacted positively to two specific stimuli with emoji in the experiment. One of them was Figure 3 which is a stimulus with an autumn leaf. Many participants stated that the leaf reinforced the information and matched well with the topic of autumn outfits. Likewise, they thought that Figure 1 which contains an emoji of a pizza, also matched the other content of the stimulus in a very pleasant way.

The no-emoji population also thought pictures were an excellent addition to search results and thought more search results should have included them. Some participants thought that some stimuli felt more like advertising than important information about the search topic. A few from the no-emoji population also thought that some stimuli were too long and included too much information.

4.3 Post-experiment Question

The experiment included a post-experiment question on whether the participants believed emojis in search results were a good idea. The results show that seven out of 20 participants (35%) believed that emojis in search results was a good idea. Those who believed that emojis in search results was a good idea, thought emojis improved the content and gave it more context. Those who were positive towards emojis also said that emojis could make it easier to search through multiple search results because of the distinctive visual aspects that emojis acquire.

Meanwhile, those who did not believe that emojis in search results was a good idea, thought that emojis could be distracting and make the search result feel unprofessional. Other participants who also were negative towards emojis thought that emojis in conjunction with text would make search results messier and unnecessarily complex.

However, both sides emphasized that using emojis in search results would first and foremost need to be used in the right way for it to have a positive impact. Some participants gave examples of ways this could be done. For example, they thought it was important that emojis are used in a complementary way and that only a few emojis are used simultaneously.

5 Discussion

This study examines how user desirability is affected by emojis in search results. The experiment shows that the emoji version, on average receives 0.13 more positive words than no emoji. However, there is no significant statistical difference at $p \leq 0.05$. The result from the post-experiment question suggests that only 35% believe emojis in search results is a good idea.

The two reaction words that got the most percentage from both populations were *clear* and *helpful*. This result can most likely be attributed to the nature of search results. The word *helpful* is, by its very nature, something a search result is; it helps a user get to something they searched after. That is why *helpful* gained over a tenth of all reactions in both populations. On the other hand, the word *clear* might be based on the fact that when participants did not know what to answer, they answered *clear* because it is a neutral word that fits most search snippets. As long as the participant understood the text, they could say it was clear. There is, however, some considerable difference between the populations. The no-emoji population chose the word *clear* more often. It is possible that both populations focused on different things. The emoji population might have focused more on the visual aspects, and the no-emoji population on the actual meaning of the search result. This is in line with earlier research⁷ which found that people tend to focus more on the visual aspects when an emoji is present.

⁷ <https://www.nngroup.com/articles/emojis-email/>, Emojis in Email Subject Lines: Advantage or Impediment?, accessed 2023-01-03

Because of this, the no-emoji population likely chose the word *clear* more often than the emoji population.

Both populations' different focus is also probably the reason that some of the other reaction words have a significant difference between the populations. The words *Friendly* and *Annoying* had a higher occurrence in the no-emoji population, whereas *Appealing* and *Exciting* had a more significant occurrence in the emoji population. The words *appealing* and *exciting* are both reactions that are more connected to visual design aspects and may be coupled to the focus of the population as described earlier. The words *friendly* and *annoying* are not as easily coupled to visual aspects, which might have led the emoji population to choose other reaction words. On the other hand, the no-emoji population was probably fixated on the meaning of the search results. They might have found the content more like an advertisement (*annoying*) or that they generally felt welcomed by the text (*friendly*).

During the follow-up questions, participants from both populations pointed out that they thought pictures were an excellent addition. Therefore, emojis could be desirable since they share some of the same visual qualities as pictures. There is an interesting contrast between those who say that emojis make it easier to search through multiple search results and those who say that emojis make it more complex.

Lastly, some suggestions on how to implement emojis in search results in a valuable and desirable way. Including emojis in search results might lead users to focus more on the visual design than the content of the search result. Therefore it is essential to only use emojis in a complementary way that enriches the message. Some participants in this study liked when the emoji was placed at the beginning of the title. Others believed that only a few emojis should be used simultaneously.

5.1 Drawbacks and Limitations

Due to time constraints, the number of participants was only 20, and their sociodemographic spread is relatively small compared to the user population of search engines. We did not control for brand awareness, so it is possible that participants' earlier knowledge of the brands influenced their reactions. Some users used the previous stimulus to compare the next stimulus, which can have led to a specific pattern, especially since we did not change the order of the stimuli or the reaction words.

The time constraints limited our choices of data analysis. For example, we did not investigate if there was any measurable difference between age or if gender had any effect on the answers. However, the age distribution was too small to draw any general results, and the sample size for measuring gender differences would have been too small to measure statistically with an acceptable margin of error. The limited time also meant we could only perform some tests on campus. The other tests were conducted via Zoom, which could have negatively affected the result. That is because the non-oral input from the participants was unnoticed during those tests.

Even though we used as many different emojis as possible, it is possible that the choices are not general enough. The result can therefore be different for other emojis. This study only examined single search snippets. In a real-world environment, multiple search results could significantly affect desirability more than between two single search results.

The participants might have answered differently if they knew what they searched for, but that was not the primary goal of this study. There is also a possibility that participants did not recognize or misinterpret the emojis since they have different designs across different platforms.

5.2 Future Work

Based on the observed shortcomings of this study, we suggest a more thorough study with a larger sample size and more disposable time. This would make predictions about gender and age differences a more accessible task.

It would also be interesting to explore if search results with emojis create a higher click-through rate (CRT) than no emoji in a field of multiple search results. Another intriguing topic would be to use another factor than desirability to evaluate emoji effectiveness in search results. Lastly, an in-depth study of how other elements in search results affect desirability would be interesting.

6 Conclusion

Our study reveals that users do not necessarily desire emojis in search results. Only 35% of all participants believed that emojis in search results might be a good idea. On average, search results with emojis received 0.13 more positive reaction words than those without emojis. However, this result is not statistically significant, and a larger sample size is required to draw further conclusions.

References

- [1] Bai, Q., Dan, Q., Mu, Z., Yang, M.: A systematic review of emoji: Current research and future perspectives. *Frontiers in Psychology* **10** (2019)
- [2] Cavalheiro, B.P., Prada, M., Rodrigues, D.L., Garrido, M.V., Lopes, D.: With or without emoji? perceptions about emoji use in different brand-consumer communication contexts. *Human Behavior and Emerging Technologies* **2022** (2022)
- [3] Das, G., Wiener, H.J., Kareklas, I.: To emoji or not to emoji? examining the influence of emoji on consumer reactions to advertising. *Journal of Business Research* **96** (2019) 147–156
- [4] Cramer, H., De Juan, P., Tetreault, J.: Sender-intended functions of emojis in us messaging. In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. (2016) 504–509

- [5] Bich-Carrière, L.: Say it with [a smiling face with smiling eyes]: judicial use and legal challenges with emoji interpretation in canada. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique* **32**(2) (2019) 283–319
- [6] Castellacci, F., Tveito, V.: Internet use and well-being: A survey and a theoretical framework. *Research Policy* **47**(1) (2018) 308–325
- [7] Benedek, J., Miner, T.: Measuring desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals Association* **2003**(8-12) (2002) 57
- [8] Barnum, C.M., Palmer, L.A.: More than a feeling: understanding the desirability factor in user experience. In: *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery (2010) 4703–4716

Sigal: Instantiation for SCAN

Willeke Martens

Department of Computing Science
Umeå University, Sweden
`willeke.martens@umu.se`

Abstract. Björklund et al. (2022) introduce a mathematical formalism to uniformly define symbolic, neural and neuro-symbolic learning tasks that involve composite semantics. In this work, we provide an instantiation of the formalism for a modified version of the grounded navigation task SCAN – a diagnostic task used to investigate the capacity of a model for compositional generalization. Whilst the instantiation demonstrates the strength of the framework as a closed-loop system, it also raises practical questions in terms of suitable training strategies when incorporating neural components.

1 Introduction

Compositional generalization (sometimes also called combinatorial or systematic generalization) is the cognitive ability of humans to understand and generate novel combinations of known elements [1, 2, 3]. For example, given that a person grasps the meaning of “twice” and understands the command “jump”, the person should have no problem interpreting the command “jump twice”, despite never having encountered this specific command before. Compositional generalization is commonly considered to be a key skill in unlocking true human-like machine intelligence [1, 4]. However, given that the world is compositional, the capacity of a system for compositional generalization hinges on its ability to properly disentangle concepts acquired in a compositional setting [5]. For example, the colour red is never encountered in isolation, instead, the learner might observe, e.g., a *red* ball, a *red* table and a *red* dress. The ability of a learner to imagine a red bottle depends then on whether the learner successfully managed to disentangle the concept of red from the concept of ball, table and dress. In this work, we refer to the task of acquiring concepts and representations in compositional settings as *learning composite semantics*.

In [5], Björklund et al. introduce a mathematical formalism, we henceforth call *Sigal*¹, to uniformly define symbolic, neural and neuro-symbolic tasks that involve learning composite semantics from examples. *Sigal* specifies such learning tasks in terms of a *template algebra*, a set of examples, a learning goal and a family of admissible solutions. A template algebra \mathcal{A} is defined in relation to a

¹ The abbreviation comes from *simultaneous grounding and learning*, a feature that characterises framework and is further detailed in Section 3.2.

set of symbols Σ so that \mathcal{A} provides the semantics of the operators in $\Sigma' \subseteq \Sigma$. The learning goal of the task is to find an instance of \mathcal{A} that assigns meaning to the symbols in $\Sigma \setminus \Sigma'$ based on the set of examples. The search space for the candidate functions of the symbols in $\Sigma \setminus \Sigma'$ is limited to the family of admissible solutions, which are specified symbol-wise.

The main aim of this work is to discuss the instantiation of the Sigal framework for a modified version of SCAN, a grounded navigation task originally introduced by Lake and Baroni [6]. The task consists in translating unambiguous natural language commands into action commands. Thereby, we further expand on the number of examples given in [5].

The remainder of this paper is organized as follows. Section 2 discusses common approaches to learning composite semantics, and frames the SCAN task in the context of other diagnostic tasks for compositional generalization. Next, Section 3 introduces the theoretical backbone, focusing predominantly on the Sigal formalism. Section 4 presents the instantiation of the framework for SCAN. In Section 5, we conclude the paper with a summary and raise questions for future work.

2 Related Work

In this section, we briefly discuss different approaches to learning composite semantics and relate SCAN to other diagnostic tasks that investigate a model’s capacity for compositional generalization.

2.1 Approaches to Learning Composite Semantics

We can discern three distinct approaches to learning composite semantics, namely, neural, symbolic and neuro-symbolic procedures.

Neural procedures are especially popular in areas such as natural language processing [7, 8, 9] and computer vision [10]. Whilst neural models require less effort in terms of manually extracting features and rules, experiments indicate poor results for standard neural models in terms of compositional generalization [10, 6, 11]. Although substantial progress has been made in recent years [1, 12, 13], weak comprehensibility of black-box systems raises other concerns in terms of harmful biases [14] and Clever Hans behaviour [10, 15, 16].²

Symbolic learning frameworks are commonly praised for their data efficiency, comprehensiveness and natural way to incorporate background information [1]. The inductive logic programming (ILP) approach, realised by a wide range of systems, e.g., FOIL [18], XHAIL [19], Popper [20], expresses learning problems in terms of background knowledge \mathcal{B} , positive and negative examples, and a

² A machine learning algorithm is said to display Clever Hans behaviour if it exploits invalid solving strategies [10]. Lapuschkin et al. [17] demonstrated how Fisher vectors used watermarks in images for class prediction and ignored remaining content. Detecting problems like this requires detailed analysis, which might be practically impossible in the case of large datasets [10].

hypothesis space \mathcal{H} . The aim is to induce a hypothesis $h \in \mathcal{H}$ such that \mathcal{B} and h derive the positive but not the negative examples. Probabilistic ILP [21] extends the ILP framework to explicitly deal with statistical uncertainty. A limitation of such frameworks is that the background knowledge, examples and hypothesis must be given in the form of logic programs. Whilst this allows for highly structured representations, they are often too rigid to deal with noisy and numerical data [1].

Neuro-symbolic approaches [22, 23, 24, 25] integrate neural and symbolic computing elements to leverage the learning abilities of neural methods and the rich semantics and reasoning abilities of symbolic systems [26, 27]. Raedt et al. [26] for instance extended probabilistic ILP so that predicates can be realised through neural networks. Symbolic elements can further improve neural methods in terms of data efficiency, error recovery, explainability and bias mitigation [28].

Sigal is an abstraction to formalize learning tasks involving composite semantics and seeks to accommodate a wide range of neural, symbolic or neuro-symbolic approaches discussed above [5]. As such, Sigal is a useful tool to analyse task variations, as well as relevant conditions for learning composite semantics.

2.2 SCAN in Context

To investigate the capacity of a model in terms of compositional Generalization, a series of diagnostic datasets [10, 11, 29, 30, 31, 32, 33, 34, 35] have been developed in a wide range of areas. The diagnostic datasets are usually synthetic in nature, enabling controlled and systematic experiments over different splits of training and test sets.

For example, CLEVR [10] is a popular benchmark in the context of visual question answering (VQA) systems. VQA is the multi-modal task of answering natural language questions about images.³ CLEVR consists of pairs of the form (q, i) , where i is an image containing three-dimensional geometric objects with attributes such as size, color, material and shape. The question q is strategically harvested from a set of question families to avoid question-conditional biases. A template of such a question is given by

“How many $\langle c \rangle$ $\langle m \rangle$ are there?”

where $\langle c \rangle$ and $\langle m \rangle$ are substituted with a relevant color and material in image i . The diagnostic dataset probes a system’s ability to visually ground concepts such as color and material in images, whilst also evaluating higher level reasoning abilities such as counting.⁴

The dataset SCAN was developed by Lake and Baroni [6] with similar intentions in the context of grounded navigation. The set contains pairs of the

³ Multi-modality stands in this context for the inclusion of several modes of data, namely text and images.

⁴ Björklund et al. provide a rough sketch of a suitable instantiation of Sigal for CLEVR in [5].

form (c, a) , where c is a natural language statement consisting of primitive commands, modifiers and connecting words harvested from a limited vocabulary. Action sequence a is a sequence of actions corresponding to the correct translation of c . The aim of SCAN is to learn the correct semantic interpretation of the primitive commands, modifiers and connecting words by means of examples. A rigorous introduction of SCAN follows in Section 4.1.

3 Preliminaries

This section focuses on introducing the theoretic backbone of the Sigal framework.

3.1 Typed alphabets, Terms and Algebras

This section provides an overview of the essential syntactic and semantic components required to introduce the Sigal framework. The definitions are based on [5].

Note 1. To indicate that $i \in \{1, 2, \dots, k\}$ for some $k \in \mathbb{N}$ we write $i \in [k]$ for brevity.

Syntactic Components. The syntactic components of Sigal determine how symbols and operators can be combined within the framework. In other words, the components induce the set of well-formed terms and formulae of a particular instantiation of Sigal.

Definition 1 (Γ -Typed Alphabet Σ). *Let Γ be a finite set of types. The finite set $\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_n\}$ is a Γ -Typed Alphabet if each symbol $\sigma \in \Sigma$ is associated with a function signature s over Γ^+ .*

Let Σ be a Γ -typed alphabet such that $\sigma \in \Sigma$ and σ is associated with signature s given by

$$s = \gamma_1 \gamma_2 \dots \gamma_k \gamma,$$

for $\gamma_i, \gamma \in \Gamma$ and $i \in [k]$. This is equivalent to writing

$$\sigma: \gamma_1 \gamma_2 \dots \gamma_k \rightarrow \gamma,$$

and indicates that σ has arity k and evaluation type γ . Further, input argument i has type γ_i for $i \in [k]$. If σ has arity 0, i.e., is a constant, we write $\sigma: \gamma$.

Definition 2 (Leaf Alphabet Σ). *A Γ -typed alphabet Σ is a leaf alphabet if all symbols $\sigma \in \Sigma$ are associated with a signature $\sigma: \gamma$ for some $\gamma \in \Gamma$.*

Given a Γ -typed alphabet Σ , let $X_l = \{x_1, x_2, \dots, x_l\}$ for some $l \in \mathbb{N}$ be a leaf Γ -typed alphabet representing variables such that $\Sigma \cap X_l = \emptyset$. The set of terms over γ for some $\gamma \in \Gamma$, i.e., $T_{\Sigma \cup X_l}^\gamma$, is defined through simultaneous induction as follows:

1. $\sigma \in T_{\Sigma \cup X_l}^\gamma$ if $\sigma \in \Sigma \cup X_l$ and $\sigma: \gamma$,
2. $\sigma[t_1, t_2, \dots, t_k] \in T_{\Sigma \cup X_l}^\gamma$ if $\sigma \in \Sigma$, $\sigma: \gamma_1 \gamma_2 \dots \gamma_k \rightarrow \gamma$ and $t_i \in T_{\Sigma \cup X_l}^{\gamma_i}$ for $i \in [k]$.

Example 1. Let $\Gamma = \{\alpha, \beta\}$, $X_2 = \{x_1: \alpha, x_2: \alpha\}$ and

$$\Sigma = \{+: \alpha\alpha \rightarrow \alpha, -: \alpha\alpha \rightarrow \alpha, \leq: \alpha\alpha \rightarrow \beta\}.$$

All $\sigma \in \Sigma$ have arity 2 as indicated by their function signature. The set X_2 is a leaf Γ -typed alphabet that represents variables. The set of terms over $T_{\Sigma \cup X_2}^\beta$ contains elements of the form

$$\leq [t_1, t_2],$$

with $t_1, t_2 \in T_{\Sigma \cup X_2}^\alpha$. For example, the term

$$\leq [+ [x_1, - [x_2, x_1]], x_2].$$

is a member of $T_{\Sigma \cup X_2}^\beta$.

Semantic Components. The semantic components of Sigal determine how well-formed terms and formulae can be evaluated for a particular instantiation of the framework.

Definition 3 (Σ -Algebra \mathcal{A}). *Given a Γ -typed alphabet Σ , a Σ -algebra \mathcal{A} provides an interpretation of Σ and is denoted by a pair*

$$\mathcal{A} = \langle (\mathbb{A}_\gamma)_{\gamma \in \Gamma}, (\sigma_\mathcal{A})_{\sigma \in \Sigma} \rangle.$$

The algebra assigns a domain \mathbb{A}_γ to each type $\gamma \in \Gamma$ and associates each symbol $\sigma \in \Sigma$ with a meaning as follows:

1. *given $\sigma \in \Sigma$ and that $\sigma: \gamma$, then $\sigma_\mathcal{A} \in \mathbb{A}_\gamma$,*
2. *given $\sigma \in \Sigma$ and that $\sigma: \gamma_1 \gamma_2 \dots \gamma_k \rightarrow \gamma$, then $\sigma_\mathcal{A}$ is a function*

$$\sigma_\mathcal{A}: \mathbb{A}_{\gamma_1} \times \mathbb{A}_{\gamma_2} \times \dots \times \mathbb{A}_{\gamma_k} \rightarrow \mathbb{A}_\gamma.$$

Example 2. Let Γ , X_2 and Σ be defined as in Example 1. Now, let the Σ -algebra $\mathcal{A} = \langle (\mathbb{A}_\gamma)_{\gamma \in \Gamma}, (\sigma_\mathcal{A})_{\sigma \in \Sigma} \rangle$ be such that $\mathbb{A}_\alpha = \mathbb{Z}$ and $\mathbb{A}_\beta = \{T, F\}$. Further, \mathcal{A} defines $+_\mathcal{A}: \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}$ as standard integer addition and $-_\mathcal{A}: \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}$ as standard integer subtraction. The operation $\leq_\mathcal{A}: \mathbb{Z} \times \mathbb{Z} \rightarrow \{T, F\}$ returns F if the first integer is strictly larger than the second integer, and T otherwise.

Given a Γ -typed alphabet Σ , evaluation type $\gamma \in \Gamma$ and Σ -algebra \mathcal{A} , a term $t \in T_\Sigma^\gamma$ can be recursively evaluated with respect to \mathcal{A} as follows:

$$\text{val}_\mathcal{A}(t) = \begin{cases} \sigma_\mathcal{A} & \text{if } t = \sigma \text{ and } \sigma: \gamma, \\ \sigma_\mathcal{A}(\text{val}_\mathcal{A}(t_1), \text{val}_\mathcal{A}(t_2), \dots, \text{val}_\mathcal{A}(t_k)) & \text{if } t = \sigma[t_1, t_2, \dots, t_k], \end{cases}$$

with $t_i \in T_\Sigma^{\gamma_i}$ for $i \in [k]$, given that $t: \gamma_1 \gamma_2 \dots \gamma_k \rightarrow \gamma$.

To evaluate terms t that contain variables $x \in X_l$, a distinct function $\phi: X_l \rightarrow (\mathbb{A}_\gamma)_{\gamma \in \Gamma}$ is required, such that every $x \in X_l$ with $x: \gamma$ for some $\gamma \in \Gamma$ is mapped to an element in \mathbb{A}_γ .

Example 3. Let Γ , X_2 and Σ be defined as in Example 1 and Σ -algebra \mathcal{A} be defined as in Example 2. To evaluate the term

$$t = \leq [+ [x_1, -[x_2, x_1]], x_2],$$

we first define $\phi : X_2 \rightarrow \mathbb{Z}$ so that $\phi(x_1) = 8$ and $\phi(x_2) = 9$. Now,

$$\text{val}_{\mathcal{A}}^{\phi}(-[x_2, x_1]) = 9 - 8 = 1,$$

and

$$\text{val}_{\mathcal{A}}^{\phi}([x_1, -[x_2, x_1]]) = \text{val}_{\mathcal{A}}^{\phi}(x_1) + \text{val}_{\mathcal{A}}^{\phi}(-[x_2, x_1]) = 8 + 1 = 9.$$

Finally, since $\text{val}_{\mathcal{A}}^{\phi}(x_2) = 9$ is equal to $\text{val}_{\mathcal{A}}^{\phi}([x_1, -[x_2, x_1]])$, t evaluates to 1.

Definition 4 (Template Algebra \mathcal{A}). A template Σ -Algebra \mathcal{A} for some Γ -typed alphabet Σ , is a Σ -Algebra except that there may exist symbols $\sigma \in \Sigma$ such that $\sigma_{\mathcal{A}}$ is undefined. A Σ -Algebra \mathcal{A}' is an instance of \mathcal{A} if

$$\sigma_{\mathcal{A}} = \sigma_{\mathcal{A}'},$$

for all $\sigma \in \Sigma$ where $\sigma_{\mathcal{A}}$ is defined.

3.2 Sigal Framework

The Sigal framework formalises learning and grounding problems in terms of typed alphabets and template algebras previously defined in Section 3.1.

Consider a Γ -typed alphabet Σ , a template Σ -algebra \mathcal{A} and the leaf Γ -typed alphabet $X_l = \{x_1, x_2, \dots, x_l\}$, for some $l \in \mathbb{N}$, such that $\Sigma \cap X_l = \emptyset$. Let $\tau \in \Gamma$ be a designated linearly ordered *evaluation type* on which an associative and commutative summation operation $\oplus : \mathbb{A}_{\tau} \times \mathbb{A}_{\tau} \rightarrow \mathbb{A}_{\tau}$ is defined. Note that this operator can thus be extended in the canonical way to any nonempty set of arguments. Further, let $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ be a sample set where each sample s_i , for $i \in [n]$, is a pair

$$(t_i, O_i),$$

with $t_i \in T_{\Sigma \cup X_l}^{\tau}$ and $O_i \subseteq \bigcup_{\gamma \in \Gamma} \mathbb{A}_{\gamma}$.

The goal of the learning task is to determine the meaning of the symbols in $\Sigma' \subseteq \Sigma$ that are left undefined with respect to \mathcal{A} , i.e., the goal of the learning task is to determine an optimal instance \mathcal{A}' of \mathcal{A} with respect to a sample set \mathcal{S} . In particular, the learning goal is to find \mathcal{A}' that optimises the sum

$$\bigoplus_{s \in \mathcal{S}} \bigoplus_{\phi \in \Phi} \text{val}_{\mathcal{A}'}^{\phi}(t),$$

with $\text{opt} \in \{\min, \max\}$ and Φ being the set of all injective mappings $\phi : X_l \rightarrow O$ such that $\phi(x_i) \in \mathbb{A}_{\gamma_i}$ for $x_i \in X_l$ and $x_i : \gamma_i$ with $i \in [l]$. The problem of grounding, i.e., mapping the variables in X_l to elements in O , is an integral subtask of determining a suitable instance \mathcal{A}' within the framework.

To limit the search space for suitable instances \mathcal{A}' , the formalisation of a learning task also specifies a set \mathcal{F}_{σ} of candidate functions

$$\sigma' : \mathbb{A}_{\gamma_1} \times \mathbb{A}_{\gamma_2} \times \dots \times \mathbb{A}_{\gamma_k} \rightarrow \mathbb{A}_{\gamma},$$

for every undefined $\sigma : \gamma_1 \gamma_2 \dots \gamma_k \rightarrow \gamma$ in $\Sigma' \subseteq \Sigma$ in \mathcal{A} .

4 Formal Instantiation for SCAN

Lake and Baroni introduced the dataset SCAN [6], henceforth denoted by \mathcal{D} , to investigate the compositional generalization capabilities of models on a supervised sequence-to-sequence semantic parsing task. The aim of the learner is to translate unambiguous natural language commands into action sequences. In Section 4.1, we discuss the SCAN dataset in detail. Next, we modify the task to fit within the framework. In Section 4.3, we investigate how to formulate the optimisation goal in the context of the task. We conclude with an overview of the instantiation.

4.1 Description of SCAN Dataset

Dataset \mathcal{D} is given by

$$\mathcal{D} = \{(c_1, a_1), (c_2, a_2), \dots, (c_n, a_n)\},$$

with c_i an unambiguous natural language command and a_i a corresponding action sequence for $i \in [n]$. In other words, each SCAN command is assigned an unambiguous meaning. See Table 1 for a subset of command and action sequence pairs.

Table 1: Small subset of examples in the SCAN dataset.

Command	Action Sequence
jump twice and jump	[JUMP, JUMP, JUMP]
jump around right after walk left	[LTURN, WALK, RTURN, JUMP, RTURN, JUMP, RTURN, JUMP, RTURN, JUMP]
run opposite left and look thrice	[LTURN, LTURN, RUN, LOOK, LOOK, LOOK]

The natural language commands c_i for $i \in [n]$ are generated by a context-free grammar $G = (N, T, P, C)$, where N and T are the sets of non-terminal and terminal symbols respectively, P is the set of productions and $C \in N$ is the start symbol. The set of terminals T is given by,

$$T = \{\text{and, after, twice, thrice, opposite, around, left, right, lturn, rturn, walk, look, run, jump}\},$$

whilst the set of non-terminals N is

$$N = \{C, S, V, D, U\}.$$

The set of productions P is defined by

$$P = \{ \begin{array}{l} C \rightarrow S \text{ and } S, \quad C \rightarrow S \text{ after } S, \quad C \rightarrow S, \quad S \rightarrow V \text{ twice} \\ S \rightarrow V \text{ thrice, } S \rightarrow V, \quad V \rightarrow \text{opposite } D, \quad V \rightarrow \text{around } D \\ V \rightarrow D, \quad V \rightarrow U, \quad D \rightarrow U \text{ left, } \quad D \rightarrow U \text{ right} \\ D \rightarrow \text{lturn}, \quad D \rightarrow \text{rturn}, \quad U \rightarrow \text{walk}, \quad U \rightarrow \text{look} \\ U \rightarrow \text{run}, \quad U \rightarrow \text{jump} \end{array} \}.$$

Note that since the productions do not contain recursion, the language $L(G)$ induced by G is finite. The set of commands is the language generated by G , i.e.,

$$L(G) = \{c_1, c_2, \dots, c_n\}.$$

Further, the semantics of the natural language commands $c \in L(G)$ is induced by the interpretation \mathcal{I} over the domain of all action sequences over

$$A = \{\text{LTURN}, \text{RTURN}, \text{LOOK}, \text{JUMP}, \text{WALK}, \text{RUN}\}.$$

For example, commands of the form

$$u_1 \text{ and } u_2,$$

with $u_1, u_2 \in L(G)$, are evaluated to

$$val_{\mathcal{I}}(u_1 \text{ and } u_2) = val_{\mathcal{I}}(u_1), val_{\mathcal{I}}(u_2). \quad (1)$$

The aim of the learner is to determine the evaluation function $val_{\mathcal{I}}(\cdot)$ on the basis of a subset \mathcal{S} of \mathcal{D} , such that given a natural language command $c \in L(G)$, $val_{\mathcal{I}}(\cdot)$ produces the corresponding action sequence a .

4.2 Modification of SCAN

In order to instantiate Sigal on SCAN, we have to make some minor modifications. We illustrate how we can redefine \mathcal{D} in terms of composite functions instead of natural language expressions and we adjust the overall aim of the task accordingly.

Instead of operating directly on natural language commands over T , we express a command $c \in L(G)$ as a composite function φ over the Γ -typed alphabet Σ' , where $\Gamma = \{\gamma\}$ and Σ can be constructed from T by associating a suitable signature s with each $\sigma \in T$. For example, suppose that the command c corresponds to the natural language statement “jump around right after walk left”, then the corresponding composite function φ is given by

$$\varphi = \text{after} [\text{around} [\text{right} [\text{jump}]], \text{left} [\text{walk}]].$$

We define Σ on the basis of T as follows. Let

$$\Sigma = \{\text{and}_b, \text{after}_b, \text{twice}_a, \text{thrice}_a, \text{opposite}_a, \text{around}_a, \text{left}_a, \text{right}_a, \text{lturn}, \text{rturn}, \text{walk}, \text{look}, \text{run}, \text{jump}\},$$

so that all $\sigma \in \Sigma$ without subscript are associated with signature $\sigma: \gamma$, all symbols $\sigma \in \Sigma$ with subscript a with signature $\sigma: \gamma \rightarrow \gamma$ and all $\sigma \in \Sigma$ with subscript b with signature $\sigma: \gamma\gamma \rightarrow \gamma$.

To generate the corresponding set of composite functions to the set of natural language commands in $L(G)$, define the grammar G' as

$$G' = (N', \Sigma, P', C),$$

where Σ is defined according to above and N' is the Γ -typed leaf alphabet constructed from N . The set of productions P' is derived from P as follows. Assume that $\alpha_i \in N$ corresponds to $\alpha'_i \in N'$. Similarly, assume that $\text{op} \in T$ corresponds to $\text{op}' \in \Sigma$. Then, if the production rule $p \in P$ is a terminal rule of the form

$$\alpha_i \rightarrow \text{op},$$

add

$$\alpha'_i \rightarrow \text{op}'$$

to P' . For a production rule $p \in P$ of the form

$$\alpha_i \rightarrow \alpha_j \text{ op } \alpha_k,$$

we add

$$\alpha'_i \rightarrow \text{op}'[\alpha'_j, \alpha'_k].$$

Finally, for a production rule $p \in P$ of the form

$$\alpha_i \rightarrow \alpha_j \text{ op},$$

or

$$\alpha_i \rightarrow \text{op } \alpha_j,$$

we add

$$\alpha_i \rightarrow \text{op}'[\alpha'_j].$$

The grammar G' induces the set of composite functions $L(G')$ corresponding to the previously stated set of natural language commands $L(G)$. In particular, we can find the corresponding composite function ϕ for a specific natural language command c , by conducting the following procedure. Let $c \in L(G)$ be given by the derivation

$$C \xrightarrow{p_1}_G t_1 \xrightarrow{p_2}_G t_2 \xrightarrow{p_3}_G \dots \xrightarrow{p_n}_G t_n = c,$$

where $t_i \in T_{\Sigma \cup N}$ and $p_i \in P$ denote the intermediate term and production used at time step i for $i \in n$. To find the corresponding composite function φ , we start with our start symbol in $C \in N'$ and apply the corresponding production $p'_i \in P'$ to $p_i \in P$ to intermediate term t_i for $i \in [n]$. The process yields

$$C \xrightarrow{p'_1}_{G'} t_1 \xrightarrow{p'_2}_{G'} t_2 \xrightarrow{p'_3}_{G'} \dots \xrightarrow{p'_n}_{G'} t_n = \varphi,$$

with $t_i \in T_{T \cup N'}$ and $p'_i \in P'$.

The modified aim of the learner is to determine a Σ -algebra \mathcal{A}^* on a basis of a subset \mathcal{S} of \mathcal{D}' with

$$\mathcal{D}' = \{(\varphi_1, a_1), (\varphi_2, a_2), \dots, (\varphi_n, a_n)\},$$

with φ_i corresponding to c_i for $(c_i, a_i) \in \mathcal{D}$, so that $a_i = \text{val}_{\mathcal{A}^*}(\varphi_i)$ for $i \in [n]$.

4.3 Intuition for Defining the Optimisation Goal

In Section 4.2, we transformed the natural language based dataset \mathcal{D} to accommodate composite functions. Here, we discuss how the semantics of the composite functions can be retrieved by introducing suitable distance measures.

Consider the template Σ -algebra \mathcal{A} given by

$$\mathcal{A} = \langle \mathbb{A}_\gamma, (\sigma_{\mathcal{A}})_{\sigma \in \Sigma} \rangle,$$

so that \mathbb{A}_γ represents the domain of action sequences over

$$A = \{\text{LTURN}, \text{RTURN}, \text{LOOK}, \text{JUMP}, \text{WALK}, \text{RUN}\}.$$

An instance \mathcal{A}' of Σ -algebra \mathcal{A} , defines a mapping to an action sequence in \mathbb{A}_γ for every undefined $\sigma_{\mathcal{A}}$ in \mathcal{A} . In other words, for an undefined $\sigma_{\mathcal{A}}$ in \mathcal{A} for some $\sigma: \gamma$ in Σ , \mathcal{A}' maps to a constant action sequence $a \in \mathbb{A}_\gamma$. For an undefined $\sigma_{\mathcal{A}}$ in \mathcal{A} for some $\sigma: \gamma^+ \rightarrow \gamma$ in Σ , \mathcal{A}' maps to an action sequence $a \in \mathbb{A}_\gamma$ by manipulating the action sequences given at input. For example, a correctly learned interpretation of and according to Equation 1 would be

$$\text{and}_{\mathcal{A}'}[a_1, a_2] = a_1 \circ a_2,$$

where \circ indicates the concatenation of the two action sequences a_1 and a_2 .

Since the aim is to find an optimal instance \mathcal{A}^* of \mathcal{A} , so that for any example (φ, a) in dataset $\mathcal{S} \subseteq \mathcal{D}'$ holds that

$$\text{val}_{\mathcal{A}^*}(\varphi) = a,$$

we must measure how good of job a particular instance \mathcal{A}' of \mathcal{A} does. In other words, we need to define a distance measure $\delta: \mathbb{A}_\gamma \times \mathbb{A}_\gamma \rightarrow \mathbb{N}$ that allows the learner to evaluate the quality of the predicted sequence $\text{val}_{\mathcal{A}'}(\varphi)$ for an example $(\varphi, a) \in \mathcal{S}$. Standard measures for the distance between sequences or strings are different types of edit distances, e.g., Levenshtein's distance [36].

We follow similar a similar strategy as Björklund et al. [5] in the context of learning picture languages and suggest transforming examples (φ, a) in dataset $\mathcal{S} \subseteq \mathcal{D}'$ into

$$(\delta[\varphi, x], a),$$

for some distance measure $\delta: \gamma\gamma \rightarrow r$, and variable $x: \gamma$ to stay within the framework.

4.4 Instantiation Overview

Let $\Gamma' = \Gamma \cup \{r\}$. Further, extend the typed alphabet Σ with the function symbol $\delta: \gamma\gamma \rightarrow r$. Now, consider the template Σ -algebra \mathcal{A} given by

$$\mathcal{A} = \langle (\mathbb{A})_{\gamma \in \Gamma'}, (\sigma_{\mathcal{A}})_{\sigma \in \Sigma} \rangle,$$

so that \mathbb{A}_{γ} represents the domain of action sequences over

$$A = \{\text{LTURN}, \text{RTURN}, \text{LOOK}, \text{JUMP}, \text{WALK}, \text{RUN}\},$$

and $\mathbb{A}_r = \mathbb{N}$. The learning goal of the task is to determine an instance \mathcal{A}' of Σ -algebra \mathcal{A} , that minimizes the average sum

$$\oplus \mathcal{S} = \bigoplus_{(t,a) \in \mathcal{S}} \text{val}_{\mathcal{A}'}^{\phi}(t),$$

with $t \in T_{\Sigma \cup X}^r$ and $\phi(x) = a$.

5 Conclusion and Future Work

In this work, we presented a theoretical instantiation of the new framework Sigal [5] for the modified task of sequence-to-sequence semantic parsing on SCAN. Whilst the instantiation is suitable for a symbolic approach, it remains unclear how Sigal could accommodate effective training of neural components. In particular, further investigations have to be conducted to assess the viability of training neural components in parallel, given that the loss for a training example is only known on term-level and not its constituents. A practical implementation is a natural continuation of this work.

Acknowledgements

I wish to thank Johanna Björklund, Frank Drewes and Adam Lindström for their interesting discussions and feedback.

References

- [1] Smolensky, P., McCoy, R., Fernandez, R., Goldrick, M., Gao, J.: Neuro-compositional computing: From the central paradox of cognition to a new generation of ai systems. *AI Magazine* **43**(3) (Sep. 2022) 308–322
- [2] Montero, M.L., Ludwig, C.J., Costa, R.P., Malhotra, G., Bowers, J.: The role of disentanglement in generalisation. In: *International Conference on Learning Representations*. (2020)
- [3] Bahdanau, D., Murty, S., Noukhovitch, M., Nguyen, T.H., de Vries, H., Courville, A.C.: Systematic generalization: What is required and can it be learned? *ArXiv abs/1811.12889* (2018)

- [4] Mikolov, T., Joulin, A., Baroni, M.: A roadmap towards machine intelligence. In: *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer (2016) 29–61
- [5] Björklund, J., Lindström, A.D., Drewes, F.: An algebraic approach to learning and grounding. *arXiv preprint arXiv:2204.02813* (2022)
- [6] Lake, B., Baroni, M.: Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Dy, J., Krause, A., eds.: *Proceedings of the 35th International Conference on Machine Learning*. Volume 80 of *Proceedings of Machine Learning Research.*, PMLR (10–15 Jul 2018) 2873–2882
- [7] Kim, N., Linzen, T.: COGS: A compositional generalization challenge based on semantic interpretation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Association for Computational Linguistics (November 2020) 9087–9105
- [8] Lake, B.M.: Compositional generalization through meta sequence-to-sequence learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., eds.: *Advances in Neural Information Processing Systems*. Volume 32., Curran Associates, Inc. (2019)
- [9] Baroni, M.: Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B* **375**(1791) (2020) 20190307
- [10] Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2017) 2901–2910
- [11] Bahdanau, D., de Vries, H., O'Donnell, T.J., Murty, S., Beaudoin, P., Bengio, Y., Courville, A.: Closure: Assessing systematic generalization of clevr models. *arXiv preprint arXiv:1912.05783* (2019)
- [12] Russin, J., Jo, J., O'Reilly, R.C., Bengio, Y.: Compositional generalization in a deep seq2seq model by separating syntax and semantics. *ArXiv abs/1904.09708* (2019)
- [13] Atzmon, Y., Kreuk, F., Shalit, U., Chechik, G.: A causal view of compositional zero-shot recognition. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., eds.: *Advances in Neural Information Processing Systems*. Volume 33., Curran Associates, Inc. (2020) 1462–1473
- [14] Bender, E.M., Koller, A.: Climbing towards NLU: On meaning, form, and understanding in the age of data. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Association for Computational Linguistics (2020) 5185–5198
- [15] Agrawal, A., Batra, D., Parikh, D.: Analyzing the behavior of visual question answering models. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Association for Computational Linguistics (November 2016) 1955–1960
- [16] Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., Parikh, D.: Yin and yang: Balancing and answering binary visual questions. In: *2016 IEEE*

- Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 5014–5022
- [17] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. *Nature communications* **10**(1) (2019) 1–8
 - [18] Quinlan, J.R.: Learning logical definitions from relations. *Machine learning* **5**(3) (1990) 239–266
 - [19] Ray, O.: Nonmonotonic abductive inductive learning. *Journal of Applied Logic* **7**(3) (2009) 329–340 Special Issue: Abduction and Induction in Artificial Intelligence.
 - [20] Cropper, A., Morel, R.: Learning programs by learning from failures. *Machine Learning* **110**(4) (2021) 801–856
 - [21] Raedt, L.D., Kersting, K.: Probabilistic inductive logic programming. In: *Probabilistic inductive logic programming*. Springer (2008) 1–27
 - [22] Cambria, E., Liu, Q., Decherchi, S., Xing, F., Kwok, K.: SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, European Language Resources Association* (June 2022) 3829–3839
 - [23] Khan, M.J., Curry, E.: Neuro-symbolic visual reasoning for multimedia event processing: Overview, prospects and challenges. In: *CIKM (Workshops)*. (2020)
 - [24] Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., Wu, J.: The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584* (2019)
 - [25] Ding, D., Hill, F., Santoro, A., Reynolds, M., Botvinick, M.: Attention over learned object embeddings enables complex visual reasoning. *Advances in neural information processing systems* **34** (2021) 9112–9124
 - [26] De Raedt, L., Manhaeve, R., Dumancic, S., Demeester, T., Kimmig, A.: Neuro-symbolic= neural+ logical+ probabilistic. In: *NeSy’19@ IJCAI, the 14th International Workshop on Neural-Symbolic Learning and Reasoning*. (2019)
 - [27] Garcez, A.d., Bader, S., Bowman, H., Lamb, L.C., de Penning, L., Illumino, B., Poon, H., Gerson Zaverucha, C.: Neural-symbolic learning and reasoning: A survey and interpretation. *Neuro-Symbolic Artificial Intelligence: The State of the Art* **342** (2022) 1
 - [28] Sarker, M.K., Zhou, L., Eberhart, A., Hitzler, P.: Neuro-symbolic artificial intelligence: Current trends. *arXiv preprint arXiv:2105.05330* (2021)
 - [29] Liu, R., Liu, C., Bai, Y., Yuille, A.L.: Clevr-ref+: Diagnosing visual reasoning with referring expressions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2019) 4185–4194
 - [30] Pezzelle, S., Fernández, R.: Is the red square big? MAlViC: Modeling adjectives leveraging visual contexts. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Association for Computational Linguistics* (November 2019) 2865–2876

- [31] Kottur, S., Moura, J.M.F., Parikh, D., Batra, D., Rohrbach, M.: Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In: North American Chapter of the Association for Computational Linguistics. (2019)
- [32] Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning. In: International Conference on Learning Representations. (2020)
- [33] Girdhar, R., Ramanan, D.: Cater: A diagnostic dataset for compositional actions & temporal reasoning. In: International Conference on Learning Representations. (2020)
- [34] Sampat, S.K., Kumar, A., Yang, Y., Baral, C.: CLEVR_HYP: A challenge dataset and baselines for visual question answering with hypothetical actions over images. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, Association for Computational Linguistics (June 2021) 3692–3709
- [35] Li, Z., Søgaard, A.: QLEVR: A diagnostic dataset for quantificational language and elementary visual reasoning. In: Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, United States, Association for Computational Linguistics (July 2022) 980–996
- [36] Levenshtein, V.I., et al.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. Volume 10., Soviet Union (1966) 707–710

Reducing Gender Biases with Semi-Supervised Topic Modelling

Salome Müller

Department of Computing Science
Umeå University, Sweden
`mcs21smr@cs.umu.se`

Abstract. Many models in Natural Language Processing contain gender biases. In this work we investigate biases inherent in the data used to train such models, and seek to remove them. We use Latent Dirichlet Allocation Topic Modelling with in-set topic knowledge to visualize the gender biases. We find that the investigated Mainstream English corpus associates men with crime and women with family and communication. The model further allows us to identify sentences that feed the most into the biases. In our case study we find that for the given corpus removing 5% of these sentences suffices to remove gender biases against men and that 15% have to be removed to remove the biases against women. Thus, even though this method is easily applicable, it is not generally suited to debias a corpus as a significant fraction of data has to be removed.

1 Introduction

It has been shown repeatedly how gender biases in data can vastly influence Machine Learning algorithms trained on it [1],[2]. For example, Fitria shows how gender biases influence Google Translate [1]. When translating from Indonesian, a language without gendered pronouns, into English, the system assigns ‘he’ pronouns to doctors and ‘she’ pronouns to nurses [1]. This points to biases against people of different genders, which connect doctors to men and nurses to women, thus representing and reproducing typical stereotypes.

There have been multiple attempts in visualizing and reducing gender biases [3],[4],[5],[2]. This work attempts to reduce these biases by debiasing the input data using Latent Dirichlet Allocation Topic Modelling (LDA) [6]. LDA is a generative model which finds latent topics in a corpus as probabilities over words. More explicitly, we use LDA with seed words [7], which allows us to give the model prior knowledge. We seed two topics with gendered words to retrieve two explicitly gendered topics from the corpus.

Analyzing gendered topics retrieved for a corpus allows us to pinpoint the gender biases inherent in the corpus. Additionally, these topics allow us to identify sentences that influence these biases the most. This work is a case study that tests whether removing a certain amount of the identified sentences from the English Mainstream corpus [8] leads to a corpus with less gender biases than the initial corpus, i.e., to a corpus that does not contain the pinpointed biases.

By removing different fractions of these sentences, we find that for the given corpus, removing 5% of these sentences generates a corpus with significantly fewer biases against men. Further, to reduce the gender biases against both men and women, 15% of the sentences have to be removed. This is a considerable fraction of data, and we assume that a lot of information about family and crime are removed from the corpus as well. Additionally, it only removes the biases we pinpointed. We make not investigation in how removing the sentences influences other inherent biases, that were not detected by us.

1.1 Theoretical Grounding

Following Devinney et al., we do not mean characteristics of ‘women’ and ‘men’ tied to their biological sex when referring to gender, but a social and cultural construct [9],[8]. We see gender as constructed through acts that are repeated over time and that produce our understanding of gender characteristics [9]. In the context of this work, we divide gender into two categories, *feminine* and *masculine*. Since people of other genders, e.g., non-binary or genderless people, are underrepresented in the studied corpus [8], we only analyze the binary genders.

We use the term bias to refer to associations that are gendered in the data, but which we do not want to be gendered. We say “bias against someone”, because we see such associations as a prejudice against someone based only on their gender. Hellström et al. distinguish between different types of biases in Machine Learning [10]. We use the term *Gender Bias* to refer to a subcategory of what Hellström et al. call *historical bias*. They define *historical bias* as bias that is or was embedded in the world, opposed to bias originating from data generation and learning processes. Historically biased data represents the current or past state of the world, but includes aspects that do not correspond to how we want to model the world. An example for this could be some corpus, e.g., a collection of news articles, in which the word ‘doctor’ co-occurs more often with ‘man’ while the word ‘nurse’ co-occurs more often with ‘woman’. These co-occurrences suggest that doctors are men, while nurses are women. It is historical bias, as it represents stereotypes that appear in the world. Further, these stereotypes are indeed a bias, because we want to model a world where both ‘doctor’ and ‘nurse’ are equally likely to refer to women, men and people of other genders.

This work studies gendered words including pronouns like ‘he’ and ‘she’, as well as other gendered words, e.g., ‘mother’ and ‘father’. These words and the words associated with them contribute to reproducing ideas of femininity and masculinity, but do not necessarily reflect all people’s experiences. We use the terms ‘woman’ and ‘man’ to speak about people that are referred to with the studied gendered words. Our goal is to produce a corpus with less gender bias than the initial corpus. In such a corpus, the context in which the gendered words appear would be similar across different genders¹, e.g., feminine gendered

¹ This work does not address whether this is applicable over all domains, e.g., medical journals where the biological sex of the mentioned people produces important differences between individuals, which should not be removed.

words would appear close to the word ‘doctor’ as often as masculine gendered words.

In Section 2, we present earlier work relevant to our study. The methods and data that are used are presented in Section 3. In Section 4, we show and analyze the results of our experiments.

2 Earlier Work

There were multiple recent attempts to make gender biases in text corpora visible [3],[4],[8],[5]. For example, Hoyle et al. extract adjectives and verbs from a corpus for pairs of nouns (e.g., (‘girl’, ‘boy’)) [5]. The authors focus on how the adjectives used to describe people in a positive and negative way differ between women and men. They found that positive adjectives for women relate to their body more often than for men. Other attempts [3],[4],[8] use Latent Dirichlet Allocation Topic Modelling [6]. This model classifies the documents of a corpus as a mixture of topics, showing which words are likely to co-occur. Both in [8] and [4], the same three corpora (one in Swedish and two in English, including the Mainstream English corpus), are used to train the model, and differences of gender representation between the different corpora are studied. The findings of both papers show that women are more often related to social media and family, while the words associated with men suggest that men can be associated with a wider range of topics. They also find that Topic Modelling is suited well to detect and visualize biases.

The fact that biased data leads to biased models has been shown previously [11]. One example is how Bolukbasi et al. show how biases in a text corpus manifest in models by studying Word2vec word embeddings [2]. The Word2vec model is trained on word co-occurrence in text corpora [12]. It represents each word of the corpus as a vector, such that vectors of words with similar semantic meaning are close together. These word embeddings are found to be good for solving analogy puzzles of the form “*Paris is to France, as Tokyo is to x*” with simple vector arithmetic [13]. For the given example, the best answer would be ‘*Japan*’². By examining 300-dimensional embeddings trained on a corpus of Google News texts consisting of 3 million English words and terms, Bolukbasi et al. found that the embeddings pinpoint sexism contained in the corpus [2]. For the question “*man is to woman as computer programmer is to x*”, ‘*homemaker*’ is retrieved [2]. Such analogies can have harmful consequences [11], as a model which is trained on this data is likely to make biased inferences. For example, if such a model is used to sort applicants for a job in computer science, it may rank women lower than men, as it assumes them less suitable for the job, when the only difference is the applicant’s gender [14].

One might argue that keeping certain gender biases in the data is useful to capture statistics. While this might be desired in some cases, this is not generally true, as it has been shown that algorithms can amplify biases [11]. If

² In practice, this does not always work out as reliably as advertised. In fact, the accuracy lies below 40% [12].

an algorithm is trained on biased data, its outcomes will most likely be biased as well. These outcomes then influence users, which in turn can produce more biased data. Using this newly generated data to train a model may result in a loop between biases in data, algorithms and user interaction. One example for this are web search engines [15]. They often use popularity scores to sort results and display the most popular results at the top. Users are more likely to choose the results displayed at the top. Thus, the popularity score of the results presented at the top are increased further, not because of their content but due to their positioning. In order to break this loop and to reduce gender biases in Machine Learning models, the amount of biases in the model should be reduced as far as possible.

According to Mehrabi et al., recent advances of debiasing in Machine Learning include generating labels for datasets, similar to nutrition scores on food, and detailed sheets pointing out the datasets’ creating methods, characteristics and skews [11]. These methods are good to raise awareness of biases, but do not supply us with unbiased data. There have also been multiple advances to change data mining processes, in order to not create biased data [11]. Other work reduces gender biases in intermediate steps of Machine Learning procedures [2]. For example, Bolukbasi et al. reduce some dimensions of Word2vec word embeddings in order to reduce the gender biases in the trained model [2]. This is valuable work, as many NLP models take word embeddings as input. While these advances are all valuable contributions to debiasing Machine Learning models, we think that more research should go into debiasing the data directly, before using it to train any model. This work provides research in this area, by studying a method which was proposed by Devinney et al. [8].

3 Method

In Section 3.1, the used data corpus is introduced and the performed preprocessing is explained. Then, in Section 3.2, the mathematical background for the unsupervised LDA Topic Model is given, and we show how the model can be transformed into a semi-supervised model by using seed words. In Section 3.3, the used debiasing pipeline is shown, as well as how the pipeline was implemented. Finally, the experimental setup and the qualitative analysis we perform on the retrieved topics are explained in Section 3.4.

3.1 Data

We use the English Mainstream corpus from [8] and [4], which is provided by Devinney et al. It contains 100,000 news articles, gathered from news websites in 2019, containing over 2.5 million sentences. The data is only annotated with the website the article was scraped from, and we have no information about the demographics of the authors. Further, we are not aware of any biases apart from the analyzed gender biases.

Preprocessing The corpus’ documents are split into sentences with the NLTK sentence tokenizer. Each sentence is POS tagged with the POS tag functionality of NLTK before stopwords and words with tags irrelevant to our task are removed. The stopword list is a modified version of the NLTK stopword list that keeps third-person pronouns. The POS tags are then used to determine the most appropriate lemmas of words, and to disambiguate between different meanings of words with the same lemma. For example, ‘girl’ (NN) and ‘girls’ (NNS) will in the end map to the same lemma, ‘girl’ (NN), while ‘girly’ (ADJ) will be mapped to ‘girl’ (ADJ). We use the WordNet lemmatizer provided by NLTK. Finally, non-seed words that appear less than three times in the corpus are removed.

3.2 Topic Modelling

Latent Topic Models are a family of models that retrieve latent topics from unlabelled data. Each topic is a distribution over words. Table 1 shows a simplified example for topics that could be retrieved from a corpus of news articles. When trained on a corpus, Latent Topic Models find topics that maximize the probability to generate the given corpus. They find the topics as distributions over words and the distribution over the different topics. One model in this family is Latent Dirichlet Allocation Topic Modelling (LDA) [6], an unsupervised, fully generative probabilistic model.

Table 1. Example of three topics and corresponding words that could be retrieved from a corpus consisting of articles about *Arts*, *Law* and *Sports*.

Arts	Law	Sports
film	case	game
show	court	season
event	law	team
photo	lawyer	goal

We define the LDA model following [6], but extend the definition to make it more accessible. We split the corpus into documents on sentence level. Thus, every document consists of exactly one sentence, and we use the terms ‘document’ and ‘sentence’ interchangeably. Further, Topic Modelling uses bags-of-words, which makes the order of words in a sentence irrelevant.

Let \mathcal{D} be a corpus consisting of M documents. The number of underlying topics in \mathcal{D} is T and z_t refers to topic t for $t \in \{1, \dots, T\}$. LDA takes four hyperparameters, namely M , T , α and β , where the scalars α and β are both used to parameterize distributions. In order to generate a corpus, M documents are generated. For each document, a Dirichlet random variable θ is sampled from a Dirichlet distribution parameterized with α :

$$\theta \sim \text{Dir}(\alpha). \quad (1)$$

Let s be one of the documents in \mathcal{D} , and θ_s the respective Dirichlet random variable. The length of s , denoted by N , is sampled with a Poisson distribution.

Let $w_n \in s$ for $n \in \{1, \dots, N\}$ be a word in s . The topic z_t of w_n is sampled with a multinomial distribution parameterized with θ_d , and w_n is sampled as a multinomial probability conditioned on z_t and β :

$$z_t \sim \text{Multinomial}(\theta_s), \quad (2)$$

$$w_n \sim p(w_n|z_t, \beta). \quad (3)$$

The Topic Modelling training algorithm we use starts by randomly assigning $p(w|z_t)$ for every word and topic combination. It then iteratively updates these probabilities, in order to maximize the probability of generating the given corpus. During this process, the model learns the probabilities of topics, $p(z)$, as shown in Equation (2), and of words given a topic, $p(w|z)$, as shown in Equation (3). After training, the model further holds $p(z_t|w)$ for $t \in \{1, \dots, T\}$, the probability for z_t given any word. Since the initial probabilities $p(w|z_t)$ are assigned randomly, Topic Modelling is non-deterministic.

Seed Words The unsupervised LDA model starts without any knowledge of the underlying topics. To transform the model into a semi-supervised model, seed words are added and provided with a weight, which is called z -label [7]. This weight influences the initial probability assigned for the seed word for the respective topic. During training, the seeded topics are retrieved from the text by finding other words that appear in the same contexts as the seed words [7]. By seeding some or all topics, supervised information is added, as some topics are assigned already before training [7].

In this work, seed words are used to guide the model towards the discovery of secondary patterns in the data, i.e., to find topics that represent what words are associated with different genders. The seed words we use are the same as Devinney et al. use in [4] and [8]. They are shown in Table 2.

Table 2. Seed words for the feminine, masculine and non-binary topics, as chosen in [8].

Feminine	Masculine	Non-binary
she	he	they
woman	man	person
girl	boy	child
lady	guy	
female	male	neutral
		nonbinary
feminine	masculine	genderqueer
		enby
Miss		
Ms	Mr	Mx
Mrs		
Madam	Sir	

As Devinney et al. show in [8], people of non-binary genders are underrepresented in the English Mainstream corpus. Therefore, we focus on the binary

genders only. Nevertheless, we seed the third non-binary topic, in order to have comparable results to the findings of [8] and [4], but refrain from analyzing it.

3.3 Debiasing Pipeline

In order to debias the corpus, the sentences which contribute to the bias the most are identified and removed from the corpus. A new Topic Model is trained on the new corpus, and the differences between the new feminine and masculine topic are analyzed and compared to the initial topics.

A corpus from which $x\%$ of sentences have been removed is denoted by $\mathcal{D}^{(x)}$. The model trained on the respective corpus is denoted by $\mathcal{M}^{(x)}$, and the retrieved topics by $z_t^{(x)}$, for $t \in \{1, \dots, T\}$. We let $z_0^{(x)}$ refer to the feminine topic and $z_1^{(x)}$ to the masculine topic.

The first step is to train the initial model $\mathcal{M}^{(0)}$ on the original corpus $\mathcal{D}^{(0)}$. Analyzing the resulting explicitly gendered topics $z_0^{(0)}$ and $z_1^{(0)}$ shows the gender bias contained in the original corpus. Using these two topics, we identify the set of sentences that are most likely for these topics. Let s be a sentence of length N , and w_n the words in s for $n \in \{1, \dots, N\}$. Let further $L_t(s)$ be the log-probability of s to be produced by topic $z_t^{(0)}$:

$$L_t(s) = \log p(s|z_t^{(0)}) = \sum_{n=1}^N \log p(w_n|z_t^{(0)}). \quad (4)$$

Finally, let S_t be the set of pairs of sentences and their log-probability for $z_t^{(0)}$:

$$S_t = \{(s_i, L_t(s_i)) | s_i \in \mathcal{D}^{(0)}\}. \quad (5)$$

Then, we generate an initially empty set $S^{(x)}$. We alternately choose the sentence s from S_0 or S_1 with highest log-probability in the respective set. We remove s from S_0 or S_1 , respectively. If the sentence is not contained in $S^{(x)}$, it is added to the set. We proceed this way, until the set holds exactly $x\%$ of the sentences of $\mathcal{D}^{(0)}$. Then, $S^{(x)}$ consists of the sentences which presumably influence the masculine and feminine topic most and which are removed from $\mathcal{D}^{(0)}$ to form $\mathcal{D}^{(x)}$,

$$\mathcal{D}^{(x)} = \mathcal{D}^{(0)} \setminus S^{(x)}. \quad (6)$$

It is necessary to build $S^{(x)}$ in this manner. A sentence s may have a high probability for both z_0 and z_1 . However, we want $S^{(x)}$ to be a unique set, so that the amount of sentences that are removed from $\mathcal{D}^{(0)}$ is exactly as high as advertised. Further, we want to remove the same amount of sentences for each topic.

Finally, we train a new model $\mathcal{M}^{(x)}$ on $\mathcal{D}^{(x)}$ and analyze the differences between $z_0^{(x)}$ and $z_1^{(x)}$.

Implementation The described debiasing pipeline is implemented in Python³. It is embedded into a codebase provided by Devinney et al., called EQUITBL [16], which is used in [4] and [8]. EQUITBL is written in Python and uses Parallel Semi-Supervised Latent Dirichlet Allocation (pSSLDA)⁴ to train a Topic Model on a corpus and visualize the results. pSSLDA is an implementation by Andrzejewski following the method described in [7]. The EQUITBL codebase already provides multiple functionalities that we need, therefore our contribution is mainly the creation of a pipeline to debias a corpus. The pipeline uses pre-existing methods to preprocess the corpus, train a Topic Model and identify the sentences that should be removed. Then, it uses our implementation of a method to remove these sentences and retrains the model.

Hyperparameters We use the hyperparameter setting that is used in [8] and [4]. The document size is set on sentence level, i.e., we split the corpus into documents consisting of exactly one sentence, and M is equal the number of sentences in the corpus. We assume that 15 topics are underlying in the corpus and set $T = 15$. The z -label is set to 5, α to 0.33 and β to 0.2. Devinney et al. show that this setting is optimal for the given English Mainstream corpus [16].

3.4 Experiments

The gender biases inherent in the original corpus are identified by analyzing two Topic Models trained on the original corpus, an unsupervised one without seed words, and one with the three explicitly gendered topics. Then, 5%, 10% and 15% of the data are removed to find the interval where the gender biases against both women and men disappear. Once this interval is found, we refine the steps within this interval to find the threshold of how much data has to be removed.

Since Topic Modelling is non-deterministic, we use random seeds to obtain reproducible results. Nevertheless, we train and analyze three models at every stage of the experiments, which we compare to each other.

Qualitative Analysis Our analysis follows the analysis of [8]. We compare the top 50 words of the feminine and masculine topic, i.e., the words for which $p(w|z)$ is the highest. The non-binary topic is not analyzed, as it is not representative [8]. However, we also analyze the twelve non-seeded topics, to better understand differences between models trained on different corpora. For this, we try to find a single expression for every retrieved topic, which reflects the 50 most likely words for the topic. Note that this is a subjective task which is done by one person only, meaning that the name assigned to a topic might not be objectively speaking the most accurate. This inaccuracy is, however, not central to our task, and can therefore be neglected. Further, the topics are assigned with the background knowledge that the corpus consists of news articles. If this was not known, the topics could be harder to classify.

³ https://github.com/muesal/gender_biases

⁴ <https://github.com/davidandrzej/pSSLDA>

Analyzing the sentences that are removed when debiasing a corpus could lead to further insights, but is outside the scope of this work.

4 Results

This section first shows the topics retrieved from the original corpus. Those of the unsupervised Topic Model are presented in Section 4.1 and those of the supervised Topic Model in Section 4.2. The latter also presents the gender biases we find to be inherent in the original corpus. Then, the topics retrieved from the debiased corpora are presented. In Section 4.3, the topics of $\mathcal{D}^{(5)}$ are shown and analyzed, in Section 4.4 those of $\mathcal{D}^{(10)}$, in Section 4.5 those of $\mathcal{D}^{(12.5)}$ and, finally, in Section 4.6 those of $\mathcal{D}^{(15)}$.

4.1 Non-Seeded Model for the Original Corpus

The 15 topics retrieved through non-seeded Topic Modelling can be categorized as follows:

- | | | |
|---------------|--------------|--------------------|
| 1. Arts | 6. Arts | 11. Law |
| 2. Urban Life | 7. Sports | 12. Family |
| 3. Crime | 8. Health | 13. National Votes |
| 4. Travelling | 9. Politics | 14. Verbs |
| 5. Economy | 10. Business | 15. Education |

Three of these topics, namely *Health*, *Family* and *Education*, contain feminine seed words. Five of them contain masculine gendered seed words, *Crime*, *Law*, *Family*, *National Vote* and *Verbs*. The *Crime* topic contains a majority of the masculine gendered seed words. The *Family* topic contains more feminine than masculine seed words, and the feminine words have a higher probability than the equivalent masculine words.

The topics that contain feminine seed words all evolve around family and taking care of people. This is in accordance to the findings of Devinney et al. [8]. However, compared to their findings our model connects men surprisingly strong with *Crime*.

4.2 Seeded Model for the Original Corpus

The fifty most probable words for the feminine and masculine topic retrieved from the original corpus can be seen in Table 3. For the feminine topic, we see a lot of words that are related to either family (words in *italics*) or communication (words in **bold**). The masculine topic is dominated by words that are related to crime (words in **bold**). We also see a few words that refer to cars (words in *italics*). However, these words could refer to crime as well, as they could be used when referring to car accidents and theft.

Table 3. The 50 most probable words for the feminine and the masculine topics, trained on the original corpus. Sorted by their probability $p(w|z_i)$. **Bold** and *italic* words within each topic belong to the same category.

Feminine: she, **say**, woman, **tell**, year, *family*, **story**, him, com, old, go, their, time, continue, **friend**, know, day, **news**, life, **ask**, *mother*, *son*, get, take, come, want, girl, young, **write**, **twitter**, *home*, live, **medium**, *father*, *daughter*, advertisement, **call**, leave, **twitter**, them, see, **post**, *wife*, global, **hear**, it, we, **speak**, give, name

Masculine: he, say, **police**, year, man, Mr, **officer**, old, find, **police**, *vehicle*, **kill**, him, **death**, guy, people, **RCMP** (Royal Canadian Mounted Police), charge, **incident**, call, himself, **victim**, **shoot**, *car*, charge, *driver*, **arrest**, take, **suspect**, **crime**, *drive*, **murder**, **gun**, **injury**, leave, **die**, **scene**, believe, **shooting**, video, men, boy, time, involve, **assault**, **drug**, last, **investigation**, **dead**, face

The findings for the feminine topic are in agreement with the findings of Devinney et al. [8]. The topic is very similar to the *Family* topic of the non-seeded model. The masculine topic differs from the *Crime* topic as it has higher probabilities for the masculine seed words. The topic contains significantly more words connected to *Crime* than Devinney et al. found [8].

The remaining 12 topics retrieved from the original corpus, excluding the non-binary topic, can be classified as follows:

- | | | |
|---------------|-----------------------|----------------------|
| 4. Urban Life | 8. Education & Health | 12. Social Problems |
| 5. Economy | 9. US Administration | 13. National Votes |
| 6. Arts | 10. Business | 14. Food & Lifestyle |
| 7. Sports | 11. Law | 15. Verbs |

Apart from three differences, the topics of the seeded model are in accordance with the topics of the non-seeded model; the topics *Education* and *Health* are merged into one topic, and the *Crime* and the *Family* topic are replaced by the masculine and feminine topic, respectively.

The feminine seed words are more probable in the non-seeded *Family* topic than the masculine seed words in the *Crime* topic. This indicates, that the connection between women and family is stronger than the one between men and crime. In other words, the biases against women have a higher occurrence in the text than the biases against men. Nevertheless, the biases against men are evident as well. The debiased corpus should neither connect men more to crime than women nor women more to family and communication than men.

4.3 Removing 5% of the Data

We start by removing 5% of the sentences from the corpus to generate $\mathcal{D}^{(5)}$ and train three models on the new corpus. The masculine topic in the resulting mod-

els was about *Vote*, *Foreign Policy* and *Travel*. The feminine topics all evolved around *Family* and *Food & Lifestyle*.

This shows that removing 5% of data suffices to remove biases against men, as none of the masculine topics contains words evolving around crime. It also supports our hypothesis that more biases against women are inherent in the corpus, as the link between family and women was not removed.

4.4 Removing 10% of the Data

We continue by removing 10% of the corpus' data and training three models on $\mathcal{D}^{(10)}$. The differences between the three trained models are significant, and we analyze them carefully.

The first model still associates the feminine topic with family, but not with communication anymore. The retrieved masculine topic is very close to the previously discovered *Economics* topic. This topic is also missing from the remaining 12 topics. Instead, two topics evolve around politics, we call them *Foreign Policy* and *Home Affairs*. The second model retrieves the formerly found *Sports* topic as the feminine topic, and, as the first model, the *Economics* topic as the masculine topic. The third model associates women with a topic that seems to be a mix of the formerly retrieved *Social Problems* topic and *Foreign Policy*, while men are associated with a new topic which we call *Journalism*.

These results show that removing 10% is not enough to reliably remove the biases against women.

4.5 Removing 12.5% of the Data

Next, we remove 12.5% and analyze the respective models. The feminine topic of the first model is about *Journalism*. As such, it still contains words associated with communication, as did the feminine topic of the original corpus. The second model's feminine topic contains many words associated with family, while the feminine topic of the third model is the *Vote* topic. The masculine topics of the three models are *Sports*, *Health* and *Home affairs*, respectively.

Two of the models still retrieve feminine topics that reflect the gender biases found in the original corpus. We therefore draw the conclusion that removing 12.5% of the sentences is not enough to reliably remove all biases against women.

4.6 Removing 15% of the Data

Finally, we remove 15%, and again analyze the retrieved topics of three trained models. The retrieved feminine topics of the three trained models are *Vote*, *Sports* and *Home Affairs*. The masculine topics are *Verbs*, *Law* and *Foreign Policy*, respectively.

The order in which topics are assigned is non-deterministic, and varies between multiple models. The fact that the retrieved explicitly gendered topics are different for all three models indicates that they are independent of the seed

words. This implies that a significant amount of gender biases was removed, and the contexts in which the binary gendered seed words appear are similar over the two genders. This implication is further emphasized by the fact that the third model assigns a political topic to both the feminine and masculine topic. We conclude from this that removing 15% of all data is enough to remove the biases that could be seen in the original corpus.

It should be noted that we did not only remove the biases against both genders, but also all other sentences that are connected with the respective topics, *Family* and *Crime*. Our final corpus presumably does not contain many sentences about crime, family or communication. In other words, we did not merely debias the corpus but deleted information about two formerly present topics. This, together with the fact that having to remove 15% of data is in many cases not feasible, speaks against using topic modelling the way we did to debias a corpus.

5 Conclusion

In this work, we investigated how much data must be removed from a corpus to reduce the existing gender biases significantly. We trained Latent Dirichlet Allocation Topic Models on the English Mainstream corpus, a corpus consisting of English news articles, and used these models to identify the sentences that feed into biases the most. We then removed different amounts of these sentences, to analyze whether the smaller corpus still holds the gender biases we found to be inherent in the original corpus. By proceeding this way, we found that removing 5% of text is enough to significantly reduce the found biases against masculine people. To reduce the found biases against feminine people, 15% of the data must be removed. When removing fewer data, the gender biases are reduced, but still clearly visible.

The gap between the amount of data that has to be removed, in order to reduce biases for different genders, is significant. This gap could show a possible way to reduce gender biases with keeping more data. By removing different amounts of sentences for the two binary genders, the gender biases could be eliminated for both genders, while removing less data in total.

Further work could also aim to verify the quality of the remaining data, and investigate to what extent the biases were truly removed. We suggest to train other NLP models, e.g., Word2vec, on both the original and the debiased corpus, and to investigate differences between the resulting models. There, it should also be investigated how the representation of other words that are very probable for the gendered topics changes. This could indicate whether valuable information was lost, and how removing 15% of the data, a considerably large chunk, influences the corpus.

In addition, our experiments should be repeated on more corpora to verify that the suggested method does not only work on the corpus studied here. If this verification shows our method to work properly, we recommend it as a debiasing method, in cases where it is feasible to remove 15% of the data. When tools to

train the topic model are at hand, implementing a pipeline to debias a corpus is reasonably uncomplicated, and debiasing a corpus becomes straightforward.

Lastly, we want to point out that the studied method is not the only way to debias a corpus with Topic Models. Instead of removing the identified sentences, one could also add more sentences. For example, if the sentence “she is a nurse” is identified as a sentence that should be removed, instead the sentences “he is a nurse” and “they are a nurse” could be added.

References

1. Fitria, T.N.: Gender bias in translation using google translate: Problems and solution. *Language Circle: Journal of Language and Literature* **15** (2021)
2. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., eds.: *Advances in Neural Information Processing Systems*. Volume 29., Curran Associates, Inc. (2016)
3. Dahllöf, M., Berglund, K.: Faces, fights, and families: Topic modeling and gendered themes in two corpora of swedish prose fiction. In: DHN 2019 Copenhagen, Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries Copenhagen, March 6-8 2019. (2019) 92–111 Constanza Navaretta et al.
4. Devinney, H., Björklund, J., Björklund, H.: Crime and relationship: Exploring gender bias in NLP corpora. In: SLTC 2020–The Eighth Swedish Language Technology Conference, 25–27 November 2020, Online. (2020)
5. Hoyle, A.M., Wolf-Sonkin, L., Wallach, H., Augenstein, I., Cotterell, R.: Unsupervised discovery of gendered language through latent-variable modeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, Association for Computational Linguistics (July 2019) 1706–1716
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *Journal of machine Learning research* **3**(Jan) (2003) 993–1022
7. Andrzejewski, D., Zhu, X.: Latent dirichlet allocation with topic-in-set knowledge. In: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. (2009) 43–48
8. Devinney, H., Björklund, J., Björklund, H.: Semi-supervised topic modeling for gender bias discovery in english and swedish. In: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing. (2020) 79–92
9. Butler, J.: *Gender trouble: Feminism and the subversion of identity*. Routledge (2011)
10. Hellström, T., Dignum, V., Bensch, S.: Bias in machine learning - what is it good for? In Saffioti, A., Serafini, L., Lukowicz, P., eds.: *Proceedings of the 1st International Workshop on New Foundations for Human-Centered AI co-located with ECAI (2020)*, CEUR (2020) 3–10
11. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* **54**(6) (2021) 1–35
12. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR. (2013)

13. Mikolov, T., Yih, W.T., Zweig, G.: Linguistic regularities in continuous space word representations. (2013)
14. Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.W., Wang, W.Y.: Mitigating gender bias in natural language processing: Literature review. (2020) Give good definitions for (gender) biases in NLP.
15. Lerman, K., Hogg, T.: Leveraging position bias to improve peer recommendation. PloS one **9**(6) (2014) e98914
16. Devinney, H., Björklund, J., Björklund, H.: Explore, query, and understand implicit textual bias in language data: A topic modeling package for bias analysis. To be published. (2023)

Appendix

The following tables show the 20 most probable words for various topics⁵.

Crime and Family topic of $\mathcal{D}^{(0)}$ Table 4 shows the 20 most probable words with POS tags and their probability for the *Crime* and *Family* topics of the non-seeded Topic Model trained on the original corpus.

Gendered Topics of $\mathcal{D}^{(0)}$ Table 5 shows the 20 most probable words with POS tags and their probability for the feminine and masculine topics of the non-seeded Topic Model trained on the original corpus.

Feminine Topic of $\mathcal{D}^{(5)}$ Table 6 shows the 20 most probable words with POS tags and their probability of three models for the feminine topic of $\mathcal{D}^{(5)}$.

Masculine Topic of $\mathcal{D}^{(5)}$ Table 7 shows the 20 most probable words with POS tags and their probability of three models for the masculine topic of $\mathcal{D}^{(5)}$.

Feminine Topic of $\mathcal{D}^{(10)}$ Table 8 shows the 20 most probable words with POS tags and their probability of three models for the feminine topic of $\mathcal{D}^{(10)}$.

Masculine Topic of $\mathcal{D}^{(10)}$ Table 7 shows the 20 most probable words with POS tags and their probability of three models for the masculine topic of $\mathcal{D}^{(10)}$.

Feminine Topic of $\mathcal{D}^{(12.5)}$ Table 10 shows the 20 most probable words, with POS tag, and their probability of three models for the feminine topic of $\mathcal{D}^{(12.5)}$.

Masculine Topic of $\mathcal{D}^{(12.5)}$ Table 11 shows the 20 most probable words with POS tags and their probability of three models for the masculine topic of $\mathcal{D}^{(12.5)}$.

Feminine Topic of $\mathcal{D}^{(15)}$ Table 12 shows the 20 most probable words with POS tags and their probability of three models for the feminine topic of $\mathcal{D}^{(15)}$.

Masculine Topic of $\mathcal{D}^{(15)}$ Table 13 shows the 20 most probable words with POS tags and their probability of three models for the masculine topic of $\mathcal{D}^{(15)}$.

⁵ Tables 4 – 13 showing the 50 most probable words and POS tags in their respective categories are available at https://github.com/muesal/gender_biases

Table 4. The 20 most probable words, with POS tag, and their probability for the Crime and Family topics of the non-seeded Topic Model trained on the original corpus.

Crime		Family	
policeNN	0.02456	shePRP	0.05908
sayVB	0.02355	sayVB	0.03335
yearNN	0.01624	storyNN	0.01381
manNN	0.01389	getVB	0.01341
officerNN	0.01135	tellVB	0.01316
oldJJ	0.01079	familyNN	0.01279
findVB	0.00848	himPRP	0.01266
theyPRP	0.00830	yearNN	0.01226
policeNNP	0.00743	theirPRP	0.01034
killVB	0.00719	hePRP	0.00963
womanNN	0.00692	goVB	0.00951
hePRP	0.00666	timeNN	0.00870
vehicleNN	0.00645	oldJJ	0.00793
peopleNN	0.00632	friendNN	0.00760
deathNN	0.00625	todayNN	0.00736
rcmpNNP	0.00561	theyPRP	0.00711
chargeVB	0.00553	knowVB	0.00706
incidentNN	0.00541	dayNN	0.00699
victimNN	0.00508	continueVB	0.00643

Table 5. 20 most probable words, with POS tag, and their probability for the masculine and feminine topics of the seeded Topic Model trained on the original corpus.

Masculine		Feminine	
hePRP	0.04410	shePRP	0.07615
sayVB	0.02648	sayVB	0.02906
policeNN	0.02364	womanNN	0.02096
yearNN	0.01640	tellVB	0.01654
manNN	0.01543	yearNN	0.01401
mrNNP	0.01318	storyNN	0.01089
officerNN	0.01105	himPRP	0.00953
oldJJ	0.01034	comNN	0.00951
findVB	0.00831	oldJJ	0.00873
killVB	0.00747	familyNN	0.00851
policeNNP	0.00738	continueVB	0.00784
deathNN	0.00675	newsNNP	0.00666
himPRP	0.00659	goVB	0.00601
vehicleNN	0.00639	timeNN	0.00564
guyNN	0.00618	friendNN	0.00559
rcmpNNP	0.00553	dayNN	0.00526
peopleNN	0.00551	writeVB	0.00515
chargeVB	0.00544	motherNN	0.00502
incidentNN	0.00520	sonNN	0.00488
callVB	0.00495	mediumNN	0.00482

Table 6. The 20 most probable words, with POS tag, and their probability of three models for the feminine topic of $\mathcal{D}^{(5)}$.

Model 1		Model 2		Model 3	
shePRP	0.05889	shePRP	0.06280	shePRP	0.06314
sayVB	0.03285	womanNN	0.01862	womanNN	0.01872
womanNN	0.01746	foodNN	0.00656	makeVB	0.00651
theirPRP\$	0.01676	roomNN	0.00544	itPRP	0.00632
peopleNN	0.01352	useVB	0.00525	roomNN	0.00596
familyNN	0.01332	makeVB	0.00488	foodNN	0.00578
itPRP	0.01183	wearVB	0.00410	sayVB	0.00553
yearNN	0.01178	girlNN	0.00408	theirPRP\$	0.00536
lifeNN	0.01032	itPRP	0.00405	useVB	0.00449
oldJJ	0.00875	findVB	0.00383	wearVB	0.00415
manyJJ	0.00700	blackJJ	0.00369	themPRP	0.00410
timeNN	0.00699	waterNN	0.00352	girlNN	0.00410
youngJJ	0.00670	footNN	0.00344	getVB	0.00395
himPRP	0.00663	homeNN	0.00327	lookVB	0.00394
friendNN	0.00604	smallJJ	0.00325	homeNN	0.00392
liveVB	0.00573	whiteJJ	0.00315	goVB	0.00384
knowVB	0.00551	handNN	0.00308	findVB	0.00369
comeVB	0.00511	spaceNN	0.00290	handNN	0.00362
tellVB	0.00503	floorNN	0.00284	blackJJ	0.00360
goVB	0.00493	lookVB	0.00278	comeVB	0.00333

Table 7. The 20 most probable words, with POS tag, and their probability of three models for the feminine topic of $\mathcal{D}^{(5)}$.

Model 1		Model 2		Model 3	
hePRP	0.03589	hePRP	0.03352	hePRP	0.04119
governmentNN	0.01743	trumpNNP	0.01946	sayVB	0.02219
partyNN	0.01454	sayVB	0.01889	mrNNP	0.01344
sayVB	0.01393	mrNNP	0.01094	manNN	0.01207
electionNN	0.01312	manNN	0.00981	fireNN	0.00946
trudeauNNP	0.01196	presidentNNP	0.00784	areaNN	0.00764
mrNNP	0.01173	stateNNP	0.00758	waterNN	0.00719
manNN	0.01050	countryNN	0.00756	dayNN	0.00646
canadaNNP	0.00866	chinaNNP	0.00741	guyNN	0.00618
federalJJ	0.00860	unitedNNP	0.00665	himselfPRP	0.00485
leaderNN	0.00784	presidentNN	0.00632	peopleNN	0.00481
campaignNN	0.00771	kongNNP	0.00618	highJJ	0.00399
ministerNN	0.00727	hongNNP	0.00614	homeNN	0.00376
ministerNNP	0.00660	houseNNP	0.00601	findVB	0.00371
liberalNNP	0.00646	officialNN	0.00575	expectVB	0.00354
quebecNNP	0.00619	governmentNN	0.00505	yearNN	0.00352
candidateNN	0.00543	guyNN	0.00503	hourNN	0.00347
guyNN	0.00539	stateNN	0.00467	weatherNN	0.00328
politicalJJ	0.00532	donaldNNP	0.00436	morningNN	0.00323
changeNN	0.00529	protestNN	0.00420	powerNN	0.00303

Table 8. The 20 most probable words, with POS tag, and their probability of three models for the feminine topic of $\mathcal{D}^{(10)}$.

Model 1		Model 2		Model 3	
shePRP	0.04410	shePRP	0.05347	shePRP	0.05485
sayVB	0.02648	gameNN	0.01774	sayVB	0.02615
yearNN	0.02364	womanNN	0.01664	womanNN	0.01713
womanNN	0.01640	seasonNN	0.01418	peopleNN	0.01428
theirPRP\$	0.01543	goalNN	0.01176	rightNN	0.00907
familyNN	0.01318	teamNN	0.01147	theirPRP\$	0.00775
oldJJ	0.01105	firstJJ	0.01054	kongNNP	0.00760
timeNN	0.01034	yearNN	0.00997	hongNNP	0.00755
himPRP	0.00831	playVB	0.00986	groupNN	0.00719
goVB	0.00747	lastJJ	0.00759	communityNN	0.00638
lifeNN	0.00738	playerNN	0.00729	policeNN	0.00607
dayNN	0.00675	pointNN	0.00726	itPRP	0.00589
tellVB	0.00659	getVB	0.00686	manyJJ	0.00553
itPRP	0.00639	goVB	0.00603	issueNN	0.00416
homeNN	0.00618	timeNN	0.00592	takeVB	0.00410
takeVB	0.00553	yardNN	0.00579	violenceNN	0.00406
getVB	0.00551	makeVB	0.00568	canadaNNP	0.00405
friendNN	0.00544	scoreVB	0.00555	callVB	0.00395
leaveVB	0.00520	secondJJ	0.00552	protesterNN	0.00391
youngJJ	0.00495	playNN	0.00512	protestNN	0.00390

Table 9. The 20 most probable words, with POS tag, and their probability of three models for the feminine topic of $\mathcal{D}^{(10)}$.

Model 1		Model 2		Model 3	
hePRP	0.03526	hePRP	0.03613	hePRP	0.04592
yearNN	0.02608	yearNN	0.03201	mrNNP	0.01600
centNN	0.02012	centNN	0.02584	tellVB	0.01405
mrNNP	0.01229	mrNNP	0.01260	storyNN	0.01388
companyNN	0.01189	rateNN	0.01004	getVB	0.01277
marketNN	0.01155	highJJ	0.00981	manNN	0.01154
sayVB	0.00902	monthNN	0.00963	newsNN	0.01057
manNN	0.00888	marketNN	0.00938	sayVB	0.01039
lastJJ	0.00869	sayVB	0.00935	dayNN	0.01027
rateNN	0.00860	lastJJ	0.00910	todayNN	0.00984
monthNN	0.00858	manNN	0.00909	continueVB	0.00926
highJJ	0.00831	taxNN	0.00794	comNN	0.00892
priceNN	0.00791	priceNN	0.00777	writeVB	0.00879
shareNN	0.00662	expectVB	0.00675	mediumNN	0.00866
expectVB	0.00583	riseVB	0.00585	giveVB	0.00748
canadaNNP	0.00576	economyNN	0.00575	guyNN	0.00717
tradeNN	0.00574	shareNN	0.00566	newsNNP	0.00704
saleNN	0.00564	guyNN	0.00565	signNNP	0.00678
riseVB	0.00553	tradeNN	0.00565	localJJ	0.00677
guyNN	0.00551	growthNN	0.00551	himPRP	0.00637

Table 10. The 20 most probable words, with POS tag, and their probability of three models for the feminine topic of $\mathcal{D}^{(12.5)}$.

Model 1		Model 2		Model 3	
shePRP	0.05285	shePRP	0.04877	shePRP	0.04932
sayVB	0.01974	sayVB	0.02789	partyNN	0.01604
womanNN	0.01691	theirPRP\$	0.02136	sayVB	0.01590
peopleNN	0.01220	peopleNN	0.01955	womanNN	0.01580
mediumNN	0.00773	womanNN	0.01559	electionNN	0.01424
itPRP	0.00678	familyNN	0.00883	trudeauNNP	0.01329
socialJJ	0.00671	communityNN	0.00801	governmentNN	0.01062
theirPRP\$	0.00635	manyJJ	0.00800	leaderNN	0.00835
rightNN	0.00575	itPRP	0.00783	canadaNNP	0.00790
communityNN	0.00569	lifeNN	0.00674	ministerNN	0.00773
manyJJ	0.00569	themPRP	0.00597	campaignNN	0.00714
localJJ	0.00532	rightNN	0.00530	liberalNNP	0.00714
groupNN	0.00519	groupNN	0.00493	ministerNNP	0.00634
newsNN	0.00444	wePRP	0.00489	quebecNNP	0.00631
dailyJJ	0.00414	liveVB	0.00474	federalJJ	0.00607
worldNN	0.00397	youngJJ	0.00465	conservativeNNP	0.00573
getNNP	0.00394	canadaNNP	0.00453	scheerNNP	0.00557
becomeVB	0.00388	wayNN	0.00441	ndpNNP	0.00525
canadaNNP	0.00388	makeVB	0.00436	voteNN	0.00525
writeVB	0.00364	countryNN	0.00434	candidateNN	0.00515

Table 11. The 20 most probable words, with POS tag, and their probability of three models for the feminine topic of $\mathcal{D}^{(12.5)}$.

Model 1		Model 2		Model 3	
hePRP	0.03157	hePRP	0.03546	hePRP	0.03751
gameNN	0.02020	sayVB	0.02520	sayVB	0.02746
seasonNN	0.01503	schoolNN	0.01571	trumpNNP	0.02109
mrNNP	0.01133	studentNN	0.01335	mrNNP	0.01348
firstJJ	0.01091	healthNN	0.01301	houseNNP	0.00917
goalNN	0.01027	mrNNP	0.01273	manNN	0.00891
teamNN	0.00855	careNN	0.00852	himPRP	0.00789
playVB	0.00778	manNN	0.00840	presidentNN	0.00774
pointNN	0.00757	theirPRP\$	0.00802	tellVB	0.00740
manNN	0.00751	yearNN	0.00800	mediumNN	0.00703
secondJJ	0.00646	peopleNN	0.00619	presidentNNP	0.00641
scoreVB	0.00644	highJJ	0.00617	guyNN	0.00594
lastJJ	0.00627	guyNN	0.00563	callVB	0.00536
yearNN	0.00595	findVB	0.00538	formerJJ	0.00479
playerNN	0.00568	universityNNP	0.00522	whiteNNP	0.00479
yardNN	0.00530	studyNN	0.00515	democratNNP	0.00477
runNN	0.00527	programNN	0.00507	commentNN	0.00466
guyNN	0.00499	helpVB	0.00487	himselfPRP	0.00464
playNN	0.00474	healthNNP	0.00482	donaldNNP	0.00453
himPRP	0.00468	drugNN	0.00466	makeVB	0.00436

Table 12. The 20 most probable words, with POS tag, and their probability of three models for the feminine topic of $\mathcal{D}^{(15)}$.

Model 1		Model 2		Model 3	
shePRP	0.04411	shePRP	0.03281	shePRP	0.05245
partyNN	0.01568	gameNN	0.02128	sayVB	0.03430
governmentNN	0.01466	seasonNN	0.01569	trumpNNP	0.01959
womanNN	0.01456	teamNN	0.01339	womanNN	0.01727
electionNN	0.01426	firstJJ	0.01239	presidentNN	0.01232
sayVB	0.01338	womanNN	0.01080	houseNNP	0.00970
trudeauNNP	0.01284	goalNN	0.00859	tellVB	0.00930
campaignNN	0.00871	playVB	0.00851	newNNP	0.00814
leaderNN	0.00838	yearNN	0.00788	formerJJ	0.00712
federalJJ	0.00814	pointNN	0.00753	presidentNNP	0.00611
canadaNNP	0.00778	playerNN	0.00735	stateNN	0.00555
ministerNN	0.00750	secondJJ	0.00706	whiteNNP	0.00520
liberalNNP	0.00702	lastJJ	0.00683	yorkNNP	0.00514
ministerNNP	0.00626	torontoNNP	0.00629	democratNNP	0.00506
quebecNNP	0.00602	scoreVB	0.00556	reportVB	0.00472
candidateNN	0.00601	timeNN	0.00523	donaldNNP	0.00413
politicalJJ	0.00581	thirdJJ	0.00461	callNN	0.00400
ndpNNP	0.00564	runNN	0.00447	officialNN	0.00400
conservativeNNP	0.00557	winVB	0.00439	speakVB	0.00393
voteNN	0.00546	winNN	0.00421	republicanNNP	0.00391

Table 13. The 20 most probable words, with POS tag, and their probability of three models for the feminine topic of $\mathcal{D}^{(15)}$.

Model 1		Model 2		Model 3	
sayVB	0.04062	hePRP	0.03150	hePRP	0.03401
hePRP	0.03274	sayVB	0.02473	sayVB	0.02163
getVB	0.03048	mrNNP	0.01178	mrNNP	0.01270
goVB	0.02712	courtNN	0.01070	countryNN	0.01182
itPRP	0.02352	caseNN	0.01014	unitedNNP	0.01030
goodJJ	0.02023	lawNN	0.00766	stateNNP	0.00969
makeVB	0.01908	manNN	0.00722	chinaNNP	0.00882
thinkVB	0.01674	policeNN	0.00646	manNN	0.00782
wePRP	0.01597	lawyerNN	0.00579	trumpNNP	0.00642
thingNN	0.01515	himPRP	0.00533	guyNN	0.00549
timeNN	0.01298	decisionNN	0.00511	canadaNNP	0.00517
mrNNP	0.01221	guyNN	0.00508	governmentNN	0.00512
knowVB	0.01133	courtNNP	0.00508	officialNN	0.00508
wayNN	0.01121	tellVB	0.00507	iranNNP	0.00488
lotNN	0.01056	statementNN	0.00479	yearNN	0.00483
seeVB	0.01013	investigationNN	0.00459	forceNN	0.00471
himPRP	0.00994	chargeNN	0.00448	dealNN	0.00469
wantVB	0.00975	officerNN	0.00443	warNN	0.00468
comeVB	0.00837	reportNN	0.00430	militaryJJ	0.00461
lookVB	0.00768	judgeNN	0.00414	tradeNN	0.00451

Minimizing Lost Updates Under Read-Atomic Isolation With Lazy Transactions

Lucas Paes

Department of Computing Science
Umeå University, Sweden
1220958@dac.unicamp.br

Abstract. Coordination avoidance and partition-independence have long been seen as necessary properties of truly scalable, fault-tolerant database systems. Despite having usually been associated with weaker semantic guarantees, Read-Atomic Multipartition (RAMP) protocols have demonstrated it is possible maintain those properties while achieving a useful middle ground between scalability and semantic correctness. In this paper, we propose a server-side scheduling policy that aims to reduce the lost updates anomaly with RAMP protocols, and evaluate the efficacy of this solution through experimental simulations. Our results show that lazy execution of updates does minimize lost updates, but at the cost of increasing read transaction latency.

1 Introduction

Given the ever-increasing storage capacity and query load requirements, it is no longer reasonable — nor feasible — to scale a single physical machine to host a modern database system in its entirety. Due to this, it is quite common to see database systems distribute data over multiple servers (*partitions*), such that no single partition has a copy of the entire dataset.

Despite having its advantages, partitioning introduces its own set of challenges. Most notably, all partitions upon which a given transaction depends must now communicate when it is executed in order to guarantee semantic correctness of the stored and returned data. Since they are key to achieving modern scalability requirements, such communication strategies have been a topic of extensive research.

Among many approaches to this particular problem, Bailis et al. [1] suggest a relaxation of semantic guarantees — a new isolation model called Read-Atomic (RA) — that still provides transactional access while requiring no coordination between partitions at all, and is useful for a whole class of applications. They then suggest a family of protocols (RAMP) which concretely implement RA isolation.

Since their main focus is formalizing RA isolation and the RAMP family of protocols, in their original work, Bailis et al. [1] only consider the execution of read-only and write-only transactions, without investigating about how a server partition might schedule execution of such transactions internally. In this work,

we expand on their analysis by proposing a scheduling policy that reduces lost updates that might occur in a write-intensive workload without compromising RA isolation. We evaluate this policy by simulating how it behaves given different workloads and configurations, and compare it to a simple, eager execution policy. Our results show that our policy can benefit systems in which read transaction latency is less critical than having less lost update anomalies.

This paper is organized as follows: Section 2 introduces relevant concepts. Then, Section 3 outlines background information and earlier work, followed by Section 4, which formalizes our proposed scheduling policy. Finally, Section 5 introduces our simulation setup and method, Section 6 discusses simulation results and implications, and Section 7 acknowledges limitations and indicates interesting research topics for future work.

2 Definitions

Before proceeding, we first define relevant key concepts that will be used throughout this paper. We acknowledge that most of these have been previously stated by Bailis et al. [1] and Adya [2], and are not our original work. Refer to their original work for more detailed information about such concepts.

2.1 Scalability

Definition 1 (Coordination-free). *A system can be considered coordination-free if, under no circumstances, can a transaction T_1 be stalled by any other transaction T_2 in that same system.*

By definition, coordination-free systems avoid slow response times under high query volumes, since no transaction from any client C_2 can be the reason for a client C_1 's query to slow down.

Such systems are said to have *transactional availability*.

Definition 2 (Partition-independence). *A system can be considered partition-independent if, under no circumstances, does a transaction T that solely depends on data under responsibility of partitions in set of partitions P have to communicate with any partition not in P .*

Partition-independence makes a system more resilient to failures. To understand how, consider a partial failure scenario of a partition-independent system, where partitions in P_n operate normally while partitions in P_f have failed due to hardware, networking, weather, or any other non-deterministic problem. If a transaction T depends on data that is only hosted by partitions in P_n , T can still execute normally, since it will never need to communicate with any of the partitions in P_f . Were the system not partition-independent, T might have failed if a response from a failed (and unresponsive) partition was needed.

For that reason, partition-independent systems are said to have *replica availability*.

Definition 3 (Scalability). *A system can be considered scalable if it provides coordination-free and partition-independent execution of transactions.*

For the purposes relevant to this paper, we define scalability in terms of transactional and replica availability, and consider it the defining property of available and fault-tolerant distributed systems.

2.2 System model and observed anomalies

Based on the work of [2], Bailis et al. [1] declare a set of anomalies that may occur given a transaction history H , and define RA isolation as a model that does not allow such anomalies to occur. To avoid being overly repetitive, we briefly — and not very formally — refresh such definitions here. Please refer to [2] and Bailis et al. [1] for more careful and formal explanations.

Definition 4 (Transactions). *A transaction is an ordered set of read and write operations that are atomically visible to an observer. Each transaction can either end in a commit (success) or abort (failure). For convenience, we consider all transactions committed unless explicitly stated.*

Definition 5 (Versioning). *We consider a system with versioned data items, and assume x_i is a version with timestamp i of the data item x . Timestamps are taken from a totally ordered set (natural numbers), and we assume higher timestamps represent newer versions. We consider every data item has an initial version of timestamp \perp .*

We can now consider how a transaction might observe or affect another transaction’s behavior.

Definition 6 (Read-dependency). *A transaction T_j directly read-dependes on T_i if T_j reads version x_k , and x_k was written by T_i .*

Read-dependencies allow us to model transactions that read other transactions’ writes.

Definition 7 (Antidependency). *A transaction T_j directly antidepends on T_i if T_i reads version x_k , and T_j writes the subsequent version of x after k .*

Definition 8 (Write-dependency). *A transaction T_j directly write-dependes on T_i if T_i writes version x_k , and T_j writes the subsequent version of x after k .*

Antidependencies and write-dependencies respectively model transactions that read/write values which are overwritten by following transactions.

We can use such definitions to build a directed graph of dependency relationships between transactions.

Definition 9 (Direct Serialization Graph). *A Direct Serialization Graph arises from a transaction history H , and is denoted as $DSG(H)$. Each node in $DSG(H)$ is a transaction in H , and directed edges between nodes represent read/anti/write-dependencies.*

With this tooling, it is possible to analyze properties of a given $DSG(H)$ to determine whether certain anomalies can occur while processing transactions in H .

Definition 10 (Aborted Reads). *History H has an Aborted Read if it contains transactions T_a and T_c such that T_c reads x_a , x_a was written by T_a and T_a was aborted.*

Definition 11 (Intermediate Reads). *History H has an Intermediate Read if it contains committed transactions T_i and T_f such that T_f writes both x_j and x_k , but T_i reads x_j despite $j < k$.*

Definition 12 (Circular Information Flow). *History H has a Circular Information Flow if $DSG(H)$ contains a directed cycle of read-dependency and write-dependency edges.*

With these three definitions, Bailis et al. [1] define Read-Atomic isolation as follows.

Definition 13 (Read-Atomic isolation). *A system provides Read-Atomic isolation if it disallows Aborted Reads, Intermediate Reads and Circular Information Flow.*

Informally, transactions under RA isolation see a “snapshot” of the database, and thus cannot observe other uncommitted or aborted transactions.

Finally, we refresh the definition of the “Lost Update” phenomenon, from [2]:

Definition 14 (Lost Update). *History H has a Lost Update if $DSG(H)$ contains a directed cycle consisting of one or more antidependency and all edges are over data item x .*

Intuitively, lost updates happen whenever two concurrent transactions overwrite the same value, and one of them is completely discarded since it is overwritten by the other.

2.3 Partition-Local Read-Write and Lazy Transactions

Definition 15 (Partition-Local Read-Write transaction). *A Partition-Local Read-Write (PLRW) transaction is a transaction that consists of reads followed by write statements which use values that were read. If a write happens in partition P , then all reads upon which it depends must also happen in P . It may not return values to the issuing client.*

Read-write partition-local transactions are useful for operations that update values in a record based on the previous stored value. One example is incrementing or decrementing a counter, where the written value depends on the locally read value.

Definition 16 (Lazy Transactions). *A system that offers Lazy Transactions may defer transaction processing to the future, despite having returned commit promises to issuing clients.*

As emphasized by [3] in their original paper about Lazy Transactions, they allow for reduced resource contention and temporal load balancing. As we will further explain in this paper, we discovered that transactions executed lazily offer improved semantic guarantees for RAMP protocols at the cost of increased latency.

3 Earlier work

3.1 Isolation models

Isolation models have been a central point of discussion around transaction scalability versus semantic guarantees in database systems. The strongest model is full (or one-copy) serializability [2]: the entire database is seen as a single, logical copy of the data collection, and all transactions happen as if they had been serially executed one after the other. Despite always guaranteeing semantic soundness, in practice, full serializability has generally demonstrated itself hard to achieve without degrading scalability [4, 5], and is usually associated with some degree of coordination and partition-dependence. Fortunately, for certain applications, it has been demonstrated that full serializability can be relaxed without compromising correctness [5, 1, 6].

For this reason, significant efforts have been put in defining weaker — but still useful — isolation models. Academically, [2] formalize a spectrum of isolation models, and others [7, 8] provide contributions to this spectrum in the form of weaker isolation models with specific desirable properties, as well as ideas for implementing and evaluating such models in real database systems. Concretely, there has been a surge of database technologies which claim unforeseen scalability¹ [9, 10, 11], but compromise data integrity and delegate semantic checks to the application layer, to varying degrees.

Among all these efforts, most notably Bailis et al. [1] describe a novel isolation model, namely Read-Atomic (RA) isolation, which attends a relevant class of real-world applications. We better explore RA isolation and why its relevance in the next subsection.

3.2 Read-Atomic isolation and RAMP protocols

Definition 13 states that RA isolation provides clients with a snapshot view of the database. As shown by Bailis et al. [1] in their work, this is a valuable property that is sometimes sufficient for application correctness. In particular, they highlight bidirectional foreign-key enforcement, Secondary indexing and

¹ <https://scylladb.com>

materialized view maintenance as important use cases that are supported by RA isolation.

Our work adds partition-local updates — such as updating an account balance or a counter value — with reduced lost updates to these use cases. Despite not eliminating lost updates completely, reducing such anomalies in a database system is still desirable and valuable.

4 Investigation

Before continuing with our proposed solution, we define relevant notation. A PLRW transaction is denoted as a function that has a read-set and a write-set as arguments. As an example, $X(\{x\}, \{y, z\})$ denotes a PLRW transaction X that reads data item x and writes to data items y and z .

Consider the following example, in which two clients, C_1 and C_2 issue *run* and *commit* requests for the given PLRW transactions:

$$\begin{array}{ll} C_2 & \text{run}(Y(\{x\}, \{x\})) \\ C_1 & \text{run}(X(\{\}, \{x\})) \\ C_2 & \text{commit}(Y) \\ C_1 & \text{commit}(X) \end{array} \quad (1)$$

History 1 exhibits the Lost Update phenomenon, since, despite being committed X 's updates to x are overwritten by Y , which ran first.

Fortunately, we can take advantage of the “Now” and “Later” phases introduced by Faleiro, Thomson and Abadi [3] to defer execution of Y to after it has been committed, so that X runs first and its update is not lost. We now consider that clients *prepare* transactions, and that a *commit* is a promise made by partitions it will be executed. Partitions may actually *run* the transaction in the future. In such a model, History 2 is equivalent to History 1, but we do not observe a Lost Update, since X is smartly scheduled to run before Y despite their commit order.

$$\begin{array}{ll} C_2 & \text{prepare}(Y(\{x\}, \{x\})) \\ C_1 & \text{prepare}(X(\{\}, \{x\})) \\ C_2 & \text{commit}(Y) \\ C_1 & \text{commit}(X) \\ S & \text{run}(X) \\ S & \text{run}(Y) \end{array} \quad (2)$$

We now introduce a scheduling policy that defers the *run* phase depending on a few conditions.

First, assume each partition keeps track of its prepared but not yet committed transactions, so that partitions know if there are any partition-local antidependencies. This metadata can be stored server-side as a list of uncommitted write-sets or as a counter of uncommitted antidependencies for each data item.

In the former case, the metadata grows with the number of uncommitted PLRW transactions, while the latter grows with the number of data items.

We initially consider two scenarios when committing a PLRW transaction T in partition P :

No antidependency If all values read by T are not marked as being part of the write-set of any uncommitted transaction in P , T gets executed eagerly, since it will not be overwritten any prepared transactions in the event they are committed.

Antidependency If any value read by T would be updated by a scheduled transaction T_s , T 's execution is deferred until there are no antidependencies.

The first case covers the trivial scenario where no transactions interfere on each other, while the latter considers scenarios similar to History 2.

There is one problem with this approach: the only condition for deferring a transaction's execution is the existence of another transaction that overwrites its read-set. In the worst case, this could lead to a long antidependency chain that results in T 's execution being deferred indefinitely. There are two obvious solutions to this. The first solution is to add a time limit, and force T 's execution if its maximum waiting time has been exceeded. The second solution is an antidependency depth limit, which also results in T 's eager execution.

While both solutions might still lead to a potential Lost Update, this approach could still be valuable to prevent some of them in practical situations where the number of antidependencies for a particular data item is usually bounded.

After introducing our scheduling policy intuitively and rather informally, we introduce a more formal, algorithmic definition in Algorithm 1.

In the next section, we will introduce the experimental setup that was used for benchmarking and evaluated the proposed solution.

5 Method

Our benchmark was implemented in Python, and was used to simulate a partitioned database system that supports RAMP. The simulated database schema is simply a mapping of counters to keys. A random workload of equally distributed read (get counter value), write (set counter to zero) and read-write (increase counter) transactions over those counters is generated, and those transactions are submitted to the simulated partitions.

The simulator tracks read transactions that observe lost updates by comparing results produced by each scheduling policy with a serial, one by one execution of the given workload. If a transaction observes a different value than the serial execution, we consider it an incorrect read (due to a lost update). We can then calculate a percentage of correct and incorrect reads for each scheduling policy.

Moreover, we also track read transaction latency by comparing how long a given read transaction would take to complete (read its results) after it was submitted in the eager simulator versus in our policy's simulator. The introduced

Algorithm 1 Scheduling policy

```

1:  $T_d \leftarrow$  set of deferred transactions in current partition
2:  $T_r \leftarrow$  set of transactions that are running in current partition
3: procedure HASANTIDEPENDENCIES( $T$ )  $\triangleright$  Check if  $T$  has antidependencies
4:    $W_d \leftarrow \bigcup_{T' \in T_d} \text{WriteSet}(T')$ 
5:    $W_r \leftarrow \bigcup_{T' \in T_r} \text{WriteSet}(T')$ 
6:    $C \leftarrow \text{ReadSet}(T) \cap W_d \cap W_r$ 
7:   return  $|C| \neq 0$ 
8: end procedure
9: procedure EXECUTEEAGER( $T$ )  $\triangleright$  Execute a transaction eagerly
10:   Add  $T$  to  $T_r$  and remove  $T$  from  $T_d$ , atomically
11:   Create a background job to run  $T$  and yield until execution is finished
12:   Remove  $T$  from  $T_r$ 
13:   for  $T'$  in  $T_d$  do  $\triangleright$  Check if any previously deferred transactions can now run
14:     if not HASANTIDEPENDENCIES( $T$ ) then
15:       EXECUTEEAGER( $T$ )
16:     end if
17:   end for
18: end procedure
19: procedure COMMIT( $T$ )  $\triangleright$  Commit a PLRW transaction on a partition
20:   if HASANTIDEPENDENCIES( $T$ ) then
21:     Add  $T$  to  $T_d$ 
22:   else
23:     EXECUTEEAGER( $T$ )
24:   end if
25: end procedure

```

latency is then calculated as a simple subtraction between the former and the latter.

For consistency and fairness, the exact same workloads were run for the eager update database and for the simulation of our scheduling policy. Also, the system maintains a global clock, and runs events in the same clock cycle as if they happened in parallel, and thus does not model hardware, network performance and other phenomena such as operating system scheduling.

Workloads were generated for 10 partitions, with 100 average transactions per second and standard deviation of 1. Each transaction had a random duration, with average 1 and standard deviation of also 1. We varied the total amount of stored keys from 10 to 1000. The pseudocode for the simulation is in Algorithm 2.

For repeatability, all source code is available at our Github Repository ².

6 Experimental results and discussion

For reproducibility, all raw data produced by simulations can be found at our Github repository.

² <https://github.com/serramatutu/sccs-code>

Algorithm 2 RAMP simulation

```

1: procedure PREPARE ▷ Initialize the simulation
2:    $T \leftarrow$  set of (get|overwrite|increase, timestamp, key) transactions
3:    $P \leftarrow$  set of simulated partitions
4:    $K_p \leftarrow$  set of keys stored by  $P$ , all initialized to zero
5:    $time \leftarrow 0$ 
6:    $r \leftarrow a \bmod b$ 
7: end procedure
8: procedure RUN ▷ Run the simulation
9:   PREPARE
10:  while  $T$  has transactions do
11:     $T_{now} \leftarrow \{t \in T \text{ if } t.timestamp \leq time\}$ 
12:     $T = T \setminus T_{now}$ 
13:    for  $t$  in  $T_{now}$  do
14:       $p \leftarrow$  partition for  $t.key$ 
15:      execute or schedule  $t$  over  $t.key$  in  $p$  depending on the policy
16:    end for
17:    count unnecessary lost updates
18:     $time \leftarrow time + 1$ 
19:  end while
20: end procedure

```

6.1 Added latency

As visible from Figure 1, our policy introduces significant performance penalties for few keys that are affected by many writes at the same time. However, as the keyspace size increases and, thus, the write per key rate decreases, the extra latency introduced by our policy decreases quickly.

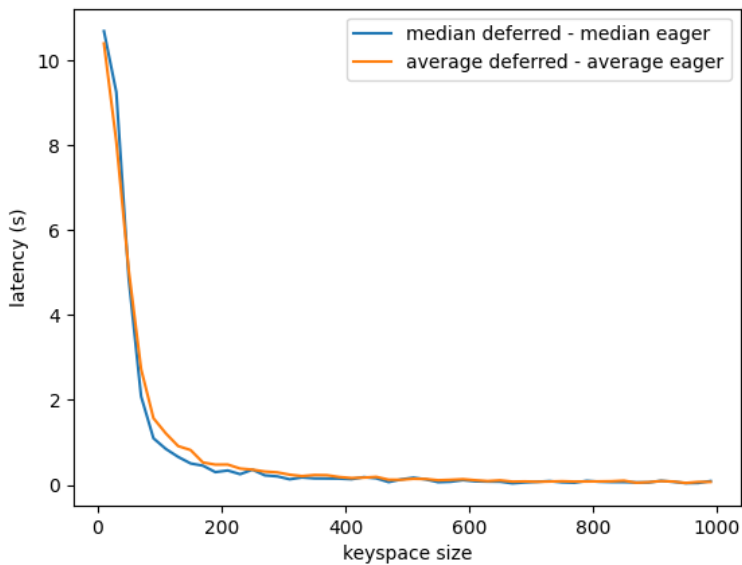
By looking more carefully at the tail of Figure 1.b, we can observe our solution stabilizes at around 50ms of added latency to read transactions (Figure 2). From this, we can conclude that, given that write-intensive workloads for a given key do not surpass a certain write per second threshold, our solution's introduced latency is controlled even if no maximum latency is set. As expected, if a maximum latency is set, our solution's maximum introduced latency caps at the set value (Figure 1.a) even on the worst case.

6.2 Lost Updates reduction

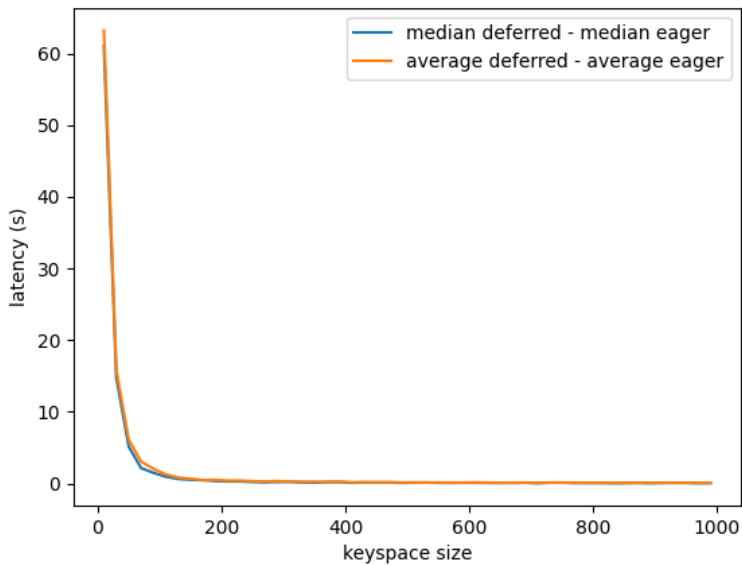
As Figure 3 shows, our policy helps decrease the amount of lost updates, and always stays ahead of eager execution in terms of correct reads.

6.3 Implications

From these results, we can conclude our solution performs well at reducing lost updates for workloads with moderate writes per second per key, at the cost of increased read transaction latency. This suggests that the proposed scheduling policy can be useful on systems with such workloads, and in which correctness is more valuable than read transaction latency.



(a) Maximum latency capped at 10s before eager execution



(b) Maximum latency uncapped

Fig. 1: Average and median read transaction latency increase by keyspace size. We subtracted the deferred latency from the eager latency to discover how much extra latency deferred transactions introduce.

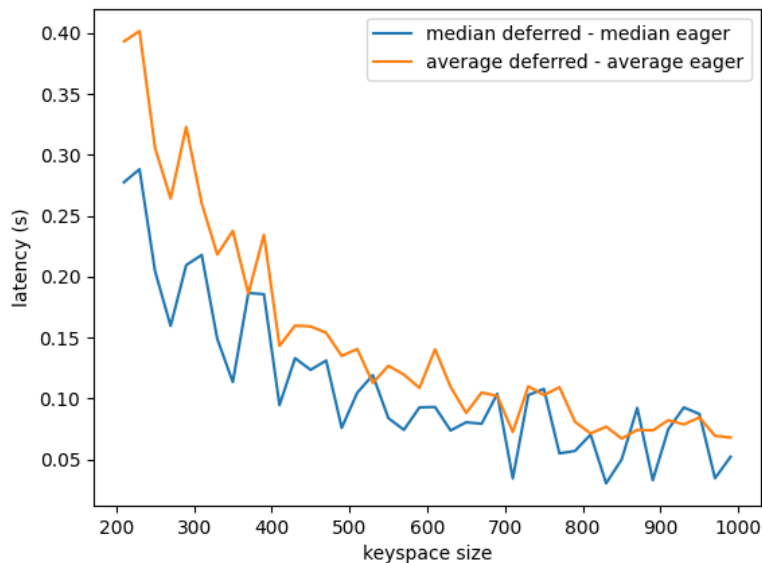
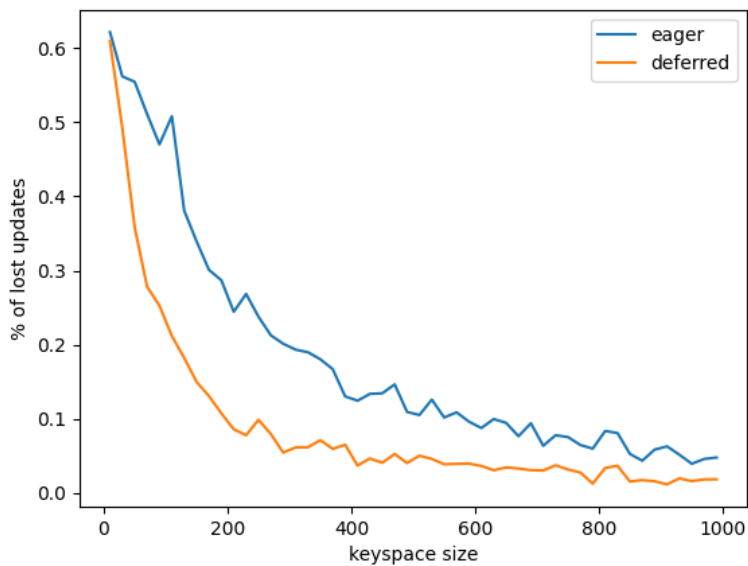
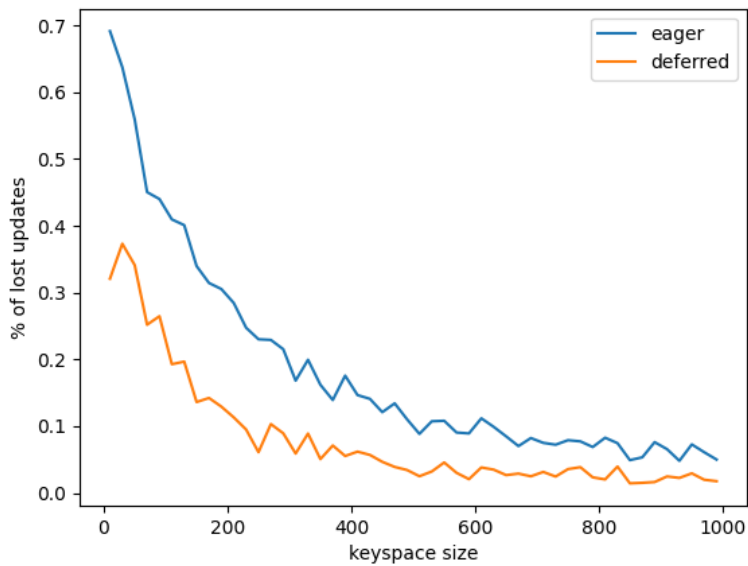


Fig. 2: Average and median read transaction latency increase by keyspace size for uncapped maximum latency. Zoomed in on the tail.



(a) Maximum latency capped at 10s before eager execution



(b) Maximum latency uncapped

Fig. 3: Percentage of incorrect reads by keyspace size.

7 Limitations and future work

Our experimental results were obtained through a simulation of database systems in Python. We acknowledge that real database systems are much more complex, providing various layers of abstraction over network, underlying filesystem structure and hardware, and thus may be affected by a plethora of external factors. This added system complexity may result in different — and perhaps contradicting — evidence than what we got through our simulation.

This suggests that a more robust benchmark of our proposed solution, involving real database system configurations, is an interesting topic for future work.

References

- [1] Bailis, P., Fekete, A., Ghodsi, A., Hellerstein, J.M., Stoica, I.: Scalable atomic visibility with RAMP transactions. *ACM Trans. Database Syst.* **41**(3) (jul 2016)
- [2] Adya, A.: Weak consistency: a generalized theory and optimistic implementations for distributed transactions. PhD thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science (1999)
- [3] Faleiro, J.M., Thomson, A., Abadi, D.J.: Lazy evaluation of transactions in database systems. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. SIGMOD '14*, New York, NY, USA, Association for Computing Machinery (2014) 15–26
- [4] Bailis, P., Davidson, A., Fekete, A., Ghodsi, A., Hellerstein, J.M., Stoica, I.: Highly available transactions: Virtues and limitations. *Proc. VLDB Endow.* **7**(3) (nov 2013) 181–192
- [5] Bailis, P., Fekete, A., Franklin, M.J., Ghodsi, A., Hellerstein, J.M., Stoica, I.: Coordination avoidance in database systems (extended version) (2014)
- [6] Shapiro, M., Preguiça, N., Baquero, C., Zawirski, M.: Conflict-free replicated data types. In: Défago, X., Petit, F., Villain, V., eds.: *Stabilization, Safety, and Security of Distributed Systems*, Berlin, Heidelberg, Springer Berlin Heidelberg (2011) 386–400
- [7] Pritchett, D.: Base: An acid alternative: In partitioned databases, trading some consistency for availability can lead to dramatic improvements in scalability. *Queue* **6**(3) (may 2008) 48–55
- [8] Kleppmann, M.: A critique of the CAP theorem. *CoRR* **abs/1509.05393** (2015)
- [9] DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Voshall, P., Vogels, W.: Dynamo: Amazon’s highly available key-value store. *ACM SIGOPS operating systems review* **41**(6) (2007) 205–220
- [10] Lakshman, A., Malik, P.: Cassandra: A decentralized structured storage system. *SIGOPS Oper. Syst. Rev.* **44**(2) (apr 2010) 35–40

- [11] Taft, R., Sharif, I., Matei, A., VanBenschoten, N., Lewis, J., Grieger, T., Niemi, K., Woods, A., Birzin, A., Poss, R., Bardea, P., Ranade, A., Darnell, B., Gruneir, B., Jaffray, J., Zhang, L., Mattis, P.: Cockroachdb: The resilient geo-distributed sql database. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. SIGMOD '20, New York, NY, USA, Association for Computing Machinery (2020) 1493–1509

Privacy-preserving mechanisms for graph databases: a preliminary study

Duarte Silva

Department of Computing Science
Umeå University, Sweden
dusi0002@ad.umu.se

Abstract. Privacy-preserving mechanisms for relational databases have been extensively studied but methods for graph databases have not. Among the privacy models, Differential Privacy has been considered a *de facto* standard mechanism. In this paper, one new privacy-preserving mechanism for graph databases called Cover-up is studied and it is proven that descriptive statistical queries applied on CU results are differentially private.

1 Introduction

Data privacy is an area of research that is concerned with protecting the privacy of individuals when a database is shared to other stakeholders. It can be said there are three different communities that work on data privacy from a technological perspective. One is the statistical disclosure control (associated with a statistical background), another is the privacy-preserving data mining (field that proceeds from databases and data mining), and finally the privacy-enhancing technologies (this community proceeds from communications and security).

Focusing on the privacy-preserving data mining (PPDM), multiple approaches have been suggested. These include k -anonymity, l -diversity, randomization, cryptographic techniques or differential privacy [1]. Typically, they are based on the concepts of privacy failure, the capacity to determine the original user data from the modified one, loss of information and estimation of the data accuracy loss [2].

Among the privacy models, Differential Privacy (ϵ -DP) has been considered a *de facto* standard mechanism for private data analysis for the past ten years, being widely used in academia as well as in industry [3]. Although currently ϵ -DP has been applied to more and more settings, initially its work was concentrated on data mining tasks, such as descriptive/summary statistical values [4], where there was a need to balance the goal of learning and releasing to the public statistical facts about some population without compromising the privacy of the individual respondents [5].

One can say that the data protection methods must adjust to the data representation. There are different approaches for how to represent data in a database. The most popular one currently being the relational databases (SQL), but other

representations do exist, such as NoSQL which includes graph databases but also other non-relational database types [6]. While there is a vast literature on privacy preserving techniques for SQL databases [7], the topic is less explored for NoSQL databases [8]. Most privacy models have been applied to data featuring graph structure [9, 10], but the application to graph databases is less investigated. However, Stokes addresses another privacy model for graph database models, called Cover-up (CU) [11].

In this paper, CU is presented, exemplified and applied on a small dataset. Since CU it is a PPDM mechanism it's important to understand if CU, in fact, produces privacy-preserving results. One way of doing this is proving if CU's results, for descriptive (or summary) statistical queries, are ϵ -DP or not, after the exemplification. This question is important because ϵ -DP has extensive literature about it, being an established privacy-preserving mechanism.

1.1 Preliminaries

The idea behind Differential Privacy (ϵ -DP) is that if the effect of making an arbitrary single substitution in the database is small enough, the query result cannot be used to infer much about any single individual, and therefore provides privacy. The primary aim is to provide global, statistical information about the data publicly available, while protecting those users privacy whose information is contained in the dataset [12].

Let a randomized algorithm $A : D \rightarrow S$, where D is a dataset and S come codomain. For a given $\epsilon > 0$, A is ϵ -DP if:

$$P(A(D_1) \in S') \leq e^\epsilon P(A(D_2) \in S') \quad (1)$$

for any set $S' \subseteq S$, and any two datasets $D_1, D_2 \subseteq D$ that differ by only one record (row).

One essential concept which is often associated with ϵ -DP is sensitivity [3]. Let $f : D \rightarrow R$, where R denotes real numbers, the sensitivity of a function (query in database) is:

$$\Delta f = \max_{x \in D_1, y \in D_2} |f(x) - f(y)| \quad (2)$$

Sensitivity is used with A . An usual A is the Laplace Mechanism [3], a randomised mechanism which adds Laplacian noise η , random variable which follows a Laplace distribution with parameters $\mu = 0$ and $b = (\frac{\Delta f}{\epsilon})$, denoted by:

$$A_L = f(x) + \eta \quad (3)$$

As for CU, its objective is to modify the label indicating the proportion of objects in the database with a given pair of properties. In this article, a dataset is a collection of properties of objects. For example, if the objects are the set of books in a library, then the types (or the attributes) of the data can be thought of as "title", "author", "number of pages", "shelf". The properties are the actual

titles, authors and so on. In Cover-up it is also assumed that every object can only have one property for each type.

To better understand CU, a definition of a data graph [11] is needed:

Definition 1 (Data graph). A data graph $G_D = (V, E)$ representing a dataset D with $s \in \mathbf{N}$ types is an s -partite graph and $n \in \mathbf{N}$ entries, such that the vertices in the partition V_r represent the properties $P = \{p_1, p_2, p_3, \dots, p_j\}$ with types $T = \{t_1, t_2, t_3, \dots, t_s\}$, with $j \in \mathbf{N}$.

The vertices are represented by $V = \{v_1, v_2, v_3, \dots, v_{s \times n}\}$, where each $v_{i'} = (t_{i'}, p_{j'})$. The data in D is represented by a set of labeled edges $E = \{(e_1; l_1), (e_2; l_2), \dots, (e_{n \times \frac{s(s-1)}{2}}; l_{n \times \frac{s(s-1)}{2}})\}$, where each $e_i = (v_{i'}, v_{i''})$, $i' \neq i''$, and $L = \{l_1, l_2, \dots, l_{n \times \frac{s(s-1)}{2}}\}$ is the set of labels. There is an edge e_i between two vertices $v_{i'}$ and $v_{i''}$, $i' \neq i''$, if there is some object with suitable¹ properties $p_{j'}$ with type t_d and $p_{j''}$ with type $t_{d'}$, and the edge is labeled with the proportion of objects having the pair of properties and types in question, $l_i = \frac{m(v_{i'}, v_{i''})}{n}$, where $m(v_{i'}, v_{i''})$ expresses the number of objects with properties $p_{j'}$ with type t_d and $p_{j''}$ with type $t_{d'}$.

Here is the definition of a protected data graph [11] G_D by CU, $G_D(P)$:

Definition 2 (Cover Up). A Cover-up privacy-preserving data graph with parameter k of the dataset D is the graph obtained from G_D using the following steps:

1. Represent each object in the dataset with a complete graph on the node set of pairs (type,property). The result is a set of disjoint cliques;
2. Identify all nodes with some given pair (t_i, p_j) ;
3. Merge all multiple edges into single edges;
4. Label every edge by $l_i = \frac{m(v_{i'}, v_{i''})}{n}$;
5. Pick an integer k (typically at least 3);
6. Modify the labels to $l'_i = \frac{m'}{n}$, where $m'_i = \arg \min_{ak} \{|m(v_{i'}, v_{i''}) - ak|\}$. In other words, m'_i is the positive integer multiple of k closest to $m(v_{i'}, v_{i''})$.
7. Remove all edges labelled with 0.

To clarify this definition, one can look at an example. Imagine a database with 3 types “name”, “age” and “profession” with 3 entries: [Johan, 40, Tailor], [Anne,

¹ This has to do with the context, reality of the dataset/database. For example, an edge labelled “practices” connecting vertexes representing, let us say, “dog” and “badminton” would be senseless.

34, Saleswoman] and [Sarah, 34, Gardener]. There are 6 nodes because there are 2 nodes for each type (3) and there are 9 edges. The edges would be then: (Johan, 40), (Johan, Tailor), (40, Tailor), (Anne, 34), (Anne, Saleswoman), (34, Saleswoman), (Sarah, 34), (Sarah, Gardener) and (34, Gardener). So, we can see that there would be two edges connected to the type "age" with property "34" ($m = 2$), (Anne,34) and (Sarah,34), which leads to the merging of the edges into ((Anne,Sarah),34), effectively creating a new node called (Anne,Sarah). Afterwards, this new edge would be labelled with $\frac{2}{3}$, as the total number of entries is 3 ($n = 3$).

2 Earlier Work

As stated in the Introduction, the field of privacy preserving techniques applied to graph databases is not very rich, but the one related to graph structured data has more research associated with it.

The first application of differentially private computation on graph data [13] was introduced in 2007. Authors showed an estimation of the cost associated with the computation of a minimum spanning tree in a differentially private manner. There are three main notions of ϵ -DP on graphs: edge-level ϵ -DP, node-level ϵ -DP and graph-level ϵ -DP. In 2012, researchers manage to exemplify an usage of edge-level ϵ -DP, using it to release edge level ϵ -DP sub-graph counts [14].

However, in 2013, it is published a paper [15] where it is shown that node level ϵ -DP offers a strictly stronger privacy guarantee than edge-level differential privacy, although being difficult to achieve when applied in certain data, such as descriptive statistics.

So far, graph-level ϵ -DP has been the least explored notion of ϵ -DP in graphs, among the three aforementioned ones. More recently however, in 2022, graph-level ϵ -DP has found applications in Machine Learning, in the field of Graph Neural Networks (GNNs) to be more exact. In [16] there is a implementation of graph-level ϵ -DP for classification tasks on several sensitive datasets, implementing the concept of graph-level ϵ -DP on GNNs and showing potential applications.

3 Exemplification of CU

Let us showcase CU in a more detailed way. Let us consider a toy database T with 3 types (for simplicity): song, disk and playlist name (see Table 1).

With this database in mind, let us run CU. Step 1 is eminently visual, but since this is a small database one can make a sketch, just to illustrate what is meant by this step (see Figure 1). Note that, for reading purposes solely, words such as Song was shortened to S, Disk to D and Playlist to P.

Step 2 asks to identify all nodes with the same pair. In this database, it would be for example (Song, Time), (Song, Eclipse), (Song, Airbag), (Song, Money) paired with (Disk, The Dark Side of the Moon) which means, in practical terms, that all of the above songs are in the "The Dark Side of the Moon" disk. Step

Song	Disk	Playlist
Time	The Dark Side of the Moon	British Songs
Eclipse	The Dark Side of the Moon	British Songs
Airbag	OK Computer	British Songs
Brain Damage	The Dark Side of the Moon	British Songs
Money	The Dark Side of the Moon	Prog Rock
Lucky	OK Computer	Prog Rock
Karma Police	OK Computer	Prog Rock
Let Down	OK Computer	Prog Rock
Exotica	Exotica	Happy Vibes
Puerto Rico	Vaya con Dios	Happy Vibes

Table 1. Entries of the toy database T .

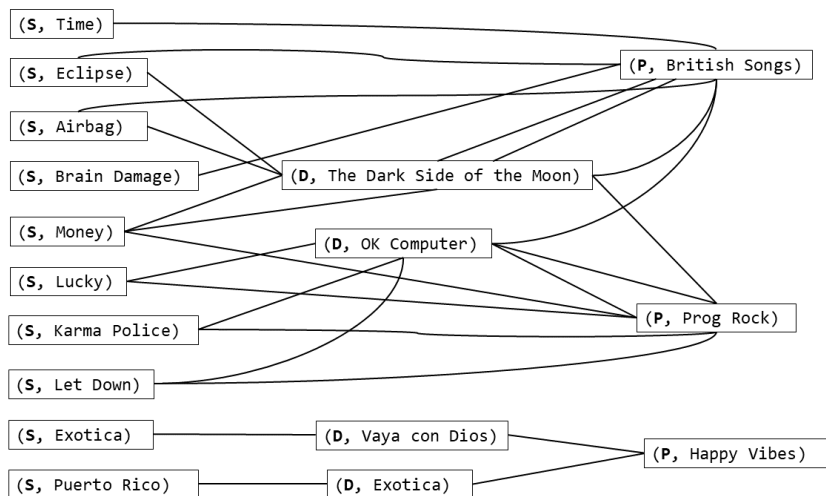


Fig. 1. One visual representation of Step 1 of CU applied on T .

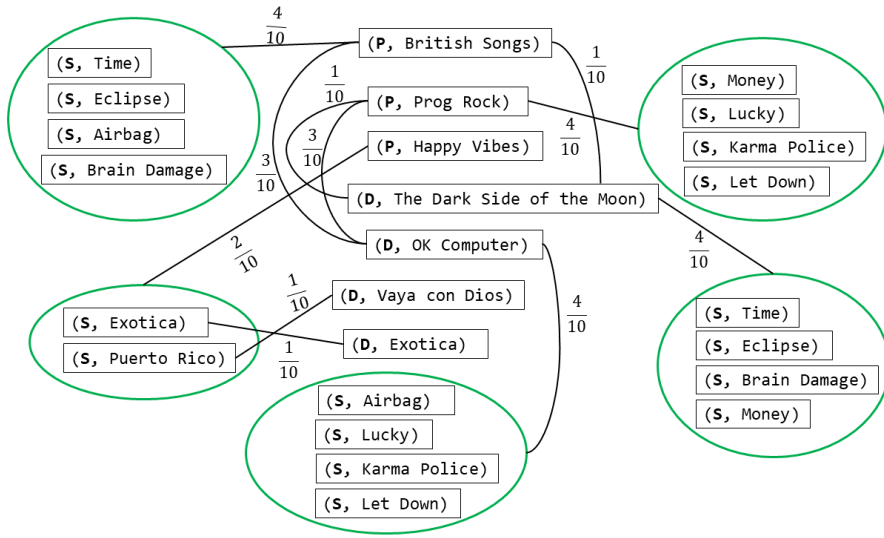


Fig. 3. One visual representation of step 4 of CU applied on T .

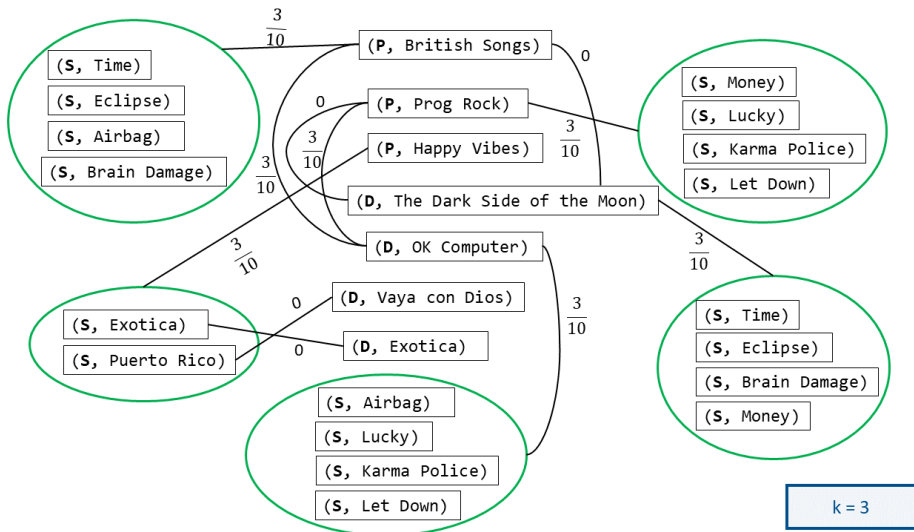


Fig. 4. One visual representation of the results of Steps 5 and 6 of CU applied on T .

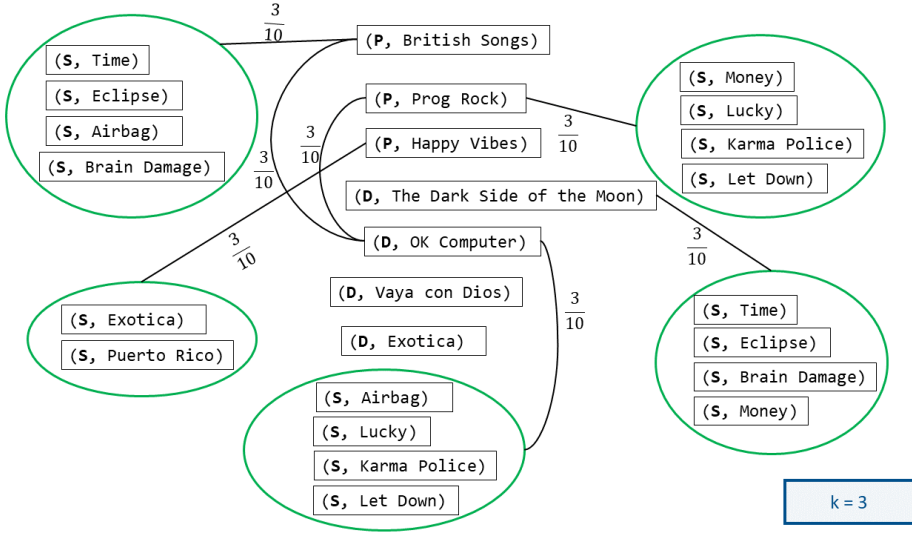


Fig. 5. One visual representation of T as a Cover-up privacy-preserving data graph.

Edge	Label
$[(\{(S, \text{Time}), (S, \text{Eclipse}), (S, \text{Brain Damage}), (S, \text{Money})\}); (D, \text{The Dark Side of the Moon})]$	$\frac{3}{10}$
$[(\{(S, \text{Airbag}), (S, \text{Lucky}), (S, \text{Karma Police}), (S, \text{Let Down})\}); (D, \text{OK Computer})]$	$\frac{3}{10}$
$[(\{(S, \text{Time}), (S, \text{Eclipse}), (S, \text{Brain Damage}), (S, \text{Airbag})\}); (P, \text{British Songs})]$	$\frac{3}{10}$
$[(\{(S, \text{Money}), (S, \text{Lucky}), (S, \text{Karma Police}), (S, \text{Let Down})\}); (P, \text{Prog Rock})]$	$\frac{3}{10}$
$[(\{(S, \text{Exotica}), (S, \text{Puerto Rico})\}); (P, \text{Happy Vibes})]$	$\frac{3}{10}$
$[(P, \text{British Songs}); (D, \text{The Dark Side of the Moon})]$	$\frac{3}{10}$
$[(P, \text{Prog Rock}); (D, \text{OK Computer})]$	$\frac{3}{10}$

Table 2. Final results of CU applied on T .

But, in the true database (see Table 1) both the song "Exotica" and "Puerto Rico" appear only once on the Playlist "Happy Vibes".

4 Descriptive statistical queries on CU results

Now, it is time to prove if, after applying CU to a given database, the results are ϵ -DP or not for a descriptive statistic query.

Theorem 1. *Let D be any database. Apply CU on D and name the final result $G_D(P)$. Let D' be a database differing D by one row. Apply CU on D' and name the final result $G_{D'}(P)$. For any descriptive statistic query f applied on $G_D(P)$ and $G_{D'}(P)$, there exists at least one ϵ such that results from f are ϵ -DP.*

Proof. Let a query f_1 : "Maximum label on the database".

Let D and D' be two databases which differ in one row, with the size of D being n ($|D| = n$) and $|D'| = n - 1$. Apply CU on both databases and denote l_{max}, l'_{max} as the maximum label on $G_D(P)$ and $G_{D'}(P)$, respectively. Take also $l'_{max} = l_{max} + c$, for some real-valued constant c .

Let us move to the randomised mechanism. Choose some labels λ, λ' such that $\Delta f_1 = \max_{\lambda \in D, \lambda' \in D'} |f_1(\lambda) - f_1(\lambda')|$ is verified. Let $A = A_L$:

$$\frac{P(A_L(D) \in S')}{P(A_L(D') \in S')} \leq e^\epsilon \quad (4)$$

The Laplacian probability cumulative density function has the form²:

$$P(X \leq x) = F(x; \mu, b) = \begin{cases} \frac{1}{2} \exp\left(\frac{x - \mu}{b}\right), & \text{if } x < \mu \\ 1 - \frac{1}{2} \exp\left(-\frac{x - \mu}{b}\right), & \text{if } x \geq \mu \end{cases} \quad (5)$$

Since $\mu = 0$ and x in our case is a label which is always non-negative, we always have that $x \geq \mu$:

$$\frac{F(l_{max}; 0, (\frac{\Delta f}{\epsilon}))}{F(l'_{max}; 0, (\frac{\Delta f}{\epsilon}))} \quad (6)$$

$$= \frac{1 - \frac{1}{2} \exp\left(-\frac{l_{max}}{(\frac{\Delta f_1}{\epsilon})}\right)}{1 - \frac{1}{2} \exp\left(-\frac{l'_{max}}{(\frac{\Delta f_1}{\epsilon})}\right)} \quad (7)$$

(plug in $l'_{max} = l_{max} + c$)

² www.randomservices.org/random/special/Laplace.html

$$= \frac{1 - \frac{1}{2} \exp\left(-\frac{l_{max}}{\left(\frac{\Delta f_1}{\epsilon}\right)}\right)}{1 - \frac{1}{2} \exp\left(-\frac{l_{max}+c}{\left(\frac{\Delta f_1}{\epsilon}\right)}\right)} \quad (8)$$

(for every non zero real number $x, y, z : \frac{x}{z} = \frac{yz}{y}$)

$$= \frac{1 - \frac{1}{2} \exp\left(-\frac{\epsilon l_{max}}{\Delta f_1}\right)}{1 - \frac{1}{2} \exp\left(-\frac{\epsilon(l_{max}+c)}{\Delta f_1}\right)} \quad (9)$$

$$\begin{aligned} & \left(\text{divide everything by } \frac{1}{2} \exp\left(-\frac{\epsilon l_{max}}{\Delta f_1}\right)\right) \\ &= \frac{2 \exp\left(\frac{\epsilon l_{max}}{\Delta f_1}\right) - 1}{2 \exp\left(\frac{\epsilon l_{max}}{\Delta f_1}\right) - \exp\left(-\frac{\epsilon c}{\Delta f_1}\right)} \end{aligned} \quad (10)$$

There are three cases to be considered: one when $c = 0$; another one with $c > 0$ and finally when $c < 0$. Let us start with $c = 0$:

$$\frac{2 \exp\left(\frac{\epsilon l_{max}}{\Delta f_1}\right) - 1}{2 \exp\left(\frac{\epsilon l_{max}}{\Delta f_1}\right) - 1} = 1 \leq e^\epsilon \quad (11)$$

Which is trivially true, since for any $\epsilon > 0$ this inequality holds. So, when $c = 0$, the query f_1 applied on $G_D(P)$ is ϵ -DP for all ϵ .

Now we want to prove, for $c > 0$, that (10) is smaller than 1 (the case when (10) is equal to 1 was already dealt with) which is smaller than e^ϵ for any ϵ . So, $\frac{-\epsilon c}{\Delta f_1} < 0$ necessarily. That is the case, because $\epsilon c > 0$ and $\Delta f_1 > 0$, implying $\frac{-\epsilon c}{\epsilon \Delta f_1} < 0$. So, for $c > 0$ the query f_1 applied on $G_D(P)$ is ϵ -DP for all ϵ as well.

Let us address $c < 0$ now. Since $c < 0$, $\exp\left(-\frac{\epsilon c}{\Delta f_1}\right) > 1$ hence (10) is also bigger than 1. This means that for $c < 0$ the query f_1 applied on $G_D(P)$ is not ϵ -DP for all ϵ . Is there any ϵ for which the query f_1 applied on $G_D(P)$, with $c < 0$, is ϵ -DP?

Let $\delta_1 = \frac{l_{max}}{\Delta f_1}$ and $\delta_2 = \frac{-c}{\Delta f_1}$. Both are positive. Plug this in (10):

$$\frac{2 \exp(\epsilon \delta_1) - 1}{2 \exp(\epsilon \delta_1) - \exp(\epsilon \delta_2)} \quad (12)$$

$$\leq \frac{2 \exp(\epsilon \delta_1) - 1}{2 \exp(\epsilon \delta_1)} \quad (13)$$

$$\leq 2 \exp(\epsilon \delta_1) - 1 \quad (14)$$

$$\leq 2 \exp(\epsilon \delta_1) \quad (15)$$

$$\leq \exp(\epsilon \delta_1 + \ln(2)) \quad (16)$$

$$\leq \exp(\epsilon^*) \quad (17)$$

So, there is an ϵ^* for which the query f_1 applied on $G_D(P)$, with $c < 0$, is ϵ -DP if $\epsilon^* \geq \epsilon \delta_1 + \ln(2)$.

The proof done above could be extended to other descriptive statistics. Meaning that if instead of querying the maximum label, one would like to query other descriptive statistic such as the first quartile label or the median label for example, the proof outline would be analogous with analogous conclusions, just changing the sensitivity function. This is the case because, one can always argue that some descriptive statistic applied on the original database (in this case it was l_{max} , but if one was querying the "median label" for instance, it could be l_{median}) can be thought as some descriptive statistic applied on the database that differs in one row of the original database plus some real-valued constant c . What can change, however, is the sensibility as the labels chosen to satisfy it may not be the same, depending on the descriptive statistic.

5 Conclusions and Future Work

A new model for representing data in terms of a graph in a privacy preserving manner was studied, this model's application being thoroughly exemplified on a small database. After this, it was proven that, for descriptive statistics queries, the results of CU are ϵ -DP for all ϵ , except when the maximum descriptive statistic of the original database D is bigger than the maximum descriptive of the D' , D and D' differing by one row. In this case, the bound on ϵ was tighter. However, it is unsure that the upper bound found for ϵ is the tight upper bound, that is, the majorant of ϵ . Further research can determine if the existing bound is the tight upper bound or not. On the other hand, more work can be done on CU. For instance, take advantage of the graph database and explore graph-related queries. Also, one can study some mechanism(s) to improve the current formulation of parameter k . Moreover, study how CU is affected by the existence of outliers and by the presence or lack of high heterogeneity in the database. It is also important to do a practical implementation of CU, implementing it on some programming language and applying it to some database.

One can also think of the possible reconfiguration of CU. In this paper, the objective was to understand if, for descriptive statistics queries, results of CU were indeed ϵ -DP or not. But, it is also interesting to understand if CU itself is ϵ -DP or not. However, in its present form, one can't evaluate this because CU is non-random. So, some randomness needs to be included in the method. A possible solution is called probabilistic privacy-preserving data graph [11], treating the labels not as simple frequencies but as random variables. But, it is still needed to study the compatibility of probability distributions on the edges of this new model and understanding how the disclosure risk can be defined for it.

References

- [1] Sachan, A., Roy, D., Arun, P.V.: An analysis of privacy preservation techniques in data mining. In Meghanathan, N., Nagamalai, D., Chaki, N., eds.: *Advances in Computing and Information Technology*, Berlin, Heidelberg, Springer Berlin Heidelberg (2013) 119–128
- [2] Xu, Z., Yi, X.: Classification of privacy-preserving distributed data mining protocols. In: 2011 Sixth International Conference on Digital Information Management. (2011) 337–342
- [3] Dwork, C.: Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., Wegener, I., eds.: *Automata, Languages and Programming*, Berlin, Heidelberg, Springer Berlin Heidelberg (2006) 1–12
- [4] Rugg, G.: *Using statistics: A gentle introduction* (2008)
- [5] Dwork, C., Smith, A.: Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality* **1**(2) (Apr. 2010)
- [6] Accottillam, T., KTV, R., G., R.: Survey on data management system for big data analytics. (2016)
- [7] Kotsogiannis, I., Tao, Y., He, X., Fanaeepour, M., Machanavajjhala, A., Hay, M., Miklau, G.: Privatesql: A differentially private sql query engine. *Proc. VLDB Endow.* **12**(11) (jul 2019)
- [8] Samaraweera, G.D., Chang, J.M.: Security and privacy implications on database systems in big data era: A survey, 33:1, pp. 239–258. ((2021))
- [9] Casas-Roma, J., Herrera-Joancomarti, J., Torra, V.: A survey of graph-modification techniques for privacy-preserving on networks. *Artificial Intelligence Review* **47**(3) (2016)
- [10] Stokes, K., Torra, V.: Reidentification and k-anonymity: a model for disclosure risk in graphs, *soft computing* 16:10, 1657-1670. (2012)
- [11] Stokes, K.: Cover-up: a probabilistic privacy-preserving graph database model. (2019)
- [12] Aldeen, Y.A., Salleh, M., Razzaque, M.A.: A comprehensive review on privacy preserving data mining. *SpringerPlus* **4**(1) (2015)
- [13] Nissim, K., Raskhodnikova, S., Smith, A.D.: Smooth sensitivity and sampling in private data analysis., *ACM* (2007) 75–84

- [14] Gupta, A., Roth, A., Ullman, J.R.: Iterative constructions and private data release. In: Theory of Cryptography - 9th Theory of Cryptography Conference, TCC 2012, Taormina, Sicily, Italy, March 19-21, 2012. Proceedings. Volume 7194 of Lecture Notes in Computer Science., Springer (2012) 339–356
- [15] Kasiviswanathan, S.P., Nissim, K., Raskhodnikova, S., Smith, A.: Analyzing graphs with node differential privacy. In Sahai, A., ed.: Theory of Cryptography, Springer Berlin Heidelberg (2013) 457–476
- [16] Mueller, T.T., Paetzold, J.C., Prabhakar, C., Usynin, D., Rueckert, D., Kaissis, G.: Differentially private graph classification with gnns. CoRR (2022)

The Uncanny Valley Effect in Zoomorphic Robots: Univariate analysis

Jiangeng Sun

Department of Computing Science
Umeå University, Sweden
`mrc20jsn@cs.umu.se`

Abstract. In human-computer interaction, the uncanny valley effect is widespread. It describes when the robot reaches the similarity of "almost" human, the human sentiment towards robots suddenly turns negative. The uncanny valley effect can cause users to have negative emotions toward robots. Therefore, it is necessary to avoid the uncanny valley effect as much as possible when designing the robot's appearance. This study will explore whether user groups with different characteristics (genders, ages, liking for animals, and interest in robots) have differences in their uncanny perception of the appearance of zoomorphic robots. After answering questions about their characteristics, one hundred sixteen participants rated the appearance of four robotic dogs with different animal similarities. It can be seen from the results that for the four robot dogs with different animal similarities, the participants' evaluation of appearance conforms to the U-shaped curve of the uncanny valley effect of zoomorphic robots. Moreover, age, gender, and liking levels for animals did not affect participants' uncanny perception of the appearance of the zoomorphic robot. However, participants with different interest levels in robots have different perceptions of the appearance of zoomorphic robots. Specifically, people who were more interested in robots scored higher on the robot's appearance, that is, the weaker their uncanny perception.

1 Introduction

The Uncanny Valley effect is a key factor affecting the interaction between humans, and robots [1]. Because of the similarities in appearance and behavior between robots and humans, humans often empathize with robots. Therefore, as robots are designed to look more and more like humans, humans will respond more and more positively to robots. However, once the robot reaches a certain level of human likeness, the human response suddenly becomes exceptionally negative. At this time, even if the robot's appearance is slightly different from that of a real human being, humans will feel that the robot looks uncanny and scary. And this negative effect will disappear when the similarity level continues to increase. That is called the uncanny valley effect. When designing a robot, the uncanny valley should be eliminated as much as possible to promote the user's satisfaction with the robot.

Research has shown that people of different ages [2] and genders ¹ have different perceptions of the uncanny valley effect of humanoid robots. Research on the uncanny valley effect inspires the design of industrial humanoid robots, it is necessary to specify different design strategies for different groups of people. However, robots are not just humanoid robots. They can also be zoomorphic robots. When humans interact with zoomorphic robots, the Uncanny Valley curve differs from humanoid robots. Therefore, exploring which characteristics will affect the user's perception of the uncanny valley effect of zoomorphic robots is also important.

This research was carried out in the form of a questionnaire survey. After answering questions about their characteristic (genders, ages, liking for animals, and interest in robots), the participants rated the appearance of four robotic dogs with different animal similarities. After the sample data was obtained, the uncanny valley curve for zoomorphic robots was confirmed. Then the statistical method used for data processing was determined through the normality test of the sample. Finally, the impact of different characteristics on the participants' ratings of the appearance of the zoomorphic robot was analyzed through this statistical method.

2 Earlier work

In 1970, Masahiro Mori, a robotics professor at the Tokyo Institute of Technology, first proposed the concept of the uncanny valley effect [1]. He pointed out that human responses to a humanoid robot may suddenly shift from empathy to revulsion as it approaches but fails to acquire a lifelike appearance.

In 2012, a project at the California Science Fair found that humans have gender differences in the uncanny valley effect¹. They found that women had better facial recognition rates and thus were more responsive to the uncanny valley effect.

In addition, research in 2020 found that humans have age-related differences in the uncanny valley effect [2]. The uncanny valley effect was found in young and middle-aged adults, but older adults do not show the uncanny valley effect for robots. Another article in March 2020 found that the Uncanny Valley curve of Zoomorphic robots is a U-shaped curve, not a cubic function like a humanoid robot [3]. However, after that, the hotspot of research is still humanoid robots, and there are still very few studies on the uncanny valley effect of animal-shaped robots. So far, the difference in human perception of the uncanny valley effect of zoomorphic robots is still a blank research spot.

From these previous studies, it is a natural idea that maybe humans still have age and gender differences in their perception of the uncanny valley effect of zoomorphic robots. At the same time, some reasonable guesses may be that people with different liking levels for animals and different interest levels

¹ <https://csef.usc.edu/History/2012/Projects/J0417.pdf>, gender differences in uncanny valley effect, accessed 2022-12-7.

in robots may perceive the uncanny valley effect differently. This study will verify these differences through the questionnaire survey and statistical analysis method.

3 Method

This study was done in the form of a social experiment. The experimental method is distributing questionnaires to people of different ages and genders. Then, after collecting sufficient questionnaires, process the samples through the SPSS platform. Questionnaires were distributed through a survey platform². Every volunteer using this platform could view and complete this public questionnaire. After completing the questionnaire, each volunteer was paid 0.5 RMB. It took three days, and finally, 418 questionnaires were received.

3.1 Quantitative survey

Questionnaires are used as a quantitative survey in this research. The advantage of quantitative surveys is that sufficient data can be collected quickly. Moreover, since participants do not need to fill in their personal information, they can objectively express their true inner thoughts. However, since the questionnaires are issued on the internet, the disadvantage is that the questionnaire response rate is low. Moreover, many questionnaires are filled carelessly and cannot be used as effective questionnaires. This phenomenon increases the data processing workload.

3.2 Questionnaire design

In order to ensure ecological validity, based on previous study [3], four robot dog pictures (Fig. 1) with different animal similarities levels were selected as stimulus materials, which were presented to participants in the online experiment.

After filling in the gender and age, the participants were asked to rate their liking level for animals and their interest level in robots using by Likert five-point scale. Then they were asked to choose which of the four robot dogs they liked the most and which they disliked the most and to rate the appearance of the four robot dogs in turn by a Ten-point Likert scale. Among them, a score of 0 means that the robot dog looks particularly weird, and a score of 10 means that it looks exceptionally adorable. The questionnaire is available on GitHub³.

3.3 Data preprocessing

Since the questionnaires were collected through online experiments in this study, it is not easy to ensure that the participants fill in the questionnaires carefully.

² <https://www.wjx.cn/>, 'Wen Juan Xing' survey platform, accessed 2022-12-7.

³ <https://github.com/hljmssjg/UVEsurvey>, project questionnaire, accessed 2022-12-7.



Fig. 1. Robot dogs with four different animal similarities (from left to right): Aibo (Sony), Animatronic Dog (Jim Henso Company), Animatronic Male Wolf (Sally Corporation), a real dog.

Therefore, the critical step is to preprocess the data before data analysis. Since the questionnaire for this research consists of ten questions, and the expected filling time is about one minute, any questionnaire that takes less than 30 seconds to fill will be considered invalid. In addition, questionnaires with illogical answers will also be considered invalid. The basic logic is that for the four robot dog pictures, it is impossible for the participants to like and dislike the same robot dog the most. In addition, the participant’s favorite and most disliked robot dogs should match subsequent ratings. For example, suppose the participant likes robot dog No. 4 the most and dislikes robot dog No. 2 the most. In that case, the participant should rate the appearance of robot dog No. 4 as the highest and the appearance of robot dog No. 2 as the lowest.

3.4 Data analysis

In this section, the first step is to check the normality of the sample. The process of data analysis is shown in Table 1. If the sample satisfies the preconditions for using a parametric test, then the parametric test should be used. Otherwise, choose the non-parametric test.

Table 1. Different methods to data analysis

Variable	Is the data normally distributed?	
	Yes	No
Age	Independent Samples t-test	Wilcoxon Mann Whitney test
Gender	One-way Anova	Kruskal-Wallis test
Love level for animals		
Interest level in robots		

4 Preliminaries

This study will use statistical methods to process the experimental data. There are several methods to determine differences between data. The general classifications of methods are parametric tests and non-parametric tests⁴.

⁴ <https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests>, non-parametric tests, accessed 2022-11-25.

4.1 Parametric test

The parametric test is one of the important parts of inferential statistics. When the data distribution is known (such as a normal distribution), parametric tests infer statistical parameters of the population distribution (such as the mean) from the sample data. When comparing means between two data sets, generally use an independent sample t-test. When the number of groups is larger than two, one-way ANOVA is used ⁴.

Although parametric tests are adequate for studying data differences, there are some prerequisites for the conditions of parametric tests. One of the prerequisites is that the data to be tested needs to be in the normal distribution.

4.2 Normality test

The normality of the data can be verified with a z-test [4] using the kurtosis and skewness of the curve. Let the sample's kurtosis be K , the skewness S , and their corresponding standard deviations SE_K and SE_S , respectively. Formulas to calculate z-score for kurtosis and skewness are shown in Equation 1:

$$\begin{aligned} Z_K &= \frac{K}{SE_K} \\ Z_S &= \frac{S}{SE_S} \end{aligned} \quad (1)$$

Statisticians pointed out that for a medium-sized sample ($50 < n < 300$), if the z-score doesn't satisfy that $Z_K \in [-3.29, 3.29]$ or $Z_S \in [-3.29, 3.29]$, which corresponds to a significance level alpha 0.05 and leads to the conclusion that the sample distribution is not normal.

However, computing z-scores is one of many criteria for normality testing. Generally, it needs to be judged by a combination of visual inspection (whether the data is bell-shaped distribution), calculation of z-score, and other methods.

4.3 Non-parametric

When the prerequisites (mentioned in Section 4.2) for parametric tests are not satisfied, it is necessary to replace the parametric test with a non-parametric one.

Mann-Whitney U Test Wilcoxon first proposed the Mann-Whitney U test in 1945 [5]. When the two sets of data don't satisfy normality and homogeneity of variances, choose this method to analyze the data by sacrificing the test efficiency. The hypotheses⁵ of this test are shown in the Definition 2:

$$\begin{aligned} H_0 &: \text{The two samples are in the same overall distribution.} \\ H_A &: \text{The two samples are in different overall distributions.} \end{aligned} \quad (2)$$

⁵ https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/bs704_nonparametric4.html, Mann-Whitney U test, accessed 2022-11-25.

The way to test the null hypothesis is to compute the rank sum, where the rank means the ranking of data values. Consider two samples D_1 and D_2 . After mixing and sorting the values of two samples, the rank of each sample is obtained. In particular, when two or more data values are equal, take the average rank. Finally, the rank sum of data set D_1 is T_1 , and the rank sum of data set D_2 is T_2 . The total data in these two sets is N , and the relationship between these two rank sums is shown in Equation 3.

$$T_1 + T_2 = \frac{N(N+1)}{2} \quad (3)$$

Assume the size of sample D_1 is n_1 , and the size of sample D_2 is n_2 , respectively. Denoted U as the test statistic for Mann-Whitney U Test, which is the smaller value between U_1 and U_2 . The formulas for calculating U is shown in Equation 4:

$$\begin{aligned} U_1 &= n_1 n_1 \frac{n_1(n_1+1)}{2} - R_1 \\ U_2 &= n_1 n_2 \frac{n_2(n_2+1)}{2} - R_2 \end{aligned} \quad (4)$$

Statisticians have shown a distribution table corresponding to the critical value of U ⁵. Assume the two-sided significance level α is 0.05, and define the sample size $N = n_1 + n_2$. Then the critical value of U can be obtained by looking up the table according to N and α . If the observed value of U is greater than the critical value of U , then the null hypothesis is retained. On the other hand, if the observed value of U is less than the critical value of U , then the null hypothesis can be rejected, and the alternative hypothesis is selected.

Kruskal-Wallis test Kruskal and Wallis first proposed the Kruskal-Wallis test in 1952 [5]. Generally, choose this method when the samples don't satisfy normality and homogeneity of variances. Kruskal Wallis test can be regarded as an extension of the Mann-Whitney test. The null hypothesis H_0 and alternative hypothesis H_A ⁶ is shown in Definition 5:

H_0 :The independent samples all have the same central tendency and are from the same population.

H_A :At least one of the independent samples does not have the same central tendency as the other samples.

(5)

Like the Mann-Whitney test, obtain the ranks in each sample by mixing and sorting. Assume that the null hypothesis is true. That is, there is no difference between the data of each group. Then the probability of rank to each group should be the same. Define the total size of sample is N , the formulas for calculating the expected value of rank sums E_R and rank variance σ^2 is shown in

⁶ <https://datatab.net/tutorial/kruskal-wallis-test>, Kruskal-Wallis test, accessed 2022-11-25.

Equation 6:

$$\begin{aligned} E_R &= \frac{N+1}{2} \\ \sigma^2 &= \frac{n^2-1}{12} \end{aligned} \quad (6)$$

Assuming there are i groups of data to be tested, the degree of freedom df is $i-1$. So the mean rank-sum in group i is \bar{R}_i . Then the formulas to calculate the H statistic for the KW test is shown in Equation 7:

$$H = \frac{n-1}{n} \cdot \sum_{i=1}^k \frac{n_i(\bar{R}_i - E_R)}{\sigma^2} \quad (7)$$

The critical H value can be obtained from the table of critical χ^2 values according to the degree of freedom df and the significance level (generally set to 0.05). Retain the null hypothesis if the calculated H value is larger than its critical value.

5 Results

After filtering out invalid questionnaires through data preprocessing, a total of 116 valid samples of the initial 418 questionnaires remained to be used in the analysis. The corresponding data is available on GitHub⁷.

5.1 Frequency analysis

Analyze the demographic characteristics of valid questionnaire samples to obtain the number and percentage of sample cases. The option with a high proportion indicates the degree of the tendency of the population. Each variable and its proportion are shown in Table 2.

Considering the variable of gender, 54 men and 62 women participated in this study, and the ratio of men to women is close to 1. For the age variable, the number of samples under the age of 20 is small, with only 3 cases accounting for 2.6 percent of the total sample size. Besides these, most people showed a relatively positive tendency toward the liking level of animals and the interest level of robots.

5.2 The uncanny valley effect

The distribution of the participants' favorite and most disliked robot dog is shown in Fig. 2.

No. 4 was most liked by 46.6% of the participants, followed by No. 1, which 37.1% most liked. No. 2 was disliked most liked by 67.2% of the participants,

⁷ https://github.com/hljmssjg/UVEsurvey/blob/master/195543715_418_116.sav, project data, accessed 2022-12-7.

Table 2. The result of frequency analysis

Variable	Category	Frequency	Percent
Gender	Male	54	46.6
	Female	62	53.4
Age	Under 20	3	2.6
	21-30	34	29.3
	31-40	18	15.5
	41-50	38	32.8
	Over 51	23	19.8
How much do you like animals?	Not at all	9	7.8
	Low level	9	7.8
	Neutral	43	37.1
	A little bit	38	32.8
	Very much	17	14.7
How much do you like robots?	Not at all	5	4.3
	Low level	6	5.2
	Neutral	45	38.8
	A little bit	28	24.1
	Very much	32	27.6

followed by No. 1, which was disliked most by 21.4%. It can be seen from Fig. 2 that the uncanny valley perception of human beings to the appearance of animal robots conforms to the U-shaped curve. Apart from the actual animal (No. 4), the robot dog with the lowest animal similarity (No. 1) was most often the participants' favorite. Therefore, these participants' perception of the uncanny valley effect of robot dog No. 1 is particularly interesting. The following explores how this perception varies for people with different characteristics. The results are shown in Table 3.

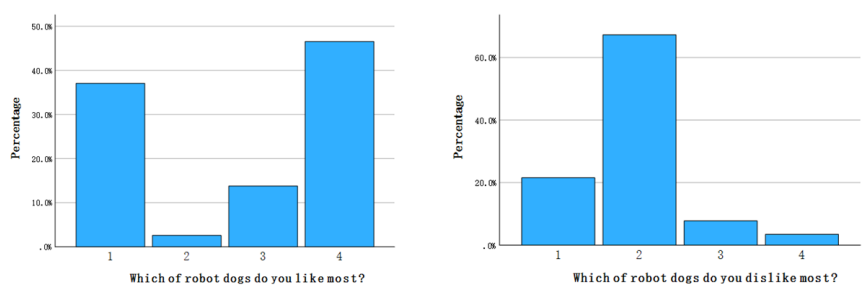


Fig. 2. Favorite and the most disliked robot dog

5.3 Normality test

Using SPSS software to verify the normality of the sample, the results are shown in Table 4.

From equation 1, the Z-score results of the kurtosis and skewness of this sample are -2.04 and -2.70 , respectively. Both absolute z-scores of kurtosis

Table 3. Frequency analysis

Variable	Category	Frequency	Percent
Gender	Male	54	46.6
	Female	62	53.4
Age	Under 20	3	2.6
	21-30	34	29.3
	31-40	18	15.5
	41-50	38	32.8
	Over 51	23	19.8
Liking levels for animals	Not at all	9	7.8
	Low level	9	7.8
	Neutral	43	37.1
	A little bit	38	32.8
	Very much	17	14.7
Interest levels in robots	Not at all	5	4.3
	Low level	6	5.2
	Neutral	45	38.8
	A little bit	28	24.1
	Very much	32	27.6

Table 4. Normality analysis

Statistics		
The impression of No. 1 robot dog		
N	Valid	116
	Missing	0
Skewness		-0.606
Std. Error of Skewness		0.225
Kurtosis		-0.909
Std. Error of Kurtosis		0.446

and skewness are less than 3.29, which means that the sample is in the normal distribution from the point of kurtosis and skewness.

The histogram of this sample and the estimated normal distribution curve are shown in Fig. 3.

The distribution of the samples is not bell-shaped, and the peak of the estimated curve does not match the maximum frequency of the samples. Therefore, it cannot be concluded that the sample is in the normal distribution.

5.4 Non-parametric test

Since the sample is not in the normal distribution, a rational choice is to use non-parametric tests.

Mann-Whitney U test In order to test whether people of different genders scored differently on the appearance of the No.1 robot dog, the results of the Mann-Whitney U test are shown in Fig. 4.

The significant difference in the ratings of the robot’s appearance by people of different genders is 0.394, which is larger than the significant level of 0.05, therefore retaining the null hypothesis (The distribution of ratings on the appearance of robot 1 was the same across genders).

Kruskal-Wallis test In order to test whether people of different ages scored differently on the appearance of the No.1 robot dog, the result of the Kruskal-Wallis test is shown in Fig. 5.

The significant difference in the ratings of the robot’s appearance by people of different ages is 0.228, which is larger than the significant level of 0.05, retaining the null hypothesis (Independent samples of different ages all have the same central tendency).

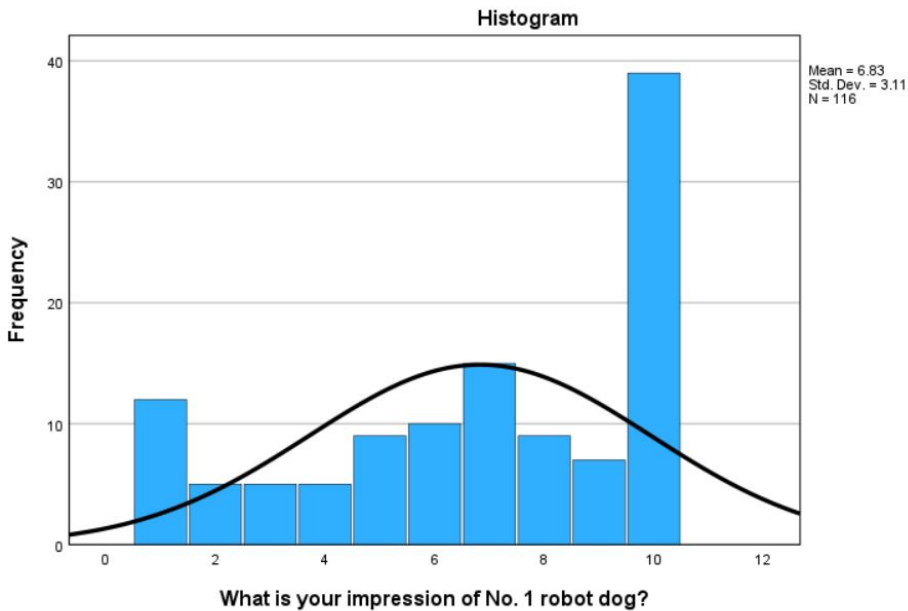


Fig. 3. The histogram of sample with estimated normal curve

Hypothesis Test Summary				
	Null Hypothesis	Test	Sig. ^{a,b}	Decision
1	The distribution of What is your impression of No. 1 robot dog? is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.394	Retain the null hypothesis.
a. The significance level is .050.				
b. Asymptotic significance is displayed.				

Fig. 4. Mann Whitney U test: gender

Hypothesis Test Summary				
	Null Hypothesis	Test	Sig. ^{a,b}	Decision
1	The distribution of What is your impression of No. 1 robot dog? is the same across categories of Age.	Independent-Samples Kruskal-Wallis Test	.228	Retain the null hypothesis.
a. The significance level is .050.				
b. Asymptotic significance is displayed.				

Fig. 5. Kruskal-Wallis test: age

In order to test whether people with different animal liking levels scored differently on the appearance of the No.1 robot dog, the result of the Kruskal-Wallis test is shown in Fig. 6.

Hypothesis Test Summary				
	Null Hypothesis	Test	Sig. ^{a,b}	Decision
1	The distribution of What is your impression of No. 1 robot dog? is the same across categories of How much do you like animals?.	Independent-Samples Kruskal-Wallis Test	.766	Retain the null hypothesis.
a. The significance level is .050.				
b. Asymptotic significance is displayed.				

Fig. 6. Kruskal-Wallis test: animal liking level

The significant difference in the ratings of the robot’s appearance by people with different animal liking levels is 0.766, which is larger than the significant level of 0.05, retaining the null hypothesis (Independent samples of different animal liking levels all have the same central tendency).

In order to test whether people with different robot interest levels scored differently on the appearance of the No.1 robot dog, the result of the Kruskal-Wallis test is shown in Fig. 7.

Hypothesis Test Summary				
	Null Hypothesis	Test	Sig. ^{a,b}	Decision
1	The distribution of What is your impression of No. 1 robot dog? is the same across categories of How much do you like robots?.	Independent-Samples Kruskal-Wallis Test	.030	Reject the null hypothesis.
a. The significance level is .050.				
b. Asymptotic significance is displayed.				

Fig. 7. Kruskal-Wallis test: robot interest level

The significant difference in the ratings of the robot’s appearance by people with different robot interest levels is 0.030, which is less than the significant level of 0.05. Therefore, the null hypothesis is rejected. Hence, people with different robot interest levels also scored differently on the appearance of the No.1 robot dog. The mean value of each group’s ratings on the appearance of robot No. 1 is shown in Table 5.

Table 5. The result of homogeneous subsets

Mean values of each group					
Level of interest in robots	Not at all	Low level	Neutral	A little bit	Very much
Mean value	7.2	3.17	6.37	7.62	7.41

In particular, for people with different robot interest levels, the rank sum distribution of each group of data is shown in Table 6.

Table 6. The result of homogeneous subsets

Based on interrst levels in robots			
		Subset	
		1	2
Sample	low level	23.417	
	neutral	53.422	53.422
	not at all	63.400	63.400
	very much		64.203
	a little bit		66.786
Test Statistic		5.947	3.742
Sig. (2-sided test)		0.051	0.291
Adjusted Sig. (2-sided test)		0.084	0.291
The significance level is 0.050.			

The rank sums of ratings on the appearance of robot No.1 by participants with different degrees of interest in robots are 23.417, 53.422, 63.400, 64.203, and 66.786. The higher the rank sum value, the higher the participants’ appearance ratings. SPSS software automatically corrects the value of significance(Sig) through Bonferroni correction(Adjusted Sig)⁸. The rank sums of the different samples are divided into two subsets. The test statistics are 5.947 and 3.742, respectively. The significance of the data between different subsets is less than 0.05. That is, there is a 95% confidence that the data in the two subsets have a statistically significant difference. However, the data in the same subset have no difference, and the corresponding significance values(Adjusted Sig) are 0.084 and 0.291, respectively. The samples of subset 1 are less interested in robots than the samples of subset 2. In conclusion, people with a higher degree of interest in robots(Subset 2) have a higher rank sum of evaluations on robot No. 1.

6 Discussion

This study provides insights into how to avoid the uncanny valley effect in the design of zoomorphic robots. Even though the results of this study may have limitations due to the number of participating samples, it reveals the underlying mechanism of the perception of the uncanny valley effect by the population with different characters. Firstly, among zoomorphic robots with different animal similarities, apart from real animals, the robot dog with the lowest animal similarity was the most preferred by the participants. This result fitted a U-shaped curve. Secondly, for this type of robot dog, the experimental results show that, unlike humanoid robots, gender, age, and liking for animals do not affect participants’ perception of the uncanny valley effect. However, interest in the robot will affect that. Specifically, people more interested in robots have a higher evaluation of robot dogs. Therefore, when the factories design a zoomorphic robot, they should determine its appearance according to the user’s interest. Future research should

⁸ https://en.wikipedia.org/wiki/Bonferroni_correction, explanation of Bonferroni correction in Wikipedia, accessed 2022-12-7.

further understand how other single factors in the population perceive the uncanny valley of zoomorphic robots and how multiple factors in the population perceive the uncanny valley of zoomorphic robots.

References

- [1] Mori, M., MacDorman, K.F., Kageki, N.: The uncanny valley [from the field]. *IEEE Robotics and Automation Magazine* **19**(2) (2012) 98–100
- [2] Tu, Y.C., Chien, S.E., Lai, Y.Y., Liu, J.C., Yeh, S.L.: The uncanny valley revisited: Age-related difference and the effect of function type. *Innovation in Aging* **3**(Supplement 1) (2019) S330–S330
- [3] Löffler, D., Dörrenbächer, J., Hassenzahl, M.: The uncanny valley effect in zoomorphic robots: The u-shaped relation between animal likeness and likeability. In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. HRI '20*, New York, NY, USA, Association for Computing Machinery (2020) 261–270
- [4] Kim, H.Y.: Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restor Dent Endod* **38**(1) (2013) 52–54
- [5] Hettmansperger, T.P., McKean, J.W.: *Robust Nonparametric Statistical Methods*, Second Edition. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis (2010)

Author Index

Engberg, Tilda, 1	Martens, Willeke, 35
Lindoff, Johanna, 11	Paes, Lucas, 69
Müller, Salome, 49	Silva, Duarte, 83
Marklund, Jakob, 23	Sun, Jiangeng, 97