QTUNA: A Corpus for Understanding How Speakers Use Quantification

Guanyi Chen[♠], Kees van Deemter[♠♡], Silvia Pagliaro[♠], Louk Smalbil[♠], Chenghua Lin[♣]
[♠]Department of Information and Computing Sciences, Utrecht University
[♡]Department of Computing Science, University of Aberdeen
[♣]Department of Computer Science, University of Sheffield
{g.chen, c.j.vandeemter}@uu.nl, s.pagliaro@students.uu.nl
l.smalbil@students.uu.nl, c.lin@sheffield.ac.uk

Abstract

A prominent strand of work in formal semantics investigates the ways in which human languages quantify over the elements of a set, as when we say "*All A are B*", "*All except two A are B*", "*Only a few of the A are B*" and so on. Our aim is to build Natural Language Generation algorithms that mimic humans' use of quantified expressions. To inform these algorithms, we conducted on a series of elicitation experiments in which human speakers were asked to perform a linguistic task that invites the use of quantified expressions. We discuss how these experiments were conducted and what corpora they gave rise to. We conduct an informal analysis of the corpora, and offer an initial assessment of the challenges that these corpora pose for Natural Language Generation. The dataset is available at: https://github.com/a-quei/qtuna.

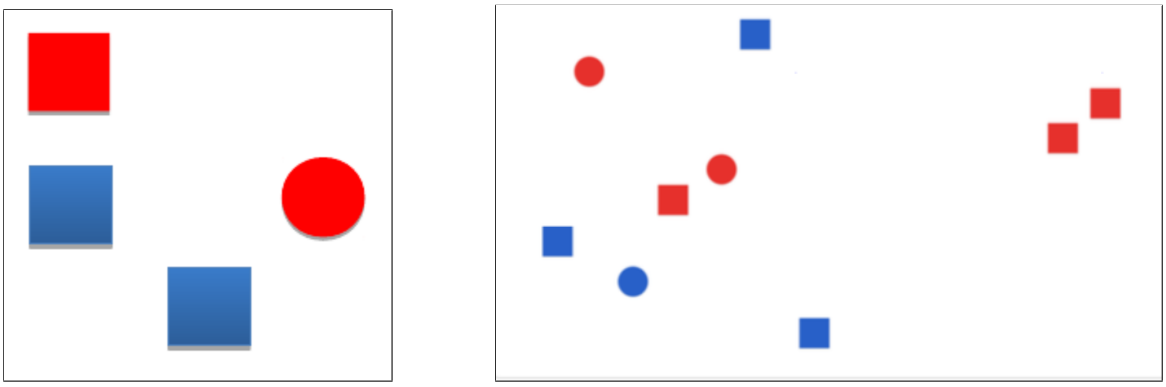
1 Introduction

A long tradition of research in the formal semantics of natural language asks how speakers quantify, as when we say "*Some A are B*", "*All except two A are B*", "*Only a few of the A are B*" and so on. This area of work is known as the theory of "Generalised Quantifiers" (GQ) (Peters and Westerstahl, 2006, GQ), because it generalises the idea of quantification beyond the standard logical quantifiers of $\forall$ and $\exists$, even including quantifiers like "*most*" or "*many*", which are not expressible in First-Order Logic (Mostowski, 1957; Barwise and Cooper, 1981; Van Benthem et al., 1986; Peters and Westerstahl, 2006). Since definite NPs can also be understood in these terms, GQ theory comprises, at least in principle, all Noun Phrases (NPs): the study of quantifiers in natural language is essentially the study of Noun Phrases.

There exists some work that can help to give this theoretical work an empirical basis. For example, there is psycholinguistic work on people's use of vague quantifiers (Moxey and Sanford, 1993), and work that investigates the links between quantifiers' logical types and human processing of quantified expressions (Szymanik and Zajenkowski, 2010; Szymanik et al., 2016, QEs). Yet there is a dearth of knowledge about human usage of QEs. For instance, what QEs, and what combinations of QEs, are uttered by a speaker in a given situation, to accomplish a given task? And, if a given NP is uttered, what information does it convey? Some questions are starting to be addressed, for example, Sorodoc et al. (2016) looked at speakers' choice between "all", "some", and "no" (see also Grefenstette (2013) and Herbelot and Vecchi (2015)). Yildirim et al. (2013) studied speakers' use (and hearers' interpretation) of the quantifiers "some" and "many", as in "Many of the candies are green". Barr et al. (2013) investigated referring expressions in which a quantifier is embedded (e.g., "the square with 11 black dots", "the square with lots of dark dashes"). Yet there has been few attempts to chart how the wider class of generalised quantifiers are used by human speakers. The present paper lies the basis for such a study, with the ultimate aim of modelling the human production of quantifiers computationally.

In the computational modelling of language production, one class of NPs has been studied widely, namely *referring* NPs (Krahmer and van Deemter, 2012), and van Deemter (2016). One line of work focuses on corpora of referring expressions (REs) that were elicited under experimentally controlled conditions (e.g., the TUNA corpus (Gatt et al., 2007; van Deemter et al., 2012a)). Such corpora were used as a gold standard for a sequence of evaluation campaigns in which generation algorithms that produce referring expressions were compared with the gold standard (Gatt and Belz, 2010). This systematic

Proceedings of The 12th International Conference on Natural Language Generation, pages 124–129,
Tokyo, Japan, 28 Oct - 1 Nov, 2019.



(a) (b)

Figure 1: Examples from (a) the $n = 4$ experiment; (b) the $n = 9$ experiment.

comparison allowed researchers to know which algorithms worked best, and to develop new algorithms that match human language production. Inspired by this line of work on referring expressions, but aiming this time to gain insights into quantified NPs, we conducted a new series of elicitation experiments, called the QTUNA experiments (where Q stands for quantification). We report on these elicitation experiments, on the resulting corpus, and on an initial analysis of the corpus.

We set up the experiments in such a way that they would tell us how quantified NPs are employed to describe an abstract visual scene. We were curious what NPs would be used, what they meant, and how they were used (e.g., how correctly and how completely did speakers manage to describe the different scenes?). We were keen to look at tasks of different difficulty levels, curious how these levels affected the use of quantifiers.

2 The QTUNA Experiment

We wanted to find out how a wide range of quantified NPs are used as part of a wider communicative task. So, instead of showing our subjects a scene and asking them how they would describe the number of so-and-so's (e.g., circles) that are red (e.g., "Many circles are red"), we asked them to describe the scene as a whole, hoping that they would use quantifiers to do this. Moreover, we made the scenes complex enough that one simple Quantified Expression would never suffice. Let's explain in more detail how we proceeded.

Each participant presented with a series of abstract visual scenes. We asked them to try to produce a description that would allow a reader to *reconstruct* the situation, except for the location of the objects. Each scene contains a certain number of objects, which is either a circle or a square and either red or blue. In order to gain insight into the question of how domain size impacts the production of QEs, we conducted three experiments, with domain size (n) of 4, 9, and 20 respectively, each containing 10 different scenes. Figure 1(a) and Figure 1(b) show two examples from the $n = 4$ and $n = 9$ experiment respectively; Figure 2 depicts how the instruction looks like.

We'd like you to describe each situation in one or more grammatically correct English sentences. (...)

1. *Based on your description, a reader will try to "reconstruct" the situation. We use the word "reconstruct" loosely here, because the only thing that matters is the different types of objects that the sheet contains. Therefore, please do not say *where* in the grid a particular object is located (e.g., "top left", "in the middle", "on the diagonal").*

2. *Each object is a circle or a square, and either red or blue. Your reader knows this.*

3. *Please do not "enumerate the different types of objects. For example, do not say "There is a red circle, two blue circles, and ...".*

4. *Every situation contain four objects. Your reader knows this in advance, and he/she will take this information into account when interpreting your description.*

Figure 2: The sketch of how a instruction looks like, taking $n = 4$ as an example. A full version of the instruction can be found in the supplementary material.

We conducted a number of pilot experiments for each of the three experiments. These taught us that if no further instructions were offered,

$n = 4$	*There are 4 squares. Every object is blue.*
$n = 4$	*More than half of the objects are blue squares. Less than half are blue circles.*
$n = 4$	*There is one red square and the rest are blue circles.*
$n = 4$	*All possible objects are shown.*
$n = 9$	*Most of the items are red circles, but there are a couple of blue squares.*
$n = 9$	*Most of the objects are blue squares. A few objects are blue circles.*
$n = 20$	*All the objects in the picture are circles and majority of them is blue.*
$n = 20$	*Both circles and squares appear in either red or blue.*

Table 1: List of example descriptions from QTUNA corpus; n indicates domain size.

then only a small range of existentially quantified sentence patterns would be used. For example, for Figure 1(a), a description like "*There are two blue squares, one red square and one red circle.*" would tend to be given. To nudge participants into using a wider range of quantified statements, we asked participants not to use *enumerations*, followed by an example. This request may have diminished the ecological validity (Schmuckler, 2001) of our experiment, but we believe that this is more than outweighed by the increased richness of the resulting descriptions.

Participants were students at the Computing department of Utrecht University. Data from 66, 63, and 58 participants were collected for the three experiments. We manually filtered out all descriptions from subjects who showed a misunderstanding of the task by committing at least 3 (what we considered to be) errors, namely by writing gibberish, by using enumerations, or by expressing locations (e.g., ".. in the top left"). The resulting corpus contains 656, 380, and 378 valid descriptions for the three domain sizes. Examples are shown in Table 1. The larger the domain size, the smaller was the proportion of valid descriptions in it, presumably because the difficulty of producing descriptions increases with domain size.

We annotated each description with a formula that encodes its semantic content. Following Barwise and Cooper (1981), we used a form in which each k-ary quantifier is a relation between 2 or more set terms as arguments. For example, "*All the objects are blue. Half of them are squares.*" is labelled as $\text{All}(O, B) \wedge \text{Half}(O, S)$, where O, B, and S stand for the set of all objects in the situation, blue objects, and squares, respectively.

3 Analysis

The corpus was analysed on the basis of hypotheses formulated before we looked into the corpus. Annotations were done by the first two authors, who discussed their initial judgements and made final decisions together. All hypotheses focus on generation, that is, on the choices that speakers made between different possible utterances.

Vague quantifiers. The larger a domain, the harder it is to see at a glance how many objects there are in each of its set-theoretic regions (A, B, $A \cup B$, $A \cap B$, $A - B$, $B - A$, and the domain of objects O as a whole). We therefore hypothesised ($\mathcal{H}_1$) that, as our 3 domains grew larger, more vague quantifiers would occur. To test this idea, we counted the number of QEs that use vague quantifiers (e.g., *many*, and *few*, which permit so-called borderline cases, where it is unclear whether the QE is true or false, see e.g., Keefe and Smith (1996)) in each sub-corpus.[1] The number of vague QEs was compared with the total number of QEs (Table 2). Chi-Square suggests an affirmative answer ($\chi^2(2) = 471.55, p < .001$).[2]

How often do speakers describe a situation completely and correctly?

We considered a description to be complete if the described situation was the only one (modulo location) that fits the description. Since producing a complete description requires more work in a larger domain, we hypothesised that larger domains would give rise to a smaller proportion of complete descriptions than smaller ones.

This hypothesis is challenging to test because speakers frequently relied on inference when describing a scene. Consider "*half of the objects are blue*". Given there are only two colours (blue and red), we infer that the other half are red. Or consider, "*Everything is blue. Most things are*

[1] A list of all the vague quantifiers in our corpora can be found in the supplementary material.

[2] As we had 6 hypotheses, all the p-values reported in this paper are those after Bonferroni correction, i.e., multiplicated by 6.

Hypothesis	$n = 4$	$n = 9$	$n = 20$
$\mathcal{H}_1$: #(Vague Quantifier)/#(QE)	57/1401	201/638	234/543
$\mathcal{H}_2$: #(Incompleteness)/#(Description)	46/656	137/380	261/378
$\mathcal{H}_3$: #(Wrong Description)/#(Description)	7/656	11/380	29/378
$\mathcal{H}_4$: #(Word) per description	13.02	12.75	9.53
$\mathcal{H}_4$: #(QE) per description	2.13	1.67	1.43

Table 2: Statistics with respected to some of the hypotheses in §3, where #(·) means "the number of".

square". If "*most*" means not just "more than half" but also "not all", then the above description *completely* describes a situation with 3 blue squares and 1 blue circle, despite not saying this explicitly. Instead of relying on our formalisation of the meaning of quantifiers,[3] we tackled the issue by asking annotators to say directly, for each description in each sub-corpus, whether they considered the description to be logically complete.

We likewise hypothesised ($\mathcal{H}_2$) that smaller domains would give rise to a larger proportion of logically complete descriptions (because this is easier in a smaller domain). The results in Table 2 confirm ($\chi^2(2) = 443.60, p < .001$) this.

For the same reason, we expected ($\mathcal{H}_3$) that, in larger domains, there will be more descriptions that convey *incorrect* information, because counting mistakes becomes more likely. For example, we would mark a description "*all objects are blue*" as incorrect if it describes a situation where all objects are actually red. Chi-square shows an overall association between domain size and error frequency ($\chi^2(2) = 32.85, p < .001$). The association also held between each subsequent level of size, but although the proportion of errors went up from $n = 9$ to $n = 20$ ($\chi^2(1) = 8.65, p < .01$), from $n = 4$ to $n = 9$ it fell, perhaps because *vague* QEs are used (as is frequently the case in $n = 9$ and $n = 20$, but not in $n = 4$). This reduces the proportion of QEs that are downright incorrect (e.g., annotators in most situations will have been understandably reluctant to describe a QEs of the form "many/few A are B" as incorrect).

Are larger scenes described more elaborately? We expected ($\mathcal{H}_4$) participants to produce longer descriptions in larger scenes, because there is more to describe. To test this, we calculated the length, as defined by both the number of words and the number of QEs, of each description. The results in Table 2 show the opposite of what we expected: the length of descriptions *decreased* with domain size. A plausible explanation lies in the fact that (hypothesis $\mathcal{H}_2$), speakers produced fewer *complete* descriptions in larger domains.

[3] See e.g. Coventry et al. (2010) for problems assessing the meaning of "most".

	First	Later
Shape	489	121
Colour	112	514

Table 3: The number of QEs that put shape/colour in the first/later place.

Ordering of QEs. We noticed during our pilots that speakers tended to employ two discourse structures. The first starts by describing the whole scene, e.g., "*all objects are blue*", followed by a more detailed statement, e.g., "*half of them are squares*". The second discourse structure cuts the set of objects into two parts, each of which is described separately. We hypothesised ($\mathcal{H}_5$) that when a scene is described in two parts, where one part is larger than the other, then the larger part if described before the smaller part, because this strategy lets more important information be followed by less important information. For instance, "*3/4 of A are B, 1/4 are C.*" occurs more often than "*1/4 of A are C, 3/4 are B.*". We counted the number of descriptions that describe the larger part first, and those that describe the smaller part first, obtaining the numbers of 367 and 136 descriptions respectively. This confirmed ($\chi^2(1) = 212.17, p < .001$) the hypothesis.

Differences between colour and shape. Given the well-documented primacy of colour over shape in referring expressions (Pechmann, 1989; van Deemter et al., 2012b), it seemed plausible to us that colour and shape play different roles in quantification too. Based on our pilot experiments, our last hypothesis ($\mathcal{H}_6$) was: in k-ary QEs (i.e., QEs with quantifiers that describe relations between k sets), shape occurs more often in the first argument

place (i.e., the A position in $Q(A, B)$) and colour in the second argument place (the B position). For example, we expect to see sentences like "*all circles are blue*" more often than ones like "*all blue objects are circular*". The results in Table 3 confirm this hypothesis ($\chi^2(1) = 479.59, p < .001$).

4 Discussion

The corpus also gave rise to a number of interesting *post hoc* observations. For example, we found a substantial number of 3-ary quantifiers, such as "*half of A are B, and the other half are C*", which should not be confused with "*half of A are B and half of A are C*"; the latter allows B and C have a non-empty intersection, while the former means 1/2 of A are B and $(A - B) \subseteq C$. A similar example is "*Most A are B, the others are C*"

Another unanticipated feature is the existence of higher-order quantifiers. For instance, in the $n = 4$ experiment, when all the objects were different, many participants used QEs such as "*All possible types of objects are shown*", a strikingly brief and complete description which quantifies over the Cartesian product of the colours and the shapes.

In future, two issues will be addressed: 1) What descriptions will be produced if the domain size is further increased? One might expect that, similar to the findings of this paper, the participants would produce even more vague quantifiers, more incompleteness, etc.; 2) What types of QEs are produced in other languages? We are particularly curious about Chinese, since previous corpus study for machine translation suggests there is much more variations in QEs in Chinese than in English (Wang and Piao, 2007).

Computational modelling of the human production of QEs is one major goal of building QTUNA corpus. The idea is to work analogously to the generation of *referring* expressions, where corpora of experimentally elicited descriptions (such as the TUNA corpus (Gatt et al., 2007)) have guided the construction and evaluation of Referring Expressions Generation algorithms. In the same way, the QTUNA corpus can guide the construction of algorithms that mimic the human production of *quantified* descriptions. (For example, the corpus can help us understand which quantifiers and QE patterns are most frequently used, and how elaborate a description needs to be – for example, when should the generator stop adding further QEs, because it has provided enough information already, whether or not the scene has been described completely.) Examples of such a generation algorithm, based on the corpus of the present paper, can be found in Chen et al. (2019).

Acknowledgements

We thank the anonymous reviewers for their helpful comments. We thank Larry Moss, Jakub Szymanik, and Camilo Thorne for suggestions that helped shape our work. Guanyi Chen is supported by China Scholarship Council (No.201907720022).

References

Dale Barr, Kees van Deemter, and Raquel Fernández. 2013. Generation of quantified referring expressions: Evidence from experimental data. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 157–161, Sofia, Bulgaria. Association for Computational Linguistics.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence*, pages 241–301. Springer.

Guanyi Chen, Kees van Deemter, and Chenghua Lin. 2019. Generating quantified descriptions of abstract visual scenes. In *Proceedings of the 12th International Conference on Natural Language Generation*.

Kenny R Coventry, Angelo Cangelosi, Stephen E Newstead, and Davi Bugmann. 2010. Talking about quantities in space: Vague quantifiers, context and similarity. *Language and Cognition*, 2(2):221–241.

Kees van Deemter. 2016. *Computational models of referring: a study in cognitive science*. MIT Press.

Kees van Deemter, Albert Gatt, Ielka van der Sluis, and Richard Power. 2012a. Generation of referring expressions: Assessing the incremental algorithm. *Cognitive science*, 36(5):799–836.

Kees van Deemter, Albert Gatt, Ielka van der Sluis, and Richard Power. 2012b. Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, 36(5):799–836.

Albert Gatt and Anja Belz. 2010. Introducing shared tasks to nlg: The tuna shared task evaluation challenges. In *Empirical methods in natural language generation*, pages 264–293. Springer.

Albert Gatt, Ielka van der Sluis, and Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 49–56, Saarbrücken, Germany. DFKI GmbH.

Edward Grefenstette. 2013. Towards a formal distributional semantics: Simulating logical calculi with tensors. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 1–10, Atlanta, Georgia, USA. Association for Computational Linguistics.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Lisbon, Portugal. Association for Computational Linguistics.

Rosanna Keefe and Peter Smith. 1996. *Vagueness: A reader*. MIT press.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Andrzej Mostowski. 1957. On a generalization of quantifiers. *Fundamenta Mathematicae*, 44(2):12–36.

Linda M Moxey and Anthony J Sanford. 1993. *Communicating quantities: A psychological perspective.* Lawrence Erlbaum Associates, Inc.

Thomas Pechmann. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27(1):89–110.

Stanley Peters and Dag Westerstahl. 2006. *Quantifiers in language and logic*. Oxford University Press.

Mark A Schmuckler. 2001. What is ecological validity? a dimensional analysis. *Infancy*, 2(4):419–436.

Ionut Sorodoc, Angeliki Lazaridou, Gemma Boleda, Aurélie Herbelot, Sandro Pezzelle, and Raffaella Bernardi. 2016. "look, some green circles!": Learning to quantify from images. In *Proceedings of the 5th Workshop on Vision and Language*, pages 75–79, Berlin, Germany. Association for Computational Linguistics.

Jakub Szymanik and Marcin Zajenkowski. 2010. Comprehension of simple quantifiers: Empirical evaluation of a computational model. *Cognitive Science*, 34(3):521–532.

Jakub Szymanik et al. 2016. *Quantifiers and cognition: Logical and computational perspectives*, volume 96. Springer.

JFAK Van Benthem et al. 1986. *Essays in logical semantics*. Springer.

Amy Y Wang and Scott Piao. 2007. Translating vagueness? a study on translations of vague quantifiers in an english-chinese parallel corpus. In *Proceedings of the Corpus Linguistics Conference*.

Ilker Yildirim, Judith Degen, Michael K. Tanenhaus, and T. Florian Jaeger. 2013. Linguistic variability and adaptation in quantifier meanings. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.