

---

# A task-based evaluation of algorithms that generate quantifiable expressions

---

*Author:*  
Ruby PEL (5853419)

*Supervisor:*  
prof. dr. C.J. VAN DEEMTER  
*Second corrector:*  
dr. R.W.F. NOUWEN

Bachelor's thesis 7.5 ECTS  
Artificial Intelligence  
Utrecht University  
July 2020



**Utrecht University**

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Theoretical background</b>	<b>3</b>
2.1	Quantified expressions . . . . .	3
2.2	QTUNA corpus . . . . .	3
2.3	Quantified Description Generation algorithms . . . . .	3
2.3.1	Incremental Algorithm for Generating Quantified Descriptions . . . . .	4
2.3.2	Greedy Algorithm for Generating Quantified Descriptions . . . . .	4
2.3.3	Evaluation by Human Judgements . . . . .	4
<b>3</b>	<b>Methodology</b>	<b>6</b>
3.1	Methodology original experiment . . . . .	6
3.2	Methodology task-based evaluation . . . . .	6
3.2.1	Participants . . . . .	6
3.2.2	Materials . . . . .	6
3.2.3	Procedure . . . . .	6
3.2.4	Pilots . . . . .	7
3.3	Hypotheses . . . . .	7
3.4	Analysis of the data . . . . .	8
<b>4</b>	<b>Results</b>	<b>9</b>
4.1	Swaps . . . . .	9
4.2	Variance in answers . . . . .	9
4.3	Correlation experts' scores and reconstructions . . . . .	10
<b>5</b>	<b>Discussion and conclusion</b>	<b>12</b>
5.1	Evaluation of hypotheses . . . . .	12
5.2	Research question . . . . .	13
5.3	Further research and limitations . . . . .	13
<b>6</b>	<b>Bibliography</b>	<b>15</b>
<b>7</b>	<b>Appendices</b>	<b>16</b>
7.1	Appendix A: Participants . . . . .	16
7.2	Appendix B: Instructions . . . . .	17
7.3	Appendix C: Answering sheet . . . . .	18

# 1 Introduction

Natural language generation is a common term in the field of artificial intelligence. It is "the process of producing meaningful phrases and sentences in the form of natural language." (Ghosh and Gunning, 2019) and can either be in speech or writing, while it mostly refers to the generation of printed text by computers. Various types of generators exist; from simple template-based to "intelligent" generators using machine learning in order to have a better understanding of human-written texts Perera and Nand (2017).

Within the various formal theories of meaning and language use, a considerable amount of research has been done to referring expressions (e.g. expressions as "those girls"). To shine a light on other theories, recent research has been done on the meaning and use of quantified expressions. These expressions say something about the number of things having a particular property, such as "all  $A$  are  $B$ " or more vague sentences as "some  $A$  are  $B$ ". This research by Chen et al. (2019b) focused on descriptions of simple scenes populated by geometrical figures, and produced a corpus consisting of human-made quantified descriptions in English. Based on this corpus, two language generation algorithms were made to mimic these human descriptions and were then evaluated by three human experts; the experts rated the descriptions generated by both the algorithm and human speakers in terms of their naturalness, informativity and correctness.

The current project the researchers are working on, asks how useful the descriptions in question are for human readers. In particular, how well (human) participants are able to reconstruct the original situations from a description. The experts' ratings give an indication, but the addition of a task-based evaluation of the algorithm helps to answer the question with more certainty.

My addition to this project is to perform this task-based evaluation by conducting and analysing an experiment with twenty human participants, where reconstruction of the original scene is the task that the participants perform. The analysis focuses on the comparison between the task-based evaluation and the judgements of the human experts. The question I, therefore, want to answer with this research is: "To what degree do the judgements by the experts on the quantified description making algorithms comply with the results from the task-based evaluation on these algorithms?"

To answer this question, I will first provide more information about quantified expressions, the QTUNA corpus, and the original experiments in section 2. In section 3, I will explain the methodology of the task-based evaluation, and I will then present these results in section 4. Finally, I will analyze, discuss and draw conclusions from these results in section 5.

## 2 Theoretical background

### 2.1 Quantified expressions

In logical languages, quantified expressions are operators that bind variables to a formula. The most common logical quantifiers are  $\forall$  ("for all") and  $\exists$  ("there exists some"). In natural language, a quantifier is sort of similar; it tells us something about the number of things having a certain property. Take, for example, the following sentences, in which the italicized words are quantifiers: "All cups are red", "Some people were late". However, in natural language there are many more expressions that are considered quantifier expressions; not just *all* and *some*, but also *few*, *most*, *everything* and more (Glanzberg, 2009). On top of that, the grammatical structure can overshadow the logical structure in a natural language sentence. This makes it much harder to study the meaning of a quantifier expression can be ambiguous in natural language. For example, in the sentence "Someone trips over a banana peel in the supermarket every week" it is not clear whether it is the same person slipping on the banana peel every week.

### 2.2 QTUNA corpus

To understand how speakers use quantified expressions, the researchers Chen et al. (2019b) asked participants to describe a visual scene. Each scene shows a set number of objects, which can either be a circle or a square and either blue or red. An example of such a scene can be found in figure 1 (Chen et al., 2019b). The participants were instructed to give a description so that a reader would be able to *reconstruct* the situation, i.e. tell the number of objects when they are given the domain size, possible shape and colour (the location was irrelevant in this case). This was then done for three different domain sizes ( $n$ ), namely 4, 9, and 20, to determine how the size would influence the human production of QEs. Every domain size consists of 10 different scenes. Examples of the descriptions are:

$n = 4$	<i>There are 4 squares. Every object is blue.</i>
$n = 9$	<i>Most of the items are red circles, but there are a couple of blue squares.</i>
$n = 20$	<i>All the objects are blue squares. A few objects are blue circles.</i>

The experiments resulted in the QTUNA corpus, which contains 656, 380, and 378 valid descriptions for every domain size given by students at the Computing department of Utrecht University.

The analysis was performed based on of three hypotheses that were formed before the experiment. This resulted in the following conclusions:

1. The larger the domain size, the more vague quantifiers were used;
2. For smaller domains, more logically complete descriptions were given / for larger domains, less logically complete descriptions were given (in proportion);
3. The length of descriptions decreased with domain size (contrary to the hypothesis);
4. Shape occurs more often in the first argument place and colour in the second argument place.

### 2.3 Quantified Description Generation algorithms

The QTUNA corpus was then used to design two NLG algorithms that generate quantified descriptions and are able to carry out the same task as the human participants in the QTUNA experiment. However, the algorithms should be able to do this for any (reasonable) domain size and not just  $n = 4, 9, \text{ and } 10$ . An exception was made for domain sizes smaller than 4 and those for which it is impossible to count the objects in a few seconds.

Both algorithms make use of the following general framework. The *generator* is given a target scene with its domain knowledge: this is a list of possible attributes (shape/colour) with their possible values (square, circle / red, blue). The generator then calls the algorithm to construct a description containing an ordered sequence of QEs in a logical form. It does so by selecting from a set of candidate patterns, based on how human beings did so in the QTUNA experiment. The candidate patterns contain a quantified pattern, for example  $All(\cdot, \cdot)$ , and a property tuple that is able to fill in the slots of this pattern, for example  $(blue, square)$ . Lastly, the algorithm calls a simple template-based surface realiser that maps the description, which is still in logical form, into natural language text. This would result in, for example, "Every blue object is square."

### 2.3.1 Incremental Algorithm for Generating Quantified Descriptions

The first algorithm is based on two observations of the QTUNA dataset: some quantifier patterns are more frequent than others, and some choices of properties to fill a given pattern are more frequent than others. This algorithm, therefore, maintains a sequence of properties and a sequence of fillers, so that it can mimic the order of different types of statements humans use. The properties are the features that an object can contain; which are a certain shape and colour. The sequence of properties is inspired by the fourth conclusion on the QTUNA corpus mentioned in the previous paragraph: the description of the shape usually comes before the description of the colour, thus the algorithm should check if it can do this as well. The sequence of quantifiers is based on the finding that humans incline to start describing the scene as a whole. The algorithm should therefore give QEs such as *all*, *half* and *most* priority.

The algorithm uses a similar method of generation as the Incremental algorithm by Dale and Reiter (1995) and is therefore called the *Incremental QDG algorithm* (QDG-IA). It generates a description by going through all the QE patterns in the order of the quantifier preference order described above and considers them one by one in this order. To prevent that certain quantified patterns with low preference will never be chosen, a probability of 0.1 was added which the algorithm can use to make a single move of a quantified pattern with low preference into a higher preference order.

### 2.3.2 Greedy Algorithm for Generating Quantified Descriptions

Another perspective on the generation of quantified descriptions is considering it as the problem of searching for the best set of QEs. A greedy algorithm can separate the best set of QEs from the largest number of distractors (poorer sets of QEs) in each iteration. The number of distractors a set of QEs can eliminate is called the discriminatory power. The second algorithm that was made does this and is called QDG-GREEDY. To ensure variation, the sets of QEs with the same and highest discriminatory power will be randomly selected.

### 2.3.3 Evaluation by Human Judgements

To further test how informative, correct and humanlike the generated descriptions by the algorithms are, four academics from Utrecht University rated them on these criteria. Instead of mentioning these criteria by name, the "judges" were asked the following questions:

- Q1 (Naturalness): On a scale of 1-5, how likely do you think it might be that this description was uttered by a human? [1=very unlikely, 5=very likely]
- Q2 (Informativity): On a scale of 1-5, do you believe the description is as informative as it can be expected to be? [1= description isnt even nearly informative enough, 5= description gives as much information as possible]

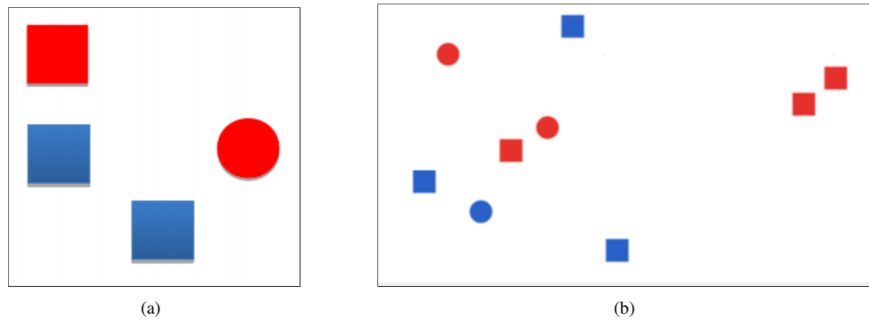


Figure 1: Examples from (a) the  $n = 4$  experiment; (b) the  $n = 9$  experiment (Chen et al., 2019b)

	Model	Naturalness	Informativity	Correctness
Experiment A	Human	3.45	4.05	4.6
	QDG-IA	2.85	3.95	4.55
	QDG-GREEDY	3.45	3.8	4.8
Experiment B	QDG-IA	3.7	3.8	4.83
	QDG-GREEDY	3.46	4.2	4.83

Table 1: Average scores for each algorithm and for human-produced descriptions, by naturalness, informativity, and correctness as annotated by the four human judges in Chen et al. (2019a)

- Q3 (Correctness): On a scale of 1-5, how correct do you consider this description to be? [1= the description is not at all correct, 5=everything the description says is correct.] (Chen et al., 2019a)

These questions were asked for 30 scene-description pairs in total for the original domain sizes  $n = 4, 9$ , and 20. Additionally, a total of 66 scene-description pairs for new domain sizes  $n = 6, 10$ , and 16 were judged. For a more detailed explanation of the experiment, see section 3.1.

Before the experiment with the judges, the researchers again came up with several hypotheses. Based on the results, they reached the following conclusions:

1. Humans and QDG-GREEDY performed similarly at naturalness, QDG-IA did slightly worse;
2. Contrary to expectations, the algorithms did not perform better at informativity, and QDG-IA did not perform better at correctness than humans. You would expect them to actually perform better at both these criteria since they are designed to optimise them. However, this could be due to the fact that the algorithms take both the semantics and pragmatics into account when judging the logical completeness of the descriptions, which the experts might disagree with;
3. Both algorithms performed equally as well on naturalness (there was no significant difference between their performance).

The results of the experiment can be found in Table 1. For my research project, I am mostly interested in the informativity and correctness scores given by the experts. You would expect the *reconstruction* of a scene to be better when these scores are higher. My hypotheses, which are based on the results of this experiment, will be discussed in section 3.3.

## 3 Methodology

### 3.1 Methodology original experiment

The original experiment by Chen et al. (2019a) was divided into two parts: experiment A and experiment B. For experiment A, they randomly selected 3 or 4 scenes from a total of 10, from each of the 3 sub-corpora of QTUNA. This resulted in 30 scene-description pairs in total. The sub-corpora each contain descriptions for 10 different scenes, each for a different domain size:  $n=4$ ,  $n=9$  and  $n=20$ . Each scene was then paired with one description by QDG-IA, one by QDG-GREEDY and one by a human being (which was randomly selected from the QTUNA corpus).

To test the generality of the algorithms they also carried out experiment B, in which three new domain sizes were the focal points:  $n=6$ ,  $n=10$  and  $n=16$ . There were 6 scenes sampled for each domain size  $n$ . Since descriptions by humans were not available for these domains, these 6 scenes were each paired with just two descriptions: one by QDG-IA and one by QDG-GREEDY.

Four experts (academics from Utrecht University) were asked to participate in both experiments to gain insight into the quality of the generated descriptions. They had to judge the 3 or 4 scenes in experiment A and 33 scene-description pairs each (from 66 in total) in experiment B, based on their naturalness, informativity and correctness.

### 3.2 Methodology task-based evaluation

To gain more insight into how useful the generated descriptions are for human readers, I designed a similar experiment in which human participants have to reconstruct the original scenes from these descriptions. Reconstruct means that the subjects have to state how many of the objects are either a circle or a square, and how many are either red or blue based on a scene’s description. The location an object does not matter.

#### 3.2.1 Participants

I recruited 20 students from Utrecht University with different educational backgrounds to participate in this task-based evaluation of both algorithms; 13 out of these 20 are or were a BSc / MSc student Artificial Intelligence (more detailed information can be found in subsection 7.1 in the appendix). The mean age of the participants was 22.6.

#### 3.2.2 Materials

The participants all had to use their computer to do the experiment on. Each participant was sent a unique online Google spreadsheet which they had to fill in. As the spreadsheets were online documents, they were automatically saved.

#### 3.2.3 Procedure

As in the original experiment, I divided the experiment into experiment A and B.

For experiment A, the original 30 scene-description pairs were used for the reconstruction. Every pair was seen by four subjects, by randomly allocating each pair four times to the 20 participants in such a way that a participant did not see the same pair twice. This means that from each description producer (human, QDG-GREEDY, or QDG-IA) ten unique pairs were reconstructed, each four times. The scene-descriptions pairs were given in ascending order of domain size (so  $n=4$ ,  $n=9$ ,  $n=20$  respectively), but in a random order within the domain size. Thus, every participant saw the pairs in a different order.

For experiment B, the original 66 scene-description pairs were used for the reconstruction. Every pair was seen twice, by randomly allocating each pair twice to the 20 participants in such

a way that a participant did not see the same pair twice. This means that from each description producer (QDG-GREEDY or QDG-1A) 33 unique pairs were reconstructed, each two times. The scene-descriptions pairs were given in ascending order of domain size (so  $n=6$ ,  $n=10$ ,  $n=16$  respectively), but in a random order within the domain size.

Each participant thus will see 6 scenes in experiment A and 2 to 3 scenes in experiment B, which makes it a total of 8 to 9 scenes per participant.

Before the experiment, the subjects were given the following instructions:

*In the experiment, you're going to read a number of descriptions. Each description describes a visual scene containing some simple geometrical objects. For each description, we'd like you to tell us what scene the description evokes: in other words, please tell us about a scene that could be described by the description.*

*Please note:*

- *Each object is a circle or a square, and is either red or blue;*
- *For each description, we will tell you how many objects the scene described by it contains (for instance, the "size" of the scene may be 4);*
- *We are not interested in the location of each object. Instead of asking you to draw the scene, we will therefore only ask you how many objects of each type it contains (for instance, 3 red circles and 1 blue square); (...)*
- *We believe that for some descriptions there is more than one "correct" answer. In those cases, please choose an answer that you consider to be consistent with the description (please choose only one answer).*

*Here are two examples, using B for blue, R for Red, S for square, and C for circle: (...)*

A full version of the instructions can be found in section 7.2. An example of an answering sheet can be found in section 7.3. The participants had to give their answers online on a spreadsheet, and each domain size had its descriptions on a separate spreadsheet. Each participant had its unique spreadsheet, such that solely the descriptions which were assigned to that participant were included in that person's spreadsheet.

### 3.2.4 Pilots

I conducted two pilot experiments and made some minor details to the layout of the instructions: using the same answering sheet layout in the examples as in the instructions. The examples were first given by using the abbreviations BS, RS, BC, RC; I changed it to images of the objects in their corresponding colour. I also made multiple sheets for every domain size instead of having them all in one; one subject indicated that he had to constantly scroll up and down to see which column represented what object.

## 3.3 Hypotheses

Based on the results of the original experiment, I formulated the following hypotheses:

1. **The larger the domain size  $n$ , the more the reconstructions will diverge.** Since for a larger  $n$ , there are often multiple answers correct for that description, but they are not the "right" answer;
2. **The larger the domain size  $n$ , the more the reconstructions will diverge from the real scene.** The reasoning for hypothesis 1 can be applied to this as well;



3. **The higher the scores given for *informativity* and *correctness* by the experts, the less the reconstructions will diverge from the real scene.** This is because I expect that the scores given by the experts mostly comply with the results of this experiment;
4. **Reconstructions based on both algorithms will be correct more often than the reconstructions based on the human descriptions,** because "both of them were explicitly designed to optimise informativity and correctness" (Chen et al., 2019a)

### 3.4 Analysis of the data

To determine how well a subject reconstructed a scene-description pair, I determined how many "swaps" a person made in their answer. Since all scenes have a certain size, adding an object in the reconstruction (to either BS, RS, BC, or RC)<sup>1</sup> always causes another object to have one less to keep the total number of objects equal to the scene size  $n$ . This results in a so-called swap. To calculate the number of swaps, I used the following formula: for each reconstruction calculate the absolute difference between the "correct" answer for each type of object and the answer given by the subject, then add these up and divide by 2. This results in the total number of switches made by the participant for that scene-description pair. For example, if the correct answer would have been

BS: 2, RS: 1, BC: 0, RC: 1

and the answer that was given by the subject was

BS: 2, RS: 1, BC: 1, RC: 0

the calculation would be as follows:

$$(|2 - 2| \text{ (BS)} + |1 - 1| \text{ (RS)} + |0 - 1| \text{ (BC)} + |1 - 0| \text{ (RC)})/2 = 1 \text{ swap}$$

In order to compare these scores for the different domain sizes, I then determined the average number of swaps for every domain size and divided it by the total possible swaps for that domain size (the total is equal to the domain size itself).

The following calculation was done in order to determine by how much the answers from all the participants diverge from one another: for each scene-description pair calculate the average answer for every object (BS, RS, BC, RC) separately, then take the standard deviation for all these means separately, and finally take the average for the standard deviation per scene. An ANOVA test with a significance level  $\alpha = .05$  was performed on these standard deviations, to determine whether their means are different.

---

<sup>1</sup>BS = blue square, RS = red square, BC = blue circle, RC = red circle

## 4 Results

### 4.1 Swaps

Table 2 shows the average percentage of swaps of the total number of possible swaps for QDG-IA, QDG-GREEDY, and human-produced descriptions for each domain size in experiment A. It also shows the total average percentages per model, and for each domain size separately.

Table 3 is a similar table, but for experiment B: it shows the average percentage of swaps of the total number of possible swaps for the QDG-IA and QDG-GREEDY descriptions for each domain size.

The percentages are somewhat higher for the QDG-IA descriptions than for the QDG-GREEDY, for domain sizes 10 and 16. When looking at the total average number of swaps for each domain size, the value again rises as the domain size increases from  $n=6$  to  $n=10$ . However, there is a slight decrease when going from  $n=10$  to  $n=16$ . This seems to be caused by the decrease in QDG-GREEDY; for QDG-IA the percentages do increase together with the domain size. However, a t-test with  $\alpha = .5$  resulted in  $p=$

	Model	n=4	n=9	n=20
Experiment A	Human	8.33%	6.25%	15%
	QDG-IA	0%	11.81%	18.33%
	QDG-GREEDY	2.08%	8.33%	15%
	Total	3.47%	8.8%	16.11%

Table 2: Average swap percentages of the total possible swaps for each algorithm and for human-produced descriptions, and the average swap percentages for  $n=4$ ,  $n=9$ , and  $n=20$ . The total average swap percentages for every model and every domain size is also included.

	Model	n=6	n=10	n=16
Experiment B	QDG-IA	1.39%	10%	10.42%
	QDG-GREEDY	1.39%	8.33%	7.81%
	Total	1.39%	9.17%	9.11%

Table 3: Average swap percentages of the total possible swaps for each algorithm, and the average swap percentages for  $n=6$ ,  $n=10$ , and  $n=16$ . The total average swap percentages for every model and every domain size is also included.

### 4.2 Variance in answers

Table 4 presents the average standard deviation in the answers for every scene per domain size in experiment A. Every scene-description pair was seen by four subjects. The p-value ( $2.2167e-06$ ) from the ANOVA test (significance level  $\alpha = .05$ ) suggested that one or more of the values in table 4 are significantly different. The posthoc Tukey test was applied to these values, and this resulted in a significant difference between pairs  $n=4$  and  $n=20$  ( $p=0.001$ ), and  $n=9$  and  $n=20$  ( $p=0.001$ ). The difference between  $n=4$  and  $n=9$  turned out to be insignificant ( $p=0.458$ ).

	Avg. $\sigma$ n=6	Avg. $\sigma$ n=10	Avg. $\sigma$ n=16
Experiment A	0 <sup>1</sup>	1 <sup>10</sup>	6.532 <sup>22</sup>
	0 <sup>2</sup>	0 <sup>11</sup>	3 <sup>23</sup>
	0 <sup>3</sup>	1.915 <sup>12</sup>	4.435 <sup>24</sup>
	0 <sup>4</sup>	0 <sup>13</sup>	2 <sup>25</sup>
	0 <sup>5</sup>	0 <sup>14</sup>	2.915 <sup>26</sup>
	0 <sup>6</sup>	2 <sup>15</sup>	2 <sup>27</sup>
	0 <sup>7</sup>	2.582 <sup>16</sup>	9.229 <sup>28</sup>
	0 <sup>8</sup>	2.309 <sup>17</sup>	4.817 <sup>29</sup>
	1 <sup>9</sup>	1 <sup>18</sup>	8.639 <sup>30</sup>
		0 <sup>19</sup>	
		1 <sup>20</sup>	
	0 <sup>21</sup>		
Avg. all scenes	0.111	0.984	4.841

Table 4: Average standard deviation in the answers for each domain size, by taking the average number of standard deviation for BS, RS, BC, RC in every scene. The scene ID’s are noted as superscript.

Table 5 presents the average standard deviation in the answers for every scene per domain size. Every scene-description pair was seen by two subjects. The p-value (0.0008) from the ANOVA test (significance level  $\alpha = .05$ ) suggested that one or more of the values in table 5 are significantly different. The posthoc Tukey test was applied to these values, and this resulted in a significant difference between pairs n=6 and n=16 (p=0.002), and n=10 and n=16 (p=0.004). The difference between n=6 and n=10 turned out to be insignificant (p=0.900).

	Avg. $\sigma$ n=6	Avg. $\sigma$ n=10	Avg. $\sigma$ n=16	
Experiment B	1.414 <sup>31</sup>	0 <sup>43</sup>	0 <sup>55</sup>	
	0 <sup>32</sup>	0 <sup>44</sup>	0 <sup>56</sup>	
	0 <sup>33</sup>	0 <sup>45</sup>	0 <sup>57</sup>	
	0 <sup>34</sup>	0 <sup>46</sup>	2.828 <sup>58</sup>	
	0 <sup>35</sup>	2.828 <sup>47</sup>	2.828 <sup>59</sup>	
	1.414 <sup>36</sup>	0 <sup>48</sup>	2.828 <sup>60</sup>	
	0 <sup>37</sup>	1.414 <sup>49</sup>	1.414 <sup>61</sup>	
	0 <sup>38</sup>	0 <sup>50</sup>	2.828 <sup>62</sup>	
	0 <sup>39</sup>	0 <sup>51</sup>	1.414 <sup>63</sup>	
	0 <sup>40</sup>	0 <sup>52</sup>	4.243 <sup>64</sup>	
	0 <sup>41</sup>	0 <sup>53</sup>	1.414 <sup>65</sup>	
	0 <sup>42</sup>	0 <sup>54</sup>	1.414 <sup>66</sup>	
	Avg. all scenes	0.236	0.354	1.768

Table 5: Average standard deviation in the answers for each domain size, by taking the average number of standard deviation for BS, RS, BC, RC in every scene. The scene ID’s are noted as superscript.

### 4.3 Correlation experts’ scores and reconstructions

Figure 2 shows all the informativity (a) and correctness (b) scores combined with the swap percentages for experiment A, figure 3 shows this for experiment B. In experiment A, informativity

has a standard error of 0.0739 and  $R^2=0.3032$ . Correctness has a standard error of 0.0772 and  $R^2=0.1391$ . In experiment B, informativity has a standard error of 0.0809 and  $R^2=0.4384$ . Correctness has a standard error of 0.0923 and  $R^2=0.2690$ .

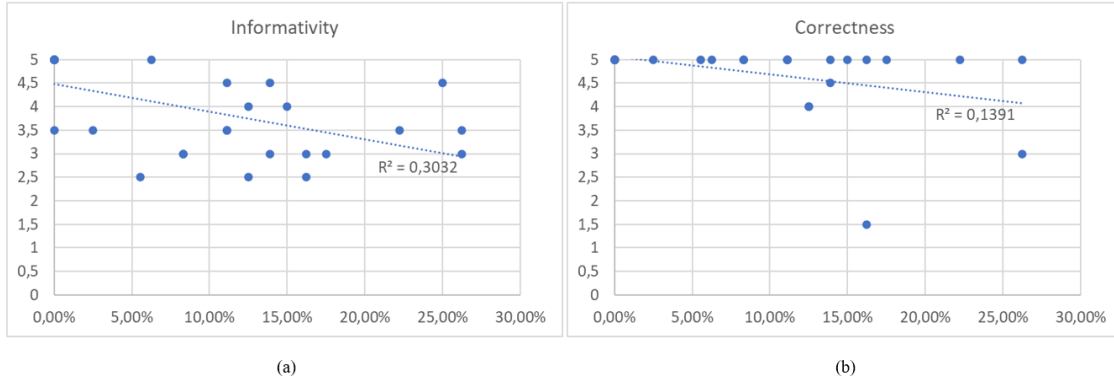


Figure 2: The informativity (a) and correctness (b) scores from the original experiments by Chen et al. (2019a) on the y axis, and my swap percentages on the x axis, along with a linear regression line.

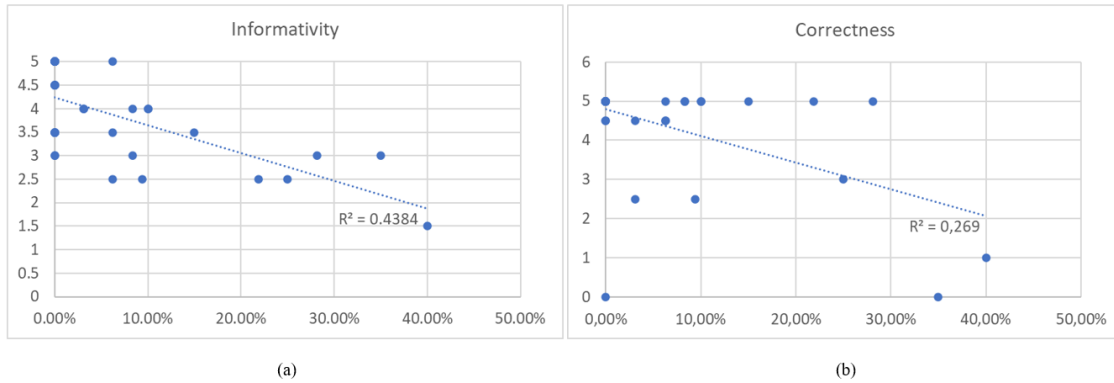


Figure 3: Combined results of the informativity and correctness scores from the original experiments by Chen et al. (2019a) on the y axis, and my swap percentages on the x axis, along with a linear regression line.

## 5 Discussion and conclusion

### 5.1 Evaluation of hypotheses

First I am going to look at the hypotheses I formed in section 3 to discuss the data even further, draw conclusions from it, and give an answer to my research question: "To what degree do the judgements by the experts on the quantified description making algorithms comply with the results from the task-based evaluation on these algorithms?". The first two hypotheses mainly serve as "sanity checks" for the experiment. The last two hypotheses are the main focus when answering the research question.

The first hypothesis states:

1. The larger the domain size  $n$ , the more the reconstructions will diverge.

Table 4 shows the standard deviation of the answers for every domain size in experiment A. Since the difference between  $n=4$  and  $n=16$  is not significant, I cannot say anything about these numbers. The difference between the significant domain sizes show that the hypothesis holds for experiment A: as the domain size increases, the answers deviate more.

Table 5 shows the standard deviation of the answers for every domain size in experiment B. Since the difference between  $n=6$  and  $n=10$  is not significant, I cannot say anything about these numbers. The difference between the significant domain sizes show that the hypothesis holds for experiment B: as the domain size increases, the answers deviate more.

In both experiments A and B, the difference between the two smaller domain sizes are those that are not significant. This could perhaps be because the difference in the maximum number of swaps is lower between these domain sizes ( $9-4=5$  for A, and  $10-6=4$  for B) than for the others ( $20-9=11$ ,  $20-4=16$  for A, and  $16-10=6$ ,  $16-6=10$  for B).

I am not able to accept this hypothesis completely, because I was not able to use all the data I acquired in its evaluation. However, with the useful data, the hypothesis can be accepted.

The second hypothesis to check the experiment states:

2. The larger the domain size  $n$ , the more the reconstructions will diverge from the real scene.

This hypothesis is supported by both experiments; the number of swaps increases when the domain size increases as well when looking at experiment A and B separately. When combining the results of both experiments, as shown in table 6, the hypothesis does not always hold. However, this is probably caused by the fact that in experiment A, a scene-description pair was viewed by twice as many subjects as in experiment B (four versus two). This means that the experiments cannot be compared this way unless these values were equal.

	Model	n=4	n=6	n=9	n=10	n=16	n=20
Experiment A & B	QDG-IA	0%	1.39%	11.81%	10%	10.42%	18.33%
	GREEDY	2.08%	1.39%	8.33%	8.33%	7.81%	15%
	Total	3.47%	1.39%	8.8%	9.17%	9.11%	16.11%

Table 6: Average percentage of swaps for each algorithm and for human-produced descriptions, and the average number of swaps for  $n=4$ ,  $n=6$ ,  $n=9$ ,  $n=10$ ,  $n=16$ , and  $n=20$ . The total average percentage of swaps for every domain size is also included.

This means that the hypotheses that served mainly as sanity checks still hold, and the results of the task-based evaluation are reliable.

The conclusions based on the following two hypotheses will tell us more about the experts'

rating compared to my experiments:

3. The higher the scores given for *informativity* and *correctness* by the experts, the less the reconstructions will diverge from the real scene.

This hypothesis can be divided up into two sub-hypotheses, namely: (a) The higher the scores given for informativity, the less the reconstructions will diverge from the real scene, and (b) The higher the scores given for correctness, the less the reconstructions will diverge from the real scene.

To be able to accept or reject the hypotheses, I performed a linear regression analysis of the data for both. The results are in figures 2 and 3. All  $R^2$  values indicate that there is some relationship between the experts' scores and the percentages of swaps, more precisely; when the correctness/informativity scores are higher, the percentages of swaps are lower. To evaluate the relationship, I use the interpretation grid by Cohen (1988):  $R^2 = 0 - 0.02$ : Very weak,  $R^2 = 0.02 - 0.16$ : Weak,  $R^2 = 0.16 - 0.26$ : Moderate, and  $R^2 > 0.26$ : Substantial. This means that in experiment A the relationship between swaps and informativity is substantial, as well as the relationships between swaps and informativity, and swaps and correctness in experiment B. The relationship between swaps and correctness in experiment A is considered weak. This means that the hypothesis can be accepted, though not with complete confidence due to the one weak relationship.

Finally, I will have a look at the fourth hypothesis:

4. Reconstructions based on both algorithms will be correct more often than the reconstructions based on the human descriptions.

If I apply this hypothesis to acquired data, it means that: the percentages of swaps have to be lower for the reconstructions based on both algorithms than for reconstructions based on the human descriptions. This can only be applied to experiment A since there was no data available for human descriptions in experiment B. An ANOVA test (with  $\alpha = .05$ ) resulted in no significant difference between the results from both algorithms and the human descriptions: for QDG-IA  $p=0.760$ , and for QDG-GREEDY  $p=0.665$ . based on these results, the hypothesis should be rejected. In the original experiment by Chen et al. (2019a) there was a similar hypothesis, which was also rejected: "(...) both algorithms perform better at informativity and correctness than humans, (...)". What is interesting about this is that it means that the scores by the experts do somewhat comply with the reconstruction scores (swap percentages).

## 5.2 Research question

The third hypothesis is the most important hypothesis to answer my research question since it addresses the relationship between the experts' scores and the results from my experiments. Despite the one case in which the relationship is weak, for the most part, it is true that: the higher the experts rated the description models on informativity and correctness, the less the reconstructions will diverge from the real scene. Let us go back to the research question: "To what degree do the judgements by the experts on the quantified description making algorithms comply with the results from the task-based evaluation on these algorithms?". I can thus say that the judgements mostly comply with the experiments' results.

## 5.3 Further research and limitations

Given the rather positive outcome on my research question and the methodology of my experiments (hypothesis 1 and 2), it would be interesting to look further into this type of task-based evaluation of quantified description generating algorithms. This form of evaluation could also be performed

once more when, for example, the algorithms have been perfected. Additionally, it might also be wise to use more participants in that case or do this exact experiment again for many more subjects. This could give more reliable results and hopefully, a significant difference between the values for smaller domain sizes. If this were to be done, it would perhaps help to do more pilot experiments to filter out potential misunderstandings or be present when conducting the experiment to check if the subject is doing what is expected. Since I had to send everyone the experiment online (due to the coronavirus), I was not present when the participants did the experiment. One participant did not understand the instructions completely; he or she did not know that there was a scene size. Some subjects miscalculated when giving the answers, and as a result, their total number of objects did not match the scene size. This meant that a few participants had to redo some or all tasks.

I also noticed I got asked a few similar questions by different participants regarding certain descriptions. They, for example, asked what was meant by a phrase as "All objects are shown"; does it mean circles and squares, or circles and squares in all possible colours? (I did not give an answer of course). It might be useful for the improvement of the algorithms if participants were able to comment on a description, or indicate that they did not quite understand what a description meant.

## 6 Bibliography

- Chen, G., van Deemter, K., and Lin, C. (2019a). Generating quantified descriptions of abstract visual scenes. In van Deemter, K., Lin, C., and Takamura, H., editors, *Proceedings of the 12th International Conference on Natural Language Generation*, pages 529–539, Tokyo, Japan. Association for Computational Linguistics.
- Chen, G., van Deemter, K., Pagliaro, S., Smalbil, L., and Lin, C. (2019b). QTUNA: A corpus for understanding how speakers use quantification. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 124–129, Tokyo, Japan. Association for Computational Linguistics.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.
- Dale, R. and Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.
- Ghosh, S. and Gunning, D. (2019). *Natural language processing fundamentals*. OCLC: 1101443873.
- Glanzberg, M. (2009). *Quantifiers*. Oxford University Press.
- Perera, R. and Nand, P. (2017). Recent advances in natural language generation: A survey and classification of the empirical literature. *Computing and Informatics*, 36(1):1–32.



## 7 Appendices





### 7.1 Appendix A: Participants

---

Participant	Age	Educational background
1	23	BSc Artificial Intelligence
2	23	BSc Artificial Intelligence
3	22	BSc Artificial Intelligence
4	21	BSc Artificial Intelligence
5	22	BSc Psychology
6	22	BSc Artificial Intelligence
7	22	BSc Philosophy
8	22	BSc Artificial Intelligence & BSc Social Sciences
9	25	MSc Artificial Intelligence
10	22	BSc Artificial Intelligence
11	23	BSc Artificial Intelligence
12	21	BSc Artificial Intelligence
13	21	BSc History
14	25	BSc Computing Science
15	20	BSc Artificial Intelligence
16	21	BSc Medicine
17	25	BSc Earth Science
18	23	BSc Artificial Intelligence
19	23	BSc Human Geography and Spatial Planning
20	25	BSc Artificial Intelligence





---

## 7.2 Appendix B: Instructions

Dear participant,									
This experiment aims to make Artificial Intelligence programs better at understanding ordinary language.									
In the experiment, you're going to read a number of descriptions. Each description describes a visual scene containing some simple geometrical objects. For each description, we'd like you to tell us what scene the description evokes: in other words, please tell us about a scene that could be described by the description.									
Please note:									
- Each object is a circle or a square, and it is either red or blue;									
- For each description, we will tell you how many objects the scene described by it contains (for instance, the "size" of the scene may be 4);									
- We are not interested in the location of each object. Instead of asking you to draw the scene, we will therefore only ask you how many objects of each type it contains (for example, 3 red circles and 1 blue square);									
- You are allowed to draw or take notes on a piece of paper — in fact, this may be a good idea! — but we want you to write your final answer on the answering sheet, using numbers;									
Simply fill in the numbers just as in the examples below (columns D, E, F and G);									
- We believe that for some descriptions there is more than one "correct" answer. In those cases, please choose an answer that you consider to be consistent with the description (please choose only one answer).									
Here are two examples:									
Example 1: [scene size 4.]									
Description: "Half of the objects are blue squares, the rest are red squares."									
In this case we would expect you to answer:	2	2	0	0					
Example 2: [scene size 20.]									
Description: "There is a mixture of squares and circles. Most of them are blue."									
Possible answer:	9	2	8	1					
Possible answer:	8	1	9	2					
Etc...									
On the bottom of this document you will see sheets labeled S=4, S=9 etc. This is where you can find the descriptions (a sheet can be empty as well). Make sure you finish every sheet.									
Please read these instructions again to make sure you know what to do.									
We expect this experiment to take you about 20 minutes. We hope you will complete the experiment for us, but if you don't want to, then you should feel free to quit at any time.									
Many thanks for your help.									

### 7.3 Appendix C: Answering sheet

This is an example of an answering sheet for  $n = 4$ .

ID	Scene size	Description				
(ignore this)						
7	4	All objects are circles. A quarter of the objects is blue.				
8	4	All of the objects are circles. A majority of the objects are red circles.				
4	4	All the objects are squares and half of them is blue.				