University of Trento
CIMeC Doctoral School in Cognitive and Brain Sciences
Track Language, Interaction and Computation

Learning the Meaning of Quantifiers from Language and Vision

Sandro Pezzelle

Supervisor:
Raffaella Bernardi

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Cognitive and Brain Sciences

November 2018

Abstract

Sandro Pezzelle

Defining the meaning of *vague* quantifiers ('few', 'most', 'all') has been, and still is, the Holy Grail of a *mare magnum* of studies in philosophy, logic, and linguistics. The way by which they are learned by children has been largely investigated in the realm of language acquisition, and the mechanisms underlying their comprehension and processing have received attention from experimental pragmatics, cognitive psychology, and neuroscience. Very often their meaning has been tied to that of numbers, amounts, and proportions, and many attempts have been made to place them on ordered scales.

In this thesis, I study quantifiers from a novel, cognitively-inspired computational perspective. By carrying out several behavioral studies with human speakers, I seek to answer several questions concerning their meaning and use: Is the choice of quantifiers modulated by the linguistic context? Do quantifiers lie on a mental, semantically-ordered scale? Which are the features of such a scale? By exploiting recent advances in computational linguistics and computer vision, I test the performance of state-of-art neural networks in performing the same tasks and propose novel architectures to model speakers' use of quantifiers in grounded contexts. In particular, I ask the following questions: Can the meaning of quantifiers be learned from visual scenes? How does this mechanism compare with that subtending comparatives, numbers, and proportions?

The contribution of this work is two-fold: On the cognitive level, it sheds new light on various issues concerning the meaning and use of such expressions, and provides experimental evidence supporting the validity of the foundational theories. On the computational level, it proposes a novel, theoretically-informed approach to the modeling of vague and context-dependent expressions from both linguistic and visual data. By carefully analyzing the performance and errors of the models, I show the effectiveness of neural networks in performing challenging, high-level tasks. At the same time, I highlight commonalities and differences with human behavior.

Acknowledgements

My first, most heartfelt thanks are for my supervisor Raffaella Bernardi$_3$[1]. At the beginning of the journey, she told me: *Doing research is like running, it's a lot more fun if you do it with someone*. She was right, running with her was the best training ever. Thanks, Raffa, for the constant and wise guidance through good and bad moments, successes and failures. It is the backbone of this work. I am very grateful to the members of my Oversight Committee Manuela Piazza, Roberto Zamparelli$_6$ and Marco Baroni$_5$ for their invaluable feedback and support, and to the reviewers Judith Degen, Lucia Specia, and Jakub Szymanik for their precious comments and observations. Thanks to Leah Mercanti and the Doctoral Program Committee for making CIMeC a great place.

In these years, I was incredibly lucky to work with many amazing researchers. I am particularly grateful to Aurélie Herbelot$_2$, Ionut Sorodoc, Ravi Shekhar, and Marco Marelli for the cheek-to-cheek working hours, the outstanding discussions, the last-minute rushes. Thanks a lot to Gemma Boleda$_4$, Angeliki Lazaridou$_8$, Moin Nabi, Shane Steinert-Threlkeld, Jakub Szymanik, Germán Kruszewski$_7$, Raquel Fernández, Denis Paperno$_1$, Nghia The Pham$_{10}$, Enver Sangineto, Tassilo Klein, Francesca Franzon, Chiara Zanini, Addison Smith, Claudio Greco, Alberto Testoni, Alexander Kuhnle, Rossella Varvara. Thanks to the SAP Machine Learning team, working with you all was a great experience.

Thanks to the members of the 'Gilda club' Marta Mangiarulo, Flavio Ragni, Addison Smith, Ionut Sorodoc, and Simone Viganò for the *pizza-pasta combo*, the political forums, the ping-pong *debacles*. You became more than friends. Thanks to Mirko Broilo, Alessia De Felice, Carola Canella, Luca Ducceschi, Demetrio Ferro, Marco Pagani, Mariagrazia Popeo, Ben Timberlake for making Rovereto (and indievano) so special.

Last but not least, I would like to say *grazie* to mamma, papà, Annica, and Marco. To my second parents Egle and Rino, and my longtime friends Andrea, Marco, Salvo for being always there. Thanks to Alice. You keep smiling, we go far.

[1] Some people have a number in subscript. This is because they are in the picture! To find who is who, count them clockwise starting from the left-most standing person. Hint: the ninth is me.

Publications

This thesis collects several articles which have been published during my PhD. As such, most of the contents of this thesis have appeared in the following publications:

- Pezzelle, S., Steinert-Threlkeld, S., Bernardi, R., & Szymanik, J. (2018). *Some of* them can Be Guessed! Exploring the Effect of Linguistic Context in Predicting Quantifiers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 114–119). (Chapter 3)

- Pezzelle, S., Bernardi, R., & Piazza, M. (2018). Probing the mental representation of quantifiers. *Cognition, 181*, 117–126. (Chapter 4)

- Pezzelle, S., Marelli, M., & Bernardi, R. (2017). Be Precise or Fuzzy: Learning the Meaning of Cardinals and Quantifiers from Vision. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, (Vol. 2, pp. 337–342). (Chapter 5)

- Pezzelle, S., Sorodoc, I. T., & Bernardi, R. (2018). Comparatives, Quantifiers, Proportions: a Multi-Task Model for the Learning of Quantities from Vision. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (Vol. 1, pp. 419–430). (Chapter 6)

Moreover, some excerpts have appeared in the following publication:

- Sorodoc, I., Pezzelle, S., Dimiccoli, M., Herbelot, A., & Bernardi, R. (2018). Learning quantification from images: A structured neural architecture. *Natural Language Engineering, 24*(3), 363–392. (Section 2.6 and 2.7)

Sarà perché, anche se non ti conoscevo, è come se ti fossi stato compagno di banco.
Sarà per quel sorriso furbetto, che tradisce un'intelligenza curiosa, instancabile.
Sarà che nei tuoi capelli arruffati, nella tua barba scompigliata e nella tua determinazione vorremmo vederci tutti un po' riflessi.
Sarà perché uno slogan, uno soltanto, ce l'abbiamo pure noi.
E lo urliamo in silenzio, a testa alta.

Verità per Giulio Regeni

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A writer should have the precision
of a poet and the imagination
of a scientist.

VLADIMIR NABOKOV

There are many ways to communicate quantities. A football commentator, after yet another goal by the visiting team, might notice that '*Most* of the supporters of the home team are leaving the stadium'. Alternatively, he might state that 'The home fans who are leaving the stadium are *seven thousands*' or that, 'In the home fans area, there are *more* empty seats than occupied ones'. Similarly, he might say that '*Three quarters* (or *seventy-five percent*) of the home fans are leaving before the game ends'. Though referring to the same event, these sentences convey different quantitative information. The first, containing the quantifier *most*, is rather 'vague': The supporters who are leaving the stadium are likely more than half of the total and probably not all. The second, in contrast, is very 'precise':[1] We know the absolute *number* of people who are leaving, though we cannot infer whether they constitute the majority of the home fans or, rather, just a small fraction. The third, by means of a *comparative*, gives us a precise (though

[1] A note on the terminology. In this thesis, the term *vague* and the concept of 'vagueness' are used to refer to quantifiers whose interpretation can be borderline and not generally-agreed (Van Deemter, 2012). Consistently, I do not consider quantifiers like 'at most 5' or 'fewer than 8' as vague since these expressions establish a clear-cut division between two sets of numbers, such that 7 is undoubtedly less than 8 (Van Deemter, 2012). Similar reasons hold for numbers, comparatives, and proportions, that I therefore consider as *precise*. As a general note, it is worth mentioning that vague expressions such as 'few' or 'many' are not *ambiguous* as words like 'bank' or 'pitcher'. While the latter have several, well-defined and different meanings, the former have a non-specific but single meaning (Tuggy, 1993). I refer the reader to Van Deemter (2012) for a detailed discussion on vagueness and its relation with ambiguity.

rather coarse) answer to the previous question. The fourth, by specifying a *fraction* (*proportion*), provides us with the exact percentage of disappointed supporters.

Being merely quantitative, the meaning of number words, comparatives, fractions and proportions is straightforward. Being *vague* and context-dependent, the meaning of quantifiers is not. The former are clearly ordered on quantitative scales: 'one', 'two', 'three'; 'less', 'same', 'more'; 10%, 50%, 90%. The latter are often claimed to be, but both the existence and the nature of such a scale is highly debated (Holyoak and Glass, 1978; Routh, 1994; Moxey and Sanford, 2000). Indeed, the notion of a 'quantifier scale' has been largely investigated by psychological and psycholinguistic work aimed at linking the meaning of these expressions to scales of numbers, amounts, proportions (see Section 2.3). Though generally shared among scholars, the intuition that quantifiers are *ordered* terms (e.g. that 'very few' refers to something less than 'few') has been repeatedly shown to be more fragile than expected. For example, Moxey and Sanford (1993b) demonstrated that any *quantitative* difference between the quantifiers 'few', 'very few', 'only a few', 'not many', and 'a few' disappears when participants are prevented from comparing one expression against the others. To account for these results, Moxey and Sanford (1993b) proposed that the difference between these expressions, rather than quantitative, might be in the *perspective* they take to this information. Intuitively, this is not the case for numbers or proportions, where an ordering between elements on solely quantitative bases can always be found. Finally, the use of quantifiers has been shown to also depend on the context (Degen and Tanenhaus, 2015), expectations (Degen and Goodman, 2014), and individual differences among speakers (Yildirim et al., 2016).

One computational way to study the meaning of these expressions is using Distributional Semantics Models (DSMs) (Landauer and Dumais, 1997; Turney and Pantel, 2010; Baroni et al., 2014). Based on the hypothesis that similar words occur in similar contexts (Harris, 1954; Firth, 1957), DSMs use large *corpora* of texts to build meaning representations that encode statistics on word associations and co-occurrences. In standard *count* DSMs, the meaning of a word is initially represented as a N-dimensional vector encoding the raw frequency of the target word in each of the N contexts. The vector is further reduced/transformed by means of various techniques such as Singular Value Decomposition (SVD) to obtain a higher order semantic representation. A more recent approach exploits neural networks to *predict* word vectors (*embeddings*) on the basis of the surrounding words (Mikolov et al., 2013; Pennington et al., 2014). In both approaches, the resulting vectors are typically used to compute the degree of semantic similarity/relatedness between pairs of words. In particular, this measure is operational-

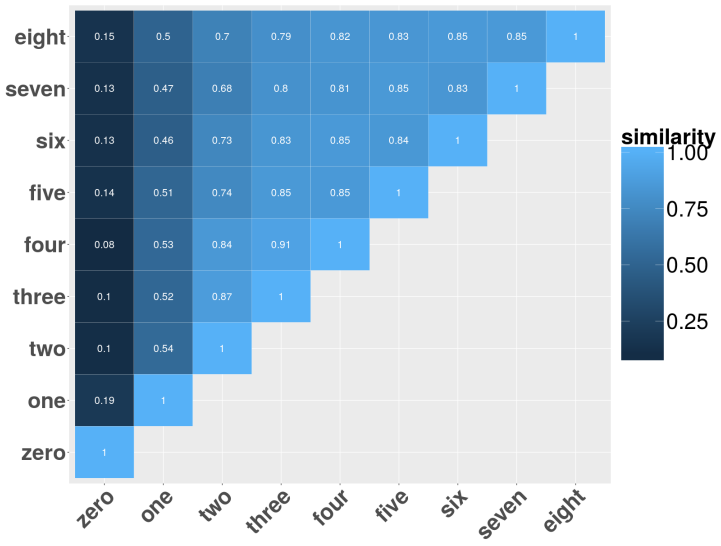

Figure 1.1: Heatmap reporting pairwise similarity between vectors of number words.

ized in terms of the cosine of the angle between the vectors: The closer the vectors, the higher their semantic similarity.

If quantifiers are semantically-ordered expressions, we should expect a measure of semantic similarity to be able to capture such a scale. For example, 'none' should be closer to 'few' compared to 'many', as well as, among numbers, 'two' should be closer to 'three' than to 'six'. The rationale is that, among ordered elements, items that are close to each other on the scale are expected to be more semantically similar compared to elements that are far. Here, I report the results of a proof-of-concept analysis performed on 9 number words (from 0 to 8) and 9 quantifiers (the same explored in Chapter 4 and Chapter 6). Word embeddings were obtained by training a state-of-the-art `word2vec` model (Mikolov et al., 2013) on the same corpus [2] and with the the best configuration of parameters used in Baroni et al. (2014). Then, pairwise similarities were computed.

As can be seen in Figure 1.1, the expected pattern is generally confirmed among numbers. Except for 'zero', which turns out to be very dissimilar from all other elements, increasing values from left to right (and from top to bottom) are observed for almost all cases. For example, the similarity between 'eight' and the other numbers starts very low (0.15 with 'zero') and slowly increases as soon as the numbers get higher: 0.5 with 'one', 0.7 with 'two', 0.79 with 'three', and so on. In contrast, the patterns of similarity among quantifiers (Figure 1.2) are much less straightforward: 'all' is closer to 'none'

[2]The corpus was previously pre-processed to ensure that multi-word quantifiers (e.g. 'the smaller part') are treated as a single word (i.e. 'the_smaller_part').

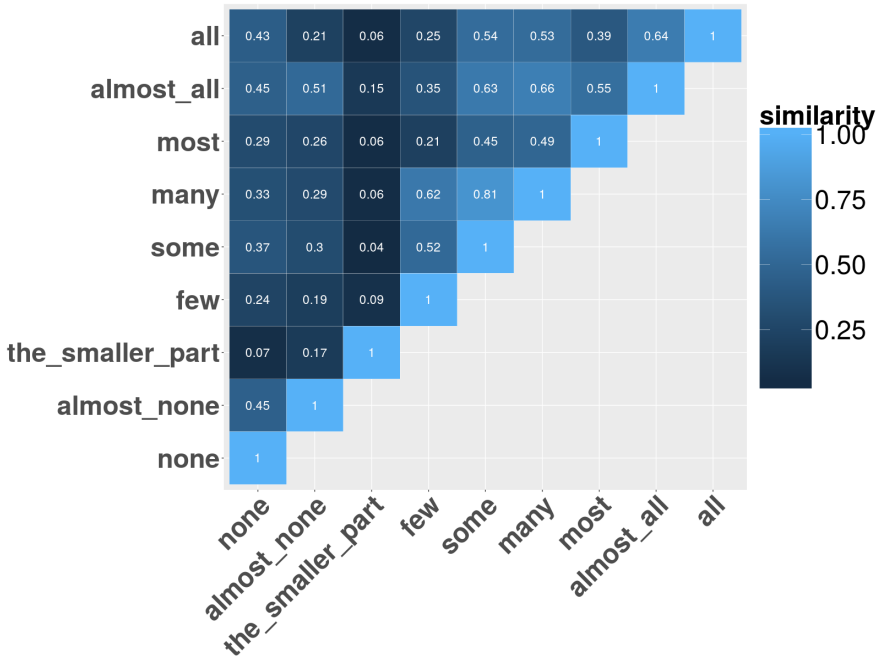

Figure 1.2: Heatmap reporting pairwise similarity between vectors of quantifier words.

(0.43) compared to 'most' (0.39), and 'few' is closer to 'many' (0.62) than to 'almost none' (0.19). Though just exploratory, such analysis confirms that numbers and quantifiers have a very different semantics. The former, perhaps except 'zero', display an almost exclusively quantitative meaning and are thus well ordered on a numerical scale. The latter, whose meaning is something more complex (and different) than numbers, amounts, proportions (Nouwen, 2010) display a much intricate pattern of similarities, possibly dependent on lexical-semantic besides quantitative factors.[3]

Coming back to our football commentator, it might be that his sentence '*Most* of the supporters of the home team are leaving the stadium' was not at all intended to tell us something about the number of people disappointed by the match. Perhaps his intention was just to express the sadness of the moment, and although the supporters leaving were just, e.g., one-fifth of the total, viewers at home were able to understand the reasons of his exaggeration. However, a more pragmatically plausible option is that the professional speaker wanted to reliably *describe* what was happening in the stadium, and after rapidly seeing that significantly more than half of the home-fans seats were empty, he said that to the microphone. If this was the case, the choice of using the quantifier 'most' was aimed at communicating a somehow 'objective' quantity that was

[3]For example, there seems to be an effect of *antonymy*, in a way that antonyms are generally similar to each other (see, e.g., 'none'-'all', 'almost none'-'almost all', etc.).

grounded in perception and, particularly, in vision. This possibility would be in line with the evidence that quantifiers' use, in visual contexts, is sensitive to quantitative information (Coventry et al., 2010; Degen and Tanenhaus, 2015) and, more in general, that quantity expressions are used and evaluated by speakers against real-life scenarios (Heim et al., 2012).

In this thesis, I investigate vague, non-numerical quantifiers ('none', 'few', 'almost all', 'many', 'all', etc.) from a novel, cognitively-inspired computational perspective. By carrying out several behavioral studies with human speakers, I seek to answer several questions concerning their meaning and use: Is the choice of quantifiers modulated by the linguistic context? Do quantifiers lie on a mental, semantically-ordered scale? Which are the features of such a scale? By exploiting recent advances in computational linguistics and computer vision, I test the performance of state-of-art neural networks in performing the same tasks and propose novel architectures to model the speakers' use of quantifiers in grounded contexts. In particular, I ask the following questions: Can the meaning of quantifiers be learned from visual scenes? How does this mechanism compare with that subtending comparatives, numbers, and proportions? The contribution of this work, thus, is two-fold: On the cognitive level, it sheds new light on various issues concerning the meaning and use of such expressions and provides experimental evidence supporting the validity of the foundational theories. On the computational level, it proposes a novel, theoretically-informed approach to the modeling of vague and context-dependent expressions from both linguistic and visual data. By carefully analyzing the performance and errors of the models, I show their effectiveness in performing challenging, high-level tasks while highlighting commonalities and differences with human behavior.

In Chapter 3, I study the role of linguistic context in modulating the choice of 9 frequently used English quantifiers. Tested in the challenging task of predicting a missing quantifier from either short or longer texts, humans and the models are shown to use different strategies, the former relying more on the information conveyed by the broader context, the latter being more effective in exploiting lexical-semantic cues. Moreover, both humans and the models make 'plausible' errors, that is, they are almost always able to grasp the 'magnitude' of the missing quantifier. This supports the idea that quantifiers, in language, are loosely ordered on some sort of quantitative scale. The characteristics of such a scale are explored in Chapter 4. By means of two behavioral experiments with human participants and a 'balanced' set of 9 Italian quantifiers, I show that quantifier words are mentally organized on an ordered, non-linear compressed scale which is sim-

ilar to that of perceptual quantities. Moreover, quantifiers turn out to be best predicted by proportional information when used to refer to objects in visual scenes. Both findings are in line with the idea that representations of quantifiers are mainly constructed by mapping them to the representations of quantities that we derive from perception. Along these lines, Chapter 5 explores the computational mechanisms underlying the learning of numbers and quantifiers from vision. I show that while numbers in the subitizing range require a model including a precise identification of the instances to be counted, quantifiers 'no', 'few', 'most', and 'all' are better learned by a model capitalizing on a fuzzy measure of similarity. Building on all this converging evidence, in Chapter 6 I use the same visual stimuli and the same 9 quantifier words explored in Chapter 4 and propose that comparatives ('more'), quantifiers ('most'), and proportions ('80%') can be jointly learned from visual scenes by means of a multi-task computational model. The motivation is that these expressions are governed by the same cognitive mechanism, which is different from that underlying numbers. By using I prove that sharing a core mechanism is beneficial for all these tasks, while numbers are shown to require a radically different operation.

In the next chapter, I briefly introduce the theoretical framework which motivates the questions explored in this work. While each of the following chapters is accompanied by a detailed and somehow more specific motivation, the aim of Chapter 2 is to provide a general overview of the problems connected with the semantics, the use, and the modeling of quantifiers. Some notions on the technical background subtending the computational models presented in the following chapters are also provided.

Chapter 2

Theoretical Framework

2.1 Formal Semantics: Relations between Sets

Studies on the semantics of quantifiers are dominated by Generalized Quantifier Theory (hence, GQT) based on the mathematical principles described by Mostowski (1957); Lindström (1966) and systematically applied to linguistics by Barwise and Cooper (1981); Keenan and Stavi (1986); van Benthem (1986).[1] The overall aim of GQT is to devise a general semantics for expressions of quantity by applying mathematical (or generalized) quantifiers to linguistics. Quantifier meanings are defined set-theoretically by means of categorical evaluation functions yielding either truth or falsity of a sentence in which a quantifier is present. As such, quantifiers are conceived as *non-referential*: They do not denote objects, but instead relations between sets of objects.

The core idea is that a quantifier like 'some' or 'every' expresses a relation between two sets. The GQT formalization includes a typology of quantifiers. In particular, noun/determiner phrases (i.e. 'some donkeys') correspond to the type (1) quantifier. This type is called (1) because it expresses an *unary* relation, that is a set. Determiner-like quantifiers like 'some' or 'every' represent the type $(1,1)$, where $(1,1)$ stands for a *binary* relation, that is a relation between two sets. To illustrate:

1. some(A, B) is true iff $\|A\| \cap \|B\| \neq \varnothing$

2. many(A, B) is true iff $\|A\| \cap \|B\| > n$, where n is some large number

[1] See Peters et al. (2006) for an exhaustive overview.

That is, the sentence 'some donkeys fly' is true if and only if the intersection of the donkeys and the flying things is not empty. That implies that the sentence always holds truth except in the case when no donkeys fly. In other words, it is true either when only one donkey out of all donkeys in the world can fly or when all of them do. In the case of 'many donkeys fly', the sentence is true if the cardinality of the flying donkeys is larger than some contextual norm n.

In formal semantics, there exists an extensive literature on quantifiers whose meanings depend on such a contextual norm, like 'few' and 'many' (Partee, 1989; Solt, 2009). Partee (1989), for example, proposes that 'few' and 'many' are ambiguous because of the nature of n, which can stand for either a *cardinal* or a *proportion*. The idea is further extended and formalized by Solt (2009) in terms of 'scale' structures. In a nutshell, the cardinal reading would arise when the involved scale does not display a clear upper bound (hence, the scale is numerical). In contrast, the presence of an upper bound would license the proportional reading (hence, the scale is made of proportions). Crucially, such formalization is not aimed at mapping the set-theoretic definition to any well-defined scale of numbers or proportions. However, it highlights two core features of quantifiers: They are vague and context-dependent (see section 2.3).

Another interesting distinction within GQT has been proposed between *first order* (FO) and *higher order* (HO) quantifiers (van Benthem, 1986). The former class includes quantifiers that are definable in first-order logic and can be computed by simple devices without cycles (e.g. finite *automata*, that is, simple idealized machines used to either accept or reject an input). In contrast, the latter class includes quantifiers which are not definable in first-order logic and require computability models using some internal memory (see Szymanik and Zajenkowski (2010)). That is, the meaning of the latter would require some more complex operations to be recognized and verified in a context compared to the former. By definition, FO include Aristotelean quantifiers such as 'no', 'some', 'all' as well as cardinal/numerical quantifiers like 'at least three', 'at most two'. The reason is that Aristotelean can be *translated* into numerical ones. For example, 'some' can be rephrased as 'at least one', 'no' as 'at most zero', and so on. In contrast, HO include both proportional quantifiers such as 'more than half', 'most' and parity quantifiers such as 'an even/odd number of', whose comprehension would require to keep some information in the memory. According to Clark (2011), GQT formalization would thus imply a direct connection between quantifiers and numbers interpretation (see section 2.4).

To sum up, GQT defines the *semantics* of quantifiers in terms of set relations. As such, the meaning of these expressions is logically unambiguous. Though extremely powerful, the formalization provided by GQT has been repeatedly shown to be poorly connected with the pragmatic *use* of quantifiers (Nouwen, 2010). For example, the interpretation of 'some' as 'at least one and possibly all' has appeared to be too broad and coarse-grained compared to speakers' use in real contexts. In the next section, I briefly review the pragmatic approach on quantifiers, which is aimed at studying the use and interpretation of quantifiers in real-communication contexts.

2.2 Pragmatics: Scalar Implicatures

The pragmatic approach focuses on the informative strength of *utterances* (i.e., units of speech) containing a quantifier. Typically, the focus is on a particular type of *implicature*, called 'scalar implicature' (Grice, 1975), which consists in the attribution of an *implicit* meaning that is neither expressed nor strictly implied by the utterance containing the quantifier. For example, in the utterance '*Some* of the home supporters left the stadium', the use of 'some' would *implicate* that '*Not all* of the home supporters left the stadium'. Crucially, this view is in contrast with GQT (see section 2.1), according to which 'some' would be logically consistent with 'all', in a way that using the former term would not exclude that 'all' of the supporters are leaving the stadium. In conversational settings, however, speakers are ordinarily required to be as informative as possible (but not more informative than required). Therefore, the choice of using a given quantifier would be determined by its position on the *implicational* scale, which ranges from informatively weaker to stronger elements. Horn (1972), for example, proposed the following scale, ordered from weaker to stronger elements: 'one', 'some/a few', 'several', 'many', 'half', 'most/a majority', 'all/every'.

Scalar implicatures have been largely investigated in experimental pragmatics, where the focus is on how they are computed by listeners in real-time language comprehension. Across the various accounts proposed, scalar implicatures have been considered as either a *default* (Levinson, 2000), a *literal-first* (Huang and Snedeker, 2009), or a *context-driven* (Breheny et al., 2006) process. According to the default view, generating the implicature would be immediate and effortless. According to the literal-first view, they would require computing the literal meaning first. According to the context-driven view, both the robustness and the speed with which a scalar implicature is computed

would depend on multiple cues that are available in the context. Evidence for the latter possibility was brought by Degen and Tanenhaus (2011), who employed a 'gumball' paradigm (i.e., visual scenes depicting a variable number of gumballs) to investigate the role of various cues in affecting the scalar implicature of 'some'. Their results showed that the syntactic form of the quantifier phrase, the availability of alternatives, and the size of the referred set affect various aspects of the processing of the scalar implicature. The same experimental paradigm was employed by Degen and Tanenhaus (2015) to explore the 'naturalness' of quantifiers and number terms when used to refer to sets containing a variable number of gumballs (ranging from 0 to 13). 'Some' turned out to be more natural in some settings (e.g., when referring to small sets) compared to others (e.g., when referring to the set containing all 13 gumballs), thus bringing new evidence in favor of the context-driven view of scalar implicatures. Further work strengthened this claim by showing that both prior knowledge (Degen and Goodman, 2014) and the availability of lexical alternatives (Degen and Tanenhaus, 2016) have an early role in the pragmatic utterance interpretation.

By supporting the hypothesis that scalar implicatures vary on the basis of various contextual factors, this line of work brings important evidence in favor of the vague and context-dependent status of quantifiers (see section 2.3). At the same time, the pragmatic approach postulates the existence of a quantifier scale whose elements are clearly ordered on the basis of their informative strength. While pragmatics is crucial to explore the use and interpretation of quantifiers, it does not directly focus on the general *semantics* of these expressions. Instead of answering the question 'What does *some* mean?', it rather focuses on questions like 'What does the use of *some* implicate in an utterance?' or 'Under which circumstances and to what extent the use of *some* implicates, e.g., *not all*?'. In the next section, I review some linguistic and psychological work aimed at studying the meaning of quantifiers from a *quantitative* perspective, namely by linking their semantics to scales of *exact* numbers or proportions. Crucially, contextual factors are often not taken into account in these accounts, based on the assumption that the meaning of quantifiers is well-defined and stable across situations.

2.3 Quantifiers, Quantities, and Contextual Effects

One of the very first attempts to link the meaning of quantifiers to *exact* quantities is represented by Graves and Hodge (1943), who normatively assigned a percentage to a

large number of quantifying expressions such as 'none' (0%), 'a part' (20%), 'not much' (10%), 'the greater part' (70%), and so on. Although the aim of the work was to help writers to properly use these expressions in English, this proposal is interesting for at least two reasons. First, it overtly assumes that the meaning of quantifiers is defined by *proportions* – not, e.g., by absolute numbers. Second, the percentage assigned to each quantifier is thought to be *fixed* and not affected by any contextual effect.

To empirically test these assumptions, Hammerton (1976) designed an experimental setup where participants were asked to assign percentages ranging from 0 to 100 to quantifiers embedded in sentences. The same set of quantifiers by Graves and Hodge (1943) was used. Overall, the results of this study replicated the previously-defined percentages, thus supporting both the validity of the proposed scale and the hypothesis of prototypical *focal* ranges associated with each quantifier.

While intriguing, such a well-defined picture has been repeatedly shown to become much less clear when taking into account a number of factors. For instance, Moxey and Sanford (1993a) demonstrated that when preventing participants from comparing one quantifier versus another (i.e., when removing lexical alternatives; see section 2.2) in the task of assigning a precise number to a given quantifier word, any difference between quantifiers 'a few', 'only a few', 'not many', 'few', and 'very few' disappeared. Moreover, they showed that the number assigned to a given quantifier heavily depends on the context, with e.g. '*lots of* stars in the sky' being matched with a rather different number compared to e.g. '*lots of* typos in this thesis'. Similarly, Newstead and Collis (1987) found that low-magnitude quantifiers such as 'few' and 'several' refer to greater percentages when describing small sets compared to larger sets. That is, the assigned percentage is affected by the cardinality of the set and thus not stable across conditions.

Since quantifiers are often used for communication purposes that are different or wider in scope compared to that of conveying quantity information, many scholars maintained that they cannot be simply considered as words that stand for numbers, amounts, proportions (Paterson et al., 2009; Nouwen, 2010). The supporting evidence is provided by the fact that in sentences like 'There are *many* people in this queue' the meaning of 'many' could depend on speaker's expectations (e.g., he/she thought there was a shorter queue) and psychological attitude (e.g. he/she does not like waiting in a queue) besides purely quantitative aspects. Moreover, quantifier meanings have been shown to depend on both listeners' adaptation to the statistics of the linguistic environment (Yildirim et al., 2013) and talker variability (Yildirim et al., 2016).

While a certain correspondence between quantifier meanings and scales of exact numbers or proportions is observed, such correspondence has been repeatedly proved to be affected by a wide range of contextual factors. Indeed, these factors have appeared to be something more than simple pragmatic add-ons to numerical information, being responsible of affecting the *semantics* of quantifiers besides their *interpretation*. Though it is generally accepted that quantities alone cannot account for the whole meaning of such expressions, however, it has been proposed that the quantitative aspects of quantifier semantics are better linked to an *approximate* – rather than exact – representation of quantities. In the next section, I discuss work aimed at exploring this connection from a cognitive and neuroscience perspective.

2.4 Quantifiers and the Brain

A crucial issue in the psychological, developmental, and neuroscience literature on quantifiers is determining which kind of numerical information, if any, underlies their comprehension and meaning. Despite their commonalities with cardinals (e.g. 'one', 'two', 'eleven') with respect to a number of syntactic, semantic and pragmatic properties (Hurewitz et al., 2006), quantifiers have been shown to differ from cardinals in several respects. First, even though they are both learned in a fairly stable order of acquisition across languages (Wynn, 1992; Katsos et al., 2016), they are handled differently by the language acquisition system. That is, children who lack exact cardinality concepts are able to understand and appropriately use quantifiers in grounded contexts (Halberda et al., 2008; Barner et al., 2009). This indicates that knowledge about (large) precise numbers is neither necessary nor sufficient for learning the meaning of quantifiers. Second, adult speakers are able to reliably answer questions involving quantifiers even in contexts that preclude counting (Pietroski et al., 2009). This evidence suggests that the semantics of quantifiers relies on a mechanism of numerosity estimation based on the Approximate Number Sense (ANS), that is a pre-verbal system for the representation of numerical magnitude (Feigenson et al., 2004; Piazza, 2010). The key feature of ANS is that it is not precise, and it becomes less precise with increasing magnitudes. Moreover, the power to discriminate among sets varies according to the numerical ratio in observance of Weber's law (Piazza and Eger, 2016).

Some interesting insight on the interplay between numerical information and quantifiers meaning has emerged from fMRI studies. The issue has been firstly investigated

by McMillan et al. (2005), who conjectured that precise number sense is required in order to understand quantifiers. To test their hypothesis, they carried out a neuroimaging study where participants were presented with a sentence containing a quantifier (e.g. 'some apples are green') followed by a visual scenes containing both target (i.e. 'green apples') and distractor objects (i.e. 'non-green apples'). Participants were asked to judge the truth of the sentence with respect to the visual stimulus. Their results showed that all quantifiers recruited right inferior parietal cortex (IPC), that is the area typically associated with numerosity processing (see for a review Kadosh et al. (2008)). These findings led the authors to claim that precise numerical information is required for understanding all types of quantifiers (see also Clark and Grossman (2007)). Similar conclusions were drawn by Heim et al. (2012), who performed a complex parametric study to investigate the neural networks involved in the comprehension and verification of proportional quantifiers. Overall, their results revealed that numerical processing is required to understand (proportional) quantifiers in grounded contexts.

A different pattern of results was found by Troiani et al. (2009), who focused on the distinction between Aristotelean (e.g. 'some', 'all') and numerical quantifiers (e.g. 'at least three', 'an odd number of'). The aim of the study was to show that the latter are associated with numerical information, whereas the former are not. Consistent with their hypotheses, only numerical quantifiers were found to be supported by a parietal-dorsolateral prefrontal network (in IPC) depending on quantity-based or numerical processing. Logical quantifiers, in contrast, turned out to be associated with rostral medial prefrontal cortex involved in elementary logic operations, and supported by a selective visual-spatial attention mechanism in posterior cingulate cortex. The authors claimed that such a dissociation is in line with the two separate learning processes reported in children acquisition of numbers and quantifiers (Hurewitz et al., 2006; Papafragou and Schwarz, 2006; Halberda et al., 2008). Consistent results were obtained by Morgan et al. (2011), who investigated the neural representation of logical/Aristotelean (e.g. 'some', 'all'), cardinal (e.g. 'at least three'), and majority (e.g. 'at least half') quantifiers in patients with corticobasal syndrome (CBS), posterior cortical atrophy (PCA), and behavioral variant frontotemporal dementia (bvFTD).

Similarly to Troiani et al. (2009), a dissociation was found between (a) cardinal (i.e. requiring quantity processing) and (b) logical-majority quantifiers (i.e. requiring executive functioning). Using a semantic distance judgment task, Wei et al. (2014) investigated brain activation for six types of materials, including Arabic digits (e.g. '1', '2'), number words (e.g. 'one', 'two'), dot arrays (e.g. '•', '• •'), and quantifiers (i.e. 'none',

'few', 'several', 'some', 'many', 'abundance', 'myriad'). Their results showed a clear dissociation between the quantity processing of quantifiers and that of numbers and numerosities. In particular, the latter stimuli elicit more activation in the right intraparietal sulcus (IPS) than quantifiers do. Also, the processing of quantifiers turned out to be more associated with brain regions for general semantic processing, namely left middle temporal gyrus and inferior frontal gyrus. This findings led the authors to claim that, consistently with the results by Troiani et al. (2009), 'pure' quantifiers are not processed in IPS, but rather in the brain's language areas.

To wrap up, McMillan et al. (2005) reported a similar activation in IPC for all quantifiers (e.g. 'at least three' and 'some') in *grounded* contexts. Similarly, Heim et al. (2012) demonstrated a role of IPS during both estimation and comparison, which are required steps for assessing the validity of a proportional quantifier. In contrast, Wei et al. (2014) reported no activation in IPS for any quantifiers in a semantic judgment task. Finer-grained dissociations were found by Troiani et al. (2009) and Morgan et al. (2011), who showed that only numerical quantifiers elicit brain areas associated with quantity processing in grounded tasks. Overall, these results might suggest that numerical information comes into play for some classes of quantifiers (i.e. numerical, proportional, parity), but not for others (i.e. Aristotelean). Moreover, it seems to be involved when a 'quantitative' interpretation of quantifiers is explicitly required, namely in grounded contexts. In the next section, I discuss some behavioral work aimed at exploring the role of quantitative information in visually-grounded quantifiers.

2.5 Quantifiers Grounded in Vision

To explore the quantitative features of quantifiers meaning, a few behavioral studies investigated their use in *grounded* contexts. A first study by Newstead and Coventry (2000) employed visual stimuli depicting a bowl and a number of black dots to test the role of object size in affecting the use of five quantifiers ('few', 'a few', 'several', 'many', 'lots of'). By manipulating the number of the dots and the size of both the dots and the bowl, they found that low-magnitude quantifiers (e.g. 'few') were more appropriate when the dots were small and the bowl was big, with an opposite trend for high-magnitude quantifiers (e.g. 'many').

Coventry et al. (2005) used visual scenes containing both striped and white fish to investigate the role of a number of perceptual factors in affecting quantifiers appropriate-

ness. They varied (a) number of target (range 3-18) and distractor objects (range 0-18), (b) spacing between objects, (c) spatial disposition of the objects in the scene (either grouped or mixed). All these factors turned out to affect quantifiers interpretation, but only when the number of targets exceeded the 'subitizing' range (i.e., the range of cardinalities, typically up to 3-4, which can be automatically and precisely enumerated; see Revkin et al. (2008)). That is, the meaning of low-magnitude quantifiers turned out to be 'stable' and somehow not affected by other factors than target cardinality.

The same set of quantifiers used by Coventry et al. (2005) was also explored by Coventry et al. (2010), who investigated (a) how judgments about quantifiers are affected by the presence of distractor objects and (b) whether the kind and function of objects affect the judgments. They employed visual stimuli where targets and distractors were either semantically similar (*men-women*) or different (*men-crocodiles*). Moreover, they manipulated the *function* of both target and distractor objects (*playing golf-not playing golf*). In all cases, a reliable effect of the number of distractors was observed. Moreover, in contrast with Coventry et al. (2005), the number of distractors was found to play a role also in the subitizing range.

Finally, an unpublished paper (van Tiel et al., in preparation) used visual stimuli to investigate whether the *focal ranges* (i.e. prototypical numbers/proportions) associated with quantifiers match the traditional semantic formalization (e.g., that 'half' is equal to exactly 50%). To test their hypotheses, the authors experimented with visual scenes where the proportion of red and black dots varied. Participants were asked to produce a quantifier to describe the scene. The results showed that the proportion of target dots associated with each quantifier did not clearly match the expected interpretation.

Overall, these studies indicate that the meaning of quantifiers in grounded contexts is mostly described by quantitative features such as either the cardinality of the sets or the proportion of target objects in the scene. On the one hand, this suggest that quantifiers are mentally represented on an ordered, quantitative scale whose representation and components, however, none of these studies explicitly investigated. On the other, these findings support the hypothesis that at least some components of the meaning of quantifiers are directly connected with approximate numerical information, partly in line with the evidence reported in section 2.4. In the next sections, I discuss computational approaches to the modeling of quantifiers in language (section 2.6) and previous work aimed at extracting quantity information from visual inputs (section 2.7).

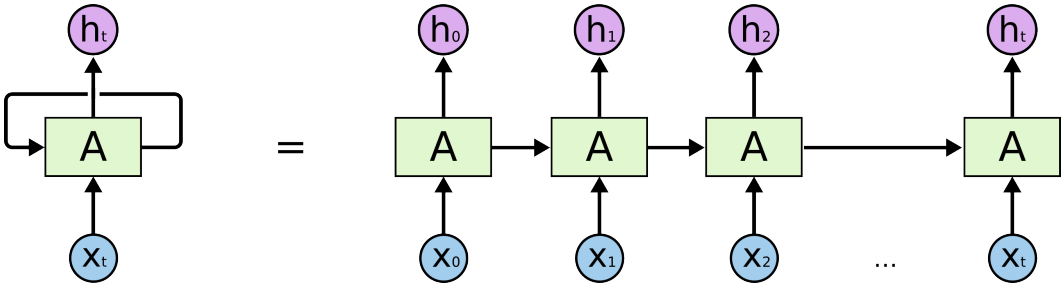

Figure 2.1: Schematic representation of an unrolled Recurrent Neural Network (RNN). Among RNNs, Long-Short Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) have become particularly popular in the Natural Language Processing (NLP) community. What makes this kind of RNNs special is their ability to learn long-term dependencies. That is, these networks can 'remember' information for long periods of time, which in natural language means that they can be used to handle long sequences of text. In a nutshell, LSTMs are structured as a chain of modules of neural network, usually called 'cells'. Each cell processes the state coming from the preceding cell, and forwards information to the following one. While processing the information, the cell 'decides' what information is important to keep (and what to forget) to perform the task. Starting from continuous representations or embeddings of words (which can be learned 'from scratch' or pre-computed using, e.g., the methods proposed by Mikolov et al. (2013); Pennington et al. (2014)), LSTMs can be trained to make predictions for virtually any NLP task, e.g. Sentiment Analysis, Question Answering, etc. Image credits: Colah's blog on Understanding LSTM Networks.

2.6 Modeling Quantifiers: Computational Linguistics

The problem of algorithmically describing logical quantifiers was first addressed by van Benthem (1986) using automata (see section 2.1). Following these first efforts, a lot of work has been done in computational formal semantics to model quantifiers in language (see e.g. Szabolcsi (2010); Keenan and Paperno (2012); Szymanik (2016) for a in-depth overview). For example, Szymanik and Zajenkowski (2010) compared the time needed for understanding different types of quantifiers and showed a psychologically-relevant distinction between quantifiers recognized by different types of automata.

Recently, distributional semantics (see Chapter 1) has turned to the problem, with Baroni et al. (2012) demonstrating that some entailment relations hold between quantifier vectors obtained from large corpora, and Herbelot and Vecchi (2015) mapping a distributional vector space to a formal space from which the quantification of a concept-property pair can be predicted. By focusing on the distributional representation of 'every', Capetola (2013), showed the limitations of such an approach in modeling the dynamic representation of quantification. One way to overcome these limitations has been

proposed by Lewis and Steedman (2013), who showed the benefits of combining distributional semantics with formal logical semantics for the representation of function words such as quantifiers. Overall, this work highlighted the limitations of the distributional approach in capturing the semantics of quantifiers (see also the results of our exploratory study in Chapter 1).

In recent years, quantifiers have received renewed attention along with the explosion of neural networks for language modeling (see Figure 2.1 for a schematic representation of a Recurrent Neural Network (RNN) and a brief description of one of the most popular architectures, namely Long-Short Term Memory (Hochreiter and Schmidhuber, 1997)). These models have been applied, for example, to solve the tasks of Natural Language Inference (Nangia et al., 2017; Ghaeini et al., 2018) and Question Answering (Andreas et al., 2016), where quantifiers were among the cases used to evaluate the models in those specific tasks. However, no previous work exploited these architectures to specifically explore quantifiers and their semantic representation.

2.7 Modeling Quantities: Computer Vision

The first attempt to model quantification mechanisms from visual inputs dates back to Dehaene and Changeux (1993). Using a forerunner neural network, this study showed that approximate numerosity could be extracted from a visual input without serial counting, bringing computational evidence to the psycholinguistic observation that infants develop numerosity abilities before being able to count. More recently, Rajapakse et al. (2005) used a similar network to reproduce the human use of quantifiers in grounded contexts. The model was trained on human annotations of images consisting of white and stripy fish (from Coventry et al. (2005)). Given an image, the model had to predict which number of fish was stripy, using the given quantifiers. The authors showed that both spacing and the number of objects played a role in the prediction. Crucially, both these studies were carried out before the revolutionary advent of Convolutional Neural Networks (CNNs)[2], which gave rise to a new era in the field of Computer Vision (see Figure 2.2 for a schematic representation of VGG-16 (Simonyan and Zisserman, 2014), one of the most popular and successful CNNs for image feature extraction).

Exploiting CNNs, a number of works in Computer Vision have proposed specific ar-

[2]See LeCun et al. (2015) for a general but detailed overview on CNNs.

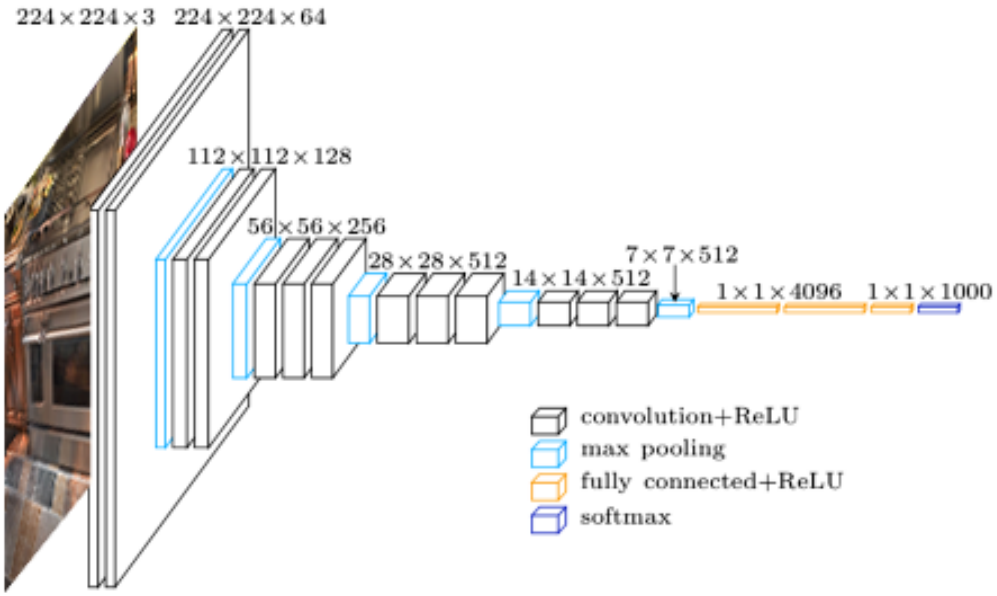

Figure 2.2: Schematic representation of VGG-16 (Simonyan and Zisserman, 2014), one of the most popular Convolutional Neural Networks (CNNs) for image feature extraction. CNNs are designed to process data that is structured in multiple arrays. This is the case of color images, which are composed of three 2D arrays containing pixel intensities in the three RGB channels. Generally speaking, the architecture of a CNN includes various types of layers: Convolutional layers followed by non-linear transformations (e.g. ReLu), pooling layers, and fully-connected layers. Convolutional layers are used to extract local features from the input image while preserving the spatial relationships between pixels. The role of pooling layers, instead, is to merge semantically similar features into one. Finally, fully-connected (FC) layers are Multi-Layer Perceptrons (MLPs) whose units are connected to every unit in the subsequent layer. FCs encode high-level features of the input image, such as information on the object class (e.g. 'dog'). Indeed, the final FC is typically used to perform object classification by means of a *softmax* activation function. Note that VGG-16 includes 13 convolutional layers, 5 pooling layers, and 3 FC layers followed by softmax. Image credits: abtosoftware.com

chitectures for counting digits (Seguí et al., 2015), people in the crowd (Zhang et al., 2015a), or penguins (Arteta et al., 2016). With a more cognitive flavor, Chattopadhyay et al. (2017) proposed a 'divide-and-conquer' strategy to split the image into subparts and count the objects in each subpart by mimicking the subitizing mechanism (see section 2.5). Inspired by the same cognitive ability is Zhang et al. (2015b), who trained a CNN to detect and count the salient objects in the image. Except Suhr et al. (2017), who built a dataset for visual reasoning to be evaluated against various types of quantity expressions including existential quantifiers, however, these works exclusively focused on exact numbers.

Focused on the modeling of approximate quantities is Stoianov and Zorzi (2012), who experimented with hierarchical generative models and showed their effectiveness in learning ANS as a statistical property of synthetic images. Tested on the task of set comparison ('more/less'), their proposed networks were shown to obtain a remarkable 93% accuracy. As for quantifiers, to our knowledge no previous studies focused on the learning of such expressions from visual scenes. Besides the studies reported in this dissertation, two other works from our group[3] tackled these issues. In particular, Sorodoc et al. (2016) proposed a model to assign the correct quantifier to synthetic scenes of colored dots, whereas Sorodoc et al. (2018) operationalized the same task in a Visual Question Answering (VQA) fashion, using real images and object-property queries (e.g. 'How many *dogs* are *black*?'). Overall, the results of these studies showed that vague quantification can be learned by neural networks, though the performance is much lower when using real images and complex queries. Moreover, in both studies, quantifiers were simplistically operationalized in terms of ranges of proportions (as in Chapter 5). In this thesis, I seek to overcome this issue by collecting (Chapter 4) and modeling human data (Chapter 6).

[3]For an overview, see quantit-clic.github.io

Chapter 3

Quantifiers and Linguistic Contexts

In this chapter, I study the role of linguistic context in modulating the choice of quantifiers. Tested in the task of predicting a missing quantifier from a *local* context (single-sentence) and a *global* context (multi-sentence) condition, humans and state-of-the-art computational models are shown to use different strategies: The former are boosted by the information conveyed by the broader context, the latter are more effective in exploiting local lexical-semantic cues. Overall, both humans and the models make 'plausible' errors, that is, they are almost always able to grasp the 'magnitude' of the missing quantifier. This supports the idea that quantifiers are loosely ordered on a quantitative scale.

3.1 Introduction

Cloze deletion test (Oller, 1973) is a typical exercise which is used to evaluate a language learner. In this task, a word is removed and learners must exploit their language abilities to understand the context and the vocabulary in order to identify the correct word. Since the comprehension of the missing word is boosted by the surrounding linguistic context, the larger the linguistic context, the easier the test becomes. Indeed, it has been recently shown that higher-ability test takers rely more on global information, with lower-ability test takers focusing more on the local context, namely information contained in the words immediately surrounding the gap (McCray and Brunfaut, 2018).

In this chapter, I exploit a cloze-test setting and explore the role of linguistic context in predicting quantifiers (see Figure 3.1). Both human and model performance is evaluated in a *local* (single-sentence) and a *global* context (multi-sentence) condition to study

<qnt> *the island's breeding birds are endemic.*

The island is one of the world's most biologically diverse areas, with many endemic species.
<qnt> *the island's breeding birds are endemic. Other endemic species include the red-bellied lemur, the indri, and the aye-aye.*

Target quantifier: ***more than half of***

Figure 3.1: Given a target sentence s_t, or s_t with the preceding and following sentence, the task is to predict the target quantifier replaced by <qnt>.

the role of context and assess the cognitive plausibility of the models. As discussed in Chapter 2, quantifiers are of central importance in linguistic semantics and its interface with cognitive science (Barwise and Cooper, 1981; Peters and Westerståhl, 2006; Szymanik, 2016). Moreover, the choice of quantifier is known to depend both on local context (e.g., positive and negative quantifiers license different patterns of anaphoric reference) and global context (the degree of positivity/negativity is modulated by discourse specificity) (Paterson et al., 2009). Finally, and more generally, the ability of predicting *function words* in the cloze test has been shown to represent a benchmark test for human linguistic competence (Smith, 1971; Hill et al., 2016a).

Our conjecture is that human performance will be boosted by more context and that this effect will be stronger for *proportional* quantifiers (e.g. 'few', 'many', 'most') than for *logical* quantifiers (e.g. 'none', 'some', 'all') because the former are more dependent on discourse context (Moxey and Sanford, 1993a; Solt, 2016). In contrast, we expect models to be very effective in exploiting the local context (Hill et al., 2016a) but to suffer with a broader context, due to their reported inability to handle longer sequences (Paperno et al., 2016). Both hypotheses are confirmed. The best models are very effective in the local context condition, where they significantly outperform humans. Moreover, model performance declines with more context, whereas human performance is boosted by the higher accuracy with proportional quantifiers like 'many' and 'most'. Finally, best-performing models and humans are found to make similar errors. In particular, they tend to confound quantifiers that denote a similar 'magnitude', namely they confound e.g. 'most with 'many', but not e.g. 'few' with 'almost all' (Bass et al., 1974; Newstead and Collis, 1987).

The contribution of this chapter is twofold. First, a new task and results for training models to learn semantically-rich function words are presented.[1] Second, the role of linguistic context in both humans and the models is carefully analyzed, with implications for cognitive plausibility and future modeling work.

3.2 Related Work

Studies on the interplay between linguistic context and *function words* date back at least to Smith (1971). In this study, it was claimed that (a) function words are easier to predict in a cloze test than content words and (b) larger context is beneficial for content words but detrimental for function words. The main reason for (a) is that predicting function words implies choosing among a limited number of options, whereas content words have much many alternatives. Strictly related, the main reason for (b) is that function words would depend more on clues that are immediately close to the deleted word rather than on the 'meaning' of the broader context (Rankin and Thomas, 1980). Though generally considered as belonging to the class of function words, quantifiers display a somehow hybrid status. Indeed, they are semantically-rich expressions whose meaning has been usually tied to some sort of quantitative information (Graves and Hodge, 1943; Bass et al., 1974; Hammerton, 1976; Newstead and Collis, 1987). As such, their choice has been shown to depend both on local and global context (Paterson et al., 2009). For example, the presence of a *local* Polarity Item (PI) like 'any' ('*none of* them has *any* constraints') restrict the choice only to those quantifiers that can license it (Krifka, 1995). Moreover, quantifiers like 'few' or 'many' are dependent on a *contextual norm* (Partee, 2008; Solt, 2009), whose cardinality can be inferred from the meaning of the (broader) surrounding context.

Computational models have been extensively tested on the cloze test. However, most previous work (see, among others, Hermann et al. (2015); Onishi et al. (2016)) has focused on content words and named entities, whereas there has been little interest in modeling function words. A notable exception is represented by Hill et al. (2016a), who evaluated a number of models in the task of predicting prepositions besides verbs, nouns and named entitities. Crucial for our purposes, they showed that Long-Short Term Memory (LSTM) models outperform humans in predicting prepositions ('on', 'at', etc.). Moreover, adding context decreases their performance. Based on this evidence, the

[1] Data and code can be found at `github.com/sandropezzelle/fill-in-the-quant`

authors claimed that LSTM predictions are almost exclusively based on local contexts. Similar conclusions can be drawn from recent work challenging computational models with larger and more sophisticated language contexts (Paperno et al., 2016; Chu et al., 2016). In these studies, state-of-the-art models were shown to fail in predicting words that require understanding the broader context.

Focusing on quantifiers, a class of semantically-rich function words, we follow a similar approach and test how the models' and humans' performance compare in the two settings. To our knowledge, we are the first investigating the effect of linguistic context in predicting these expressions.

3.3 Datasets

To test our hypotheses, we need linguistic contexts containing quantifiers. To ensure similarity in the syntactic environment of the quantifiers, we focus on partitive uses: where the quantifier is followed by the preposition 'of'. To avoid any effect of intensifiers like 'very' and 'so' and adverbs like 'only' and 'incredibly', we study only sentences in which the quantifier occurs at the beginning (see Figure 3.1). We experiment with a set of 9 quantifiers: 'a few', 'all', 'almost all', 'few', 'many', 'more than half', 'most', 'none', 'some'. This set strikes the best trade-off between number of quantifiers and their frequency in our *source* corpus, a large collection of written English including around 3B tokens.[2]

We build two datasets. One dataset – `1-Sent` – contains datapoints that only include the sentence with the quantifier (the *target* sentence, s_t). The second – `3-Sent` – contains datapoints that are 3-sentence long: the target sentence (s_t) together with both the preceding (s_p) and following one (s_f). To directly analyze the effect of the linguistic context in the task, the target sentences are exactly the same in both settings. Indeed, `1-Sent` is obtained by simply extracting all target sentences $<s_t>$ from `3-Sent` ($<s_p$, s_t, $s_f>$).

The `3-Sent` dataset is built as follows: (1) We split our source corpus into sentences and select those starting with a '*quantifier* of' construction. Around 391K sentences of this type are found. (2) We tokenize the sentences and replace the quantifier at the beginning of the sentence (the *target* quantifier) with the string `<qnt>`, to treat all

[2] A concatenation of BNC, ukWaC, and a 2009-dump of Wikipedia Baroni et al. (2014).

type	text	quantifier
PIs	*<qnt> these stories have **ever** been substantiated.*	*none of*
contrast Q	*<qnt> the population died out, but **a** select **few** with the right kind of genetic instability...*	*most of*
list	*<qnt> their major research areas are **social inequality, group dynamics, social change**...*	*some of*
quantity	*<qnt> those polled (**56%**) said that they would be willing to pay for special events...*	*more t. half of*
support Q	*<qnt> you have found this to be the case - click here for **some of** customer comments.*	*many of*
lexicalized	*<qnt> **the time**, the interest rate is set on the lender's terms...*	*most of*
syntax	*<qnt> these events **was** serious.*	*none of*
meaning	*<qnt> the original station buildings survive **as they were used** as a source of materials...*	*none of*

Table 3.1: Cues that might help human participants to predict the correct quantifier (`1-Sent`).

target quantifiers as a single token. (3) We filter out sentences longer than 50 tokens (less than 6% of the total), yielding around 369K sentences. (4) We select all cases for which both the preceding and the following sentence are at most 50-tokens long. We also ensure that the target quantifier does not occur again in the target sentence. (5) We ensure that each datapoint $<s_p, s_t, s_f>$ is unique. The distribution of target quantifiers across the resulting 309K datapoints ranges from 1152 cases ('more than half') to 93801 cases ('some'). To keep the dataset balanced, we randomly select 1150 points for each quantifier, resulting in a dataset of 10350 datapoints. This was split into train (80%), validation (10%), and test (10%) sets while keeping the balancing. Then, `1-Sent` is obtained by extracting the target sentences $<s_t>$ from $<s_p, s_t, s_f>$.

3.4 Human Evaluation

3.4.1 Method

We ran two crowdsourced experiments, one per condition. In both, native English speakers were asked to pick the correct quantifier to replace `<qnt>` after having carefully read and understood the surrounding linguistic context. When more than one quantifier sounds correct, participants were instructed to choose the one they think best for the context. To make the results of the two surveys directly comparable, the same randomly-sampled 506 datapoints from the validation sets are used. To avoid biasing responses, the 9 quantifiers were presented in alphabetical order. The surveys were carried out via CrowdFlower.[3] Each participant was allowed to judge up to 25 points. To assess the judgments, 50 unambiguous cases per setting were manually selected by the native-English author and used as a benchmark. Overall, we collected judgments from

[3] https://www.figure-eight.com/

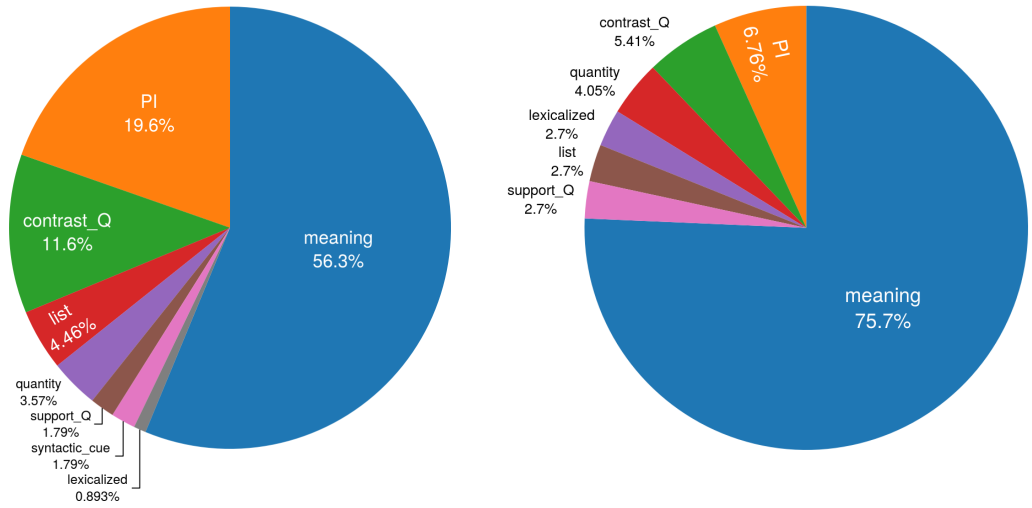

Figure 3.2: Left: Distribution of annotated cues across correcly-guessed cases in `1-Sent` (112 cases). Right: Distribution of cues across correctly-guessed cases in `3-Sent`, but not in `1-Sent` (74 cases).

205 annotators in `1-Sent` (avg. 7.4 judgments/annotator) and from 116 in `3-Sent` (avg. 13.1). Accuracy is then computed by counting cases where at least 2 out of 3 annotators agree on the correct answer (i.e., inter-annotator agreement ≥ 0.67).

3.4.2 Linguistic Analysis

Overall, the task turns out to be easier in `3-Sent` (131/506 correctly-guessed cases; 0.258 accuracy) compared to `1-Sent` (112/506; 0.221 acc.). Broader linguistic context is thus generally beneficial to the task. To gain a better understanding of the results, we analyze the correctly-predicted cases and look for linguistic cues that might be helpful for carrying out the task. Table 3.1 reports examples from `1-Sent` for each cue.

By carefully looking into the sentences used for the experiment, we identify 8 main types of cues and manually annotate the cases accordingly. Annotation is performed by one of the authors by reading the target sentences several times and checking for the presence of any of the following cues: (1) **PIs**: Polarity Items like 'ever', 'never', 'any' that are licensed by specific quantifiers (e.g., the sentence '**most of* the students have *ever* been here' is ungrammatical; see Krifka (1995)); (2) **Contrast Q**: a contasting-magnitude quantifier embedded in an adversative clause (e.g. '*few of* the Xs . . . but *most* (of the) Ys'); (3) **Support Q**: a supporting-magnitude quantifier embedded in a coordinate or subordinate clause (e.g. '*some of* Xs . . . and possibly *many* (of the)

text	quantifier
*a number of **examples of technophobic ideas** can be found in multiple forms of art, ranging from literary works such as "Frankenstein" to classic films like "Metropolis". <qnt> these works portray the darker side of technology as seen by the technophobic. As technologies become increasingly complex and difficult to understand, people are more likely to harbor anxieties relating to their use of modern technologies.*	*many of*
*you have highlighted the fact that there is **very limited business experience** within the teaching profession. <qnt> us have experienced industry over an extensive period. Apprenticeships in Germany and in other places are linked very tightly with the business community.*	*few of*
*the **weather goes smoothly** over the points of union betwixt the twin summers. <qnt> the storms are very loud or variable. The average temperature during the day, in December, was about sixty-five degrees in the shade, but on one day a little damp snow fell.*	*few of*
*by 1995 there were 120 of them, receiving tuition in: fiddle bagpipes drums tin whistle keyboard guitar voice drama at enrolment, each young person is offered the choice of tuition on up to three different instruments. <qnt> them choose an instrument they already play for their first choice and the tutors look to see a significant improvement in their proficiency at the end of the week. Tutors, however, also actively **encourage the children to try something new**.*	*most of*

Table 3.2: Examples of cases that are correctly guessed in `3-Sent` (but not in `1-Sent`). Linguistic context that appears to be particularly helpful to retrieve the correct quantifier is in **bold**.

Ys'); (4) **Quantity**: explicit quantitative information (numbers, percentages, fractions, etc.) immediately following the quantifier (e.g. '*few of* the Xs (around 10%)'); (5) **Lexicalized**: lexicalized patterns like '*most of* the time'; (6) **List**: the text immediately following the quantifier is a list introduced by verbs like 'are' or 'include' comprising at least 3 elements; (7) **Syntax**: morpho-syntactic cues, e.g. agreement (e.g. '*none of* Xs ... was happy'); (8) **Meaning**: the quantifier can only be guessed by understanding and reasoning about the context. It is worth mentioning that (8) is assigned by the annotator only if none of the cues from (1) to (7) are found.

Figure 3.2 (left) depicts the distribution of annotated cues in correctly-guessed cases of `1-Sent`. Around 44% of these cases include cues besides meaning, suggesting that almost half of the cases can be possibly guessed by means of lexical factors such as PIs, quantity information, etc. As seen in Figure 3.2 (right), the role played by the meaning becomes much higher in `3-Sent`. Of the 74 cases that are correctly guessed in `3-Sent`, but not in `1-Sent`, more than 3 out of 4 do not display cues other than meaning. In the absence of lexical cues at the sentence level, the surrounding context thus plays a crucial role, as reported in Table 3.2. By looking at these examples, it is clear that the presence of the preceding and following sentence makes the task more feasible compared to the presence of the target sentence only. This role is particularly accentuated in quantifiers like 'many', 'almost all', and 'most', where correctly-guessed cases annotated as relying on semantic information only represent 100%, 100%, and 85% cases, respectively.

3.5 Models

We test several models, that we briefly describe below. All models except `FastText` are implemented in Keras and use `ReLu` as activation function; they are trained for 50 epochs with categorical crossentropy, initialized with frozen 300-d `word2vec` embeddings (Mikolov et al., 2013) pretrained on GoogleNews.[4] A thorough ablation study is carried out for each model to find the best configuration of parameters.[5] The best configuration is chosen based on the lowest validation loss.

BoW-conc A bag-of-words (BoW) architecture which encodes a text as the *concatenation* of the embeddings for each token. This representation is reduced by a hidden layer before softmax.

BoW-sum Same as above, but the text is encoded as the *sum* of the embeddings.

FastText Simple network for text classification that has been shown to obtain performance comparable to deep learning models (Joulin et al., 2016). `FastText` represents text as a hidden variable obtained by means of a BoW representation.

CNN Simple Convolutional Neural Network (CNN) for text classification.[6] It has two convolutional layers (`Conv1D`) each followed by `MaxPooling`. A dense layer precedes softmax.

LSTM Standard Long-Short Term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997). Variable-length sequences are padded with zeros to be as long as the maximum sequence in the dataset. To avoid taking into account cells padded with zero, the 'mask zero' option is used.

[4] Available here: `http://bit.ly/1VxNC9t`

[5] We experiment with all possible combinations obtained by varying (a) optimizer: *adagrad*, *adam*, *nadam*; (b) hidden layers: 64 or 128 units; (c) dropout: 0.25, 0.5, 0.75.

[6] Adapted from: `http://bit.ly/2sFgOE1`

	1-Sent		3-Sent	
	val	*test*	*val*	*test*
chance	0.111	0.111	0.111	0.111
BoW-conc	0.270	0.238	0.224	0.207
BoW-sum	0.308	0.290	0.267	0.245
fastText	0.305	0.271	0.297	0.245
CNN	0.310	0.304	**0.298**	0.257
LSTM	0.315	0.310	0.277	0.253
bi-LSTM	0.341	**0.337**	0.279	0.265
Att-LSTM	0.319	0.324	0.287	**0.291**
AttCon-LSTM	**0.343**	0.319	0.274	0.288
Humans	0.221*	——	0.258*	——

Table 3.3: Accuracy of models and humans. Values in **bold** are the highest in the column. *Note that due to an imperfect balancing of data, chance level for humans (computed as majority class) is 0.124.

bi-LSTM The Bidirectional LSTM (Schuster and Paliwal, 1997) combines information from past and future states by duplicating the first recurrent layer and then combining the two hidden states. As above, padding and mask zero are used.

Att-LSTM LSTM augmented with an attention mechanism (Raffel and Ellis, 2016). A feed-forward neural network computes an importance weight for each hidden state of the LSTM; the weighted sum of the hidden states according to those weights is then fed into the final classifier.

AttCon-LSTM LSTM augmented with an attention mechanism using a learned *context* vector (Yang et al., 2016). LSTM states are weighted by cosine similarity to the context vector.

3.6 Results

Table 3.3 reports the accuracy of all models and humans in both conditions. We have three main results. (1) Broader context *helps* humans to perform the task, but *hurts* model performance. This can be seen by comparing the 4-point increase of human accuracy from `1-Sent` (0.22) to `3-Sent` (0.26) with the generally worse performance of all models (e.g. `AttCon-LSTM`, from 0.34 to 0.27 in *val*). (2) All models are significantly *better* than humans in performing the task at the sentence level

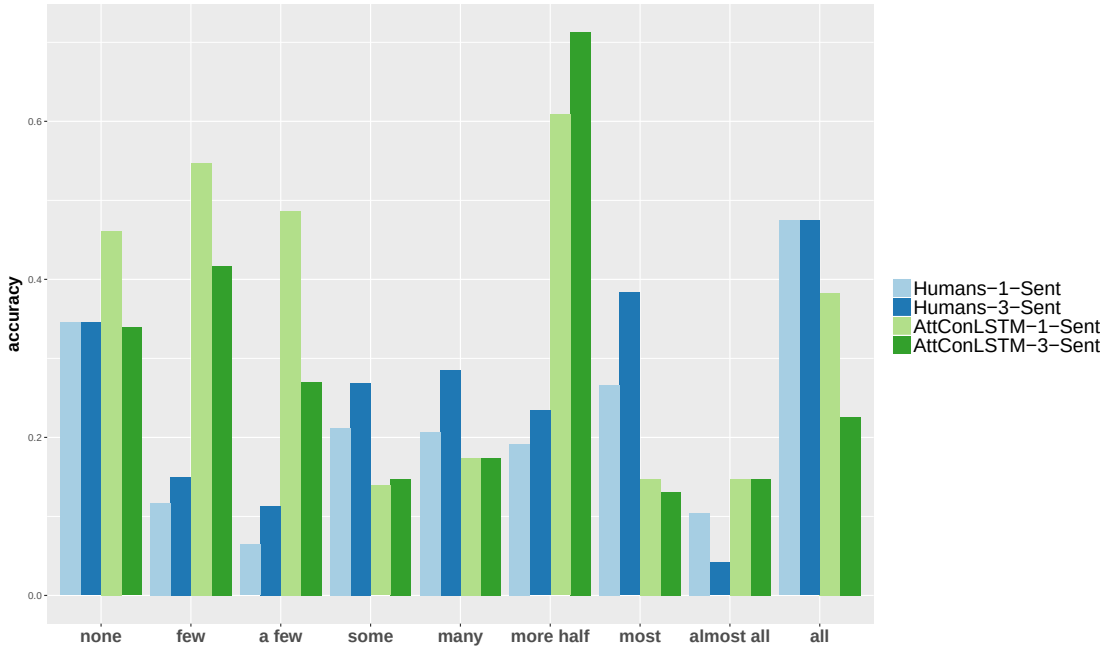

Figure 3.3: Human vs `AttCon-LSTM` accuracy (*val*) across quantifiers, loosely ordered by magnitude.

(`1-Sent`), whereas their performance is only slightly better than humans' in `3-Sent`. `AttCon-LSTM`, which is the best model in the former setting, achieves a significantly higher accuracy than humans' (0.34 vs 0.22). By contrast, in `3-Sent`, the performance of the best model is closer to that of humans (0.29 of `Att-LSTM` vs 0.26). It can be seen that LSTMs are overall the best-performing architectures, with `CNN` showing some potential in the handling of longer sequences (`3-Sent`). (3) As depicted in Figure 3.3, quantifiers that are easy/hard for humans are not necessarily easy/hard for the models. Compare 'few', 'a few', 'more than half', 'some', and 'most': while the first three are generally hard for humans but predictable by the models, the last two show the opposite pattern. Moreover, quantifiers that are guessed by humans to a larger extent in `3-Sent` compared to `1-Sent`, thus profiting from the broader linguistic context, do not experience the same boost with models. Human accuracy improves notably for 'few', 'a few', 'many', and 'most', while model performance on the same quantifiers does not.

To check whether humans and the models make similar errors, we look into the distribution of responses in `3-Sent` (*val*), which is the most comparable setting with respect to accuracy. Table 3.4 reports responses by humans (top) and `AttCon-LSTM` (bottom). Human errors generally involve quantifiers that display a similar magnitude as the correct one. To illustrate, 'some' is chosen in place of 'a few', and 'most' in place

none	**19**	1	2	0	2	0	0	0	12
few	5	**9**	2	6	5	0	3	0	2
a few	0	0	7	**17**	9	0	4	0	4
some	0	0	3	**14**	5	0	4	0	3
many	0	1	0	3	**18**	0	3	0	7
more than half	0	0	0	2	2	**11**	10	4	2
most	0	0	0	1	7	0	**23**	4	8
almost all	0	1	0	3	2	1	**7**	2	6
all	0	0	2	1	5	0	4	3	**28**
none	**39**	15	13	10	0	20	5	3	10
few	3	**48**	18	7	9	20	5	1	4
a few	7	13	**31**	18	5	15	12	8	6
some	5	18	16	17	16	**19**	9	5	10
many	2	18	18	15	**20**	17	10	6	9
more than half	2	7	2	3	10	**82**	2	1	6
most	8	14	14	12	12	**26**	15	5	9
almost all	5	9	15	10	8	**37**	15	6	10
all	7	12	10	15	21	13	7	4	**26**

Table 3.4: Responses by humans (top) and `AttCon-LSTM` (bottom) in `3-Sent` (*val*). Values in **bold** are the highest in the row.

of either 'almost all' or 'more than half'. A similar pattern is observed in the model's predictions, though we note a bias toward 'more than half'. Zooming into human responses, an interesting, bucking case is represented by the frequent choice of 'all' in place of 'none' (but never *vice versa*). On the one hand, this pattern seems to suggest an interchangeability of the quantifiers at the extremes of the quantifier scale, possibly due to their less context-dependent status in the absence of clear morpho-syntactic cues. On the other hand, the direction of the effect indicates that, when in doubt, the 'positive' interpretation is always preferred by speakers.

One last question concerns the types of linguistic cues exploited by the model (see section 3.4.2). We consider those cases which are correctly guessed by both humans and `AttCon-LSTM` in each setting and analyze the distribution of annotated cues. Though limited to a subset of datapoints, such analysis should be indicative of the overall behavior of the model: if the model genuinely understands the meaning of the text and mostly capitalizes on semantic information, we should consequently observe a higher number of cases that are annotated as containing only semantic information. In contrast, if the model learns associations between a quantifier and specific lexical items or morpho-syntactic patterns, we should observe a higher number of correct responses within datapoints displaying these cues.

Figure 3.4 (left) depicts the distribution of cues among the 44 cases that are correctly predicted by both speakers and the model in `1-Sent`. As can be observed, half of the

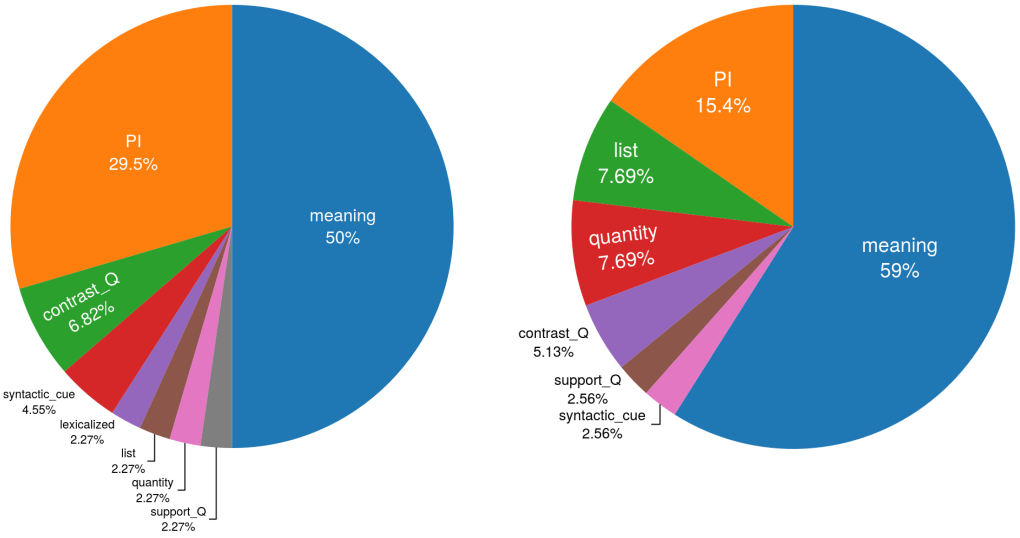

Figure 3.4: Left: Distribution of cues exploited by `AttCon-LSTM` across cases correctly-guessed by speakers in `1-Sent` (44 cases). Right: Distribution of cues across cases correctly-guessed by speakers in `3-Sent` (39 cases).

cases contain lexical or morpho-syntactic cues. That is, they might be guessed by effectively learning associations in the linguistic data. Zooming into these cases, it is worth mentioning that 83% and 80% cases of correctly-guessed 'none' and 'few', respectively, are annotated as containing cues other than meaning. A similar distribution can be observed in the rightmost panel reporting the 39 cases guessed by both humans and the model in `3-Sent`, where the non-semantic cues represent 41% cases. Though higher than in `1-Sent`, the number of cases that cannot guessed by exploiting cues other than meaning is still relatively low, especially when compared to the distribution oberved in speakers' responses (rightmost panel of Figure 3.2). Such analysis suggests that the model capitalizes more on lexical, morpho-syntactic information rather than exploiting the meaning of the context, either local or global. Since this observation is in contrast with that reported for human performance, that is observed to be significantly boosted by the meaning conveyed by the broader context, we conjecture this to be the main difference between speakers and humans. In the absence of lexical or morpho-syntactic cues, speakers use semantic information conveyed by the global context, whereas models employ this strategy to a much lesser extent.

3.7 Discussion

3.7.1 Context Dependence

In this chapter, I explored the role of linguistic context in predicting quantifiers. I showed that, for humans, the task becomes easier when a broader context is given. For the best-performing LSTMs, broader context hurts performance. This pattern mirrors evidence that predictions by these models are mainly based on local contexts, in line with Hill et al. (2016a). Corroborating our hypotheses, *proportional* quantifiers ('few', 'many', 'most') were found to be predicted by humans to a significantly higher accuracy when the broader context was provided, whereas *logical* quantifiers ('all', 'none') did not experience a similar boost. This finding supports the claim that proportional quantifiers are more context-dependent than are logical ones (Moxey and Sanford, 1993a; Solt, 2016).

It is worth mentioning that, overall, the accuracy in the task was found to be extremely low, both for humans and the models. This result could be due to several reasons, such as the difficulty of the dataset and/or the inherent overlapping use of the quantifiers employed in the study. To better investigate the former issue, the same experiment could be replicated by using linguistic contexts coming from different sources. To explore the latter, one possibility could be to experiment with a smaller and perhaps less overlapping set of quantifiers. Intuitively, the availability of less alternatives might make the task easier.

3.7.2 Mental Scale

Interestingly, humans revealed to be almost always able to grasp the 'magnitude' of the missing quantifier, even when picking up the 'wrong' one. This finding, on the one hand, is consistent with the well-reported overlapping meaning and use of these expressions (Moxey and Sanford, 1993a). On the other hand, it provides indirect evidence to the existence of a mental, ordered scale of quantifiers, an issue that has been largely debated in literature (Holyoak and Glass, 1978; Routh, 1994; Moxey and Sanford, 2000). It is worth mentioning, however, that such a scale appears rather coarse, with speakers often confounding quantifiers with similar magnitudes (e.g 'a few' with 'some' and 'almost all' with 'all'). Moreover, differently from Moxey and Sanford (1993a), in our

task the whole list of alternatives was always provided to people.

In the next chapter, I explore the nature and the characteristics of the mental scale of quantifiers by means of two behavioral experiments.

Chapter 4

Probing the Quantifier Scale: Two Behavioral Studies

In this chapter, I study the mental representation of non-numerical quantifiers by comparing their use in *abstract* and in *grounded* perceptual contexts. Using an approach similar to that used in the number domain, I test whether (and to what extent) such representation is constrained by the way we perceive the world through our senses. In two experiments, participants are asked to either judge the similarity of quantifier pairs (presented as written words) or chose among a predetermined list of quantifiers the one that best described a visual image depicting a variable number of target and non-target items. The results are rather consistent across experiments, and indicate that quantifiers are mentally organized on an ordered but non-linear compressed scale where the quantifiers that imply small quantities appear more precisely differentiated across each other compared to those implying large quantities. This fits nicely with the idea that we construct our representations of such symbols mainly by mapping them to the representations of quantities that we derive from perception.

4.1 Introduction

One of the common goals of linguists and cognitive scientists is to uncover and formally characterize how linguistic symbols are mentally represented. In this chapter, I tackle this issue by focusing on quantifiers, a class of words that had long been considered

as particularly intriguing especially by linguists due to their peculiar properties (see Chapter 2).

First, from a formal semantic perspective they are conceived as non-referential (Montague, 1973; Barwise and Cooper, 1981; Westerståhl, 1985; van Benthem, 1986; Keenan and Stavi, 1986; Szabolcsi, 2010): Differently from many other words, quantifiers do not denote objects, but instead relations between sets of objects. Second, quantifiers are widely affected by the linguistic context of use. This particularly holds for some quantifiers, like 'few' and 'many', which have therefore been proposed to be non-extensional (Keenan and Stavi, 1986; Westerståhl, 1985): The two sentences 'Many doctors attended the meeting this year' and 'Many lawyers attended the meeting this year' (even assuming that the doctors and lawyers attending the respective meetings are equal in number) might have different truth values depending on the number of doctors and lawyers who used to attend the meeting. Third, from a pragmatic perspective it has been shown how the different degree of information or logical strength of the quantifiers (that 'some' is less informative than 'all') affects the implicit information that people *infer* from an utterance (Horn, 1984). For example, listening to the sentence 'Some students were satisfied with the marks' a hearer would infer that 'Not all the students were satisifed'. Fourth, quantifiers cannot be simplistically considered as words that stand for amounts, numbers, proportions (Moxey and Sanford, 1993b, 2000; Paterson et al., 2009; Nouwen, 2010). Even when expressing approximately the same quantity (e.g. 'few' and 'a few'), quantifiers differ from each other with respect to the perspective they give to this quantity, by bringing the hearer to focus on either the target set ('a few') or the non-target set ('few'). For instance, 'few of these cars break down' is likely to bring the hearer's attention to the vast majority of cars that do not break down. 'A few of these cars break down', instead, is more likely to bring the attention to the cars that do break. This difference in the focus influences the hearer's behavior in a positive/negative way (Moxey and Sanford, 2000; Paterson et al., 2009). Consequently, quantifiers have been described in terms of probability distributions over scales (Moxey and Sanford, 1993b; Yildirim et al., 2013; Schöller and Franke, 2017). Finally, the variability of quantifiers across conditions, together with their rather elusive status with respect to the traditional linguistic classifications, have brought some researchers to take the extreme stance that devising a general semantics for these expressions might not even be possible (Nouwen, 2010).

Although a long tradition of studies convincingly proved that numerical information, such as the mechanisms of quantity estimation and comparison, is fundamental in the

comprehension of quantifiers (Heim et al., 2012; Shikhare et al., 2015; Deschamps et al., 2015),[1] cognitive science has not been successful at characterizing how humans mentally represent quantifiers. Historically, even if there has been a shared intuitive assumption that quantifiers might be internally represented on an ordered scale (which some conceived as governed by absolute quantities, e.g. Newstead et al. (1987), and other by proportions, e.g. Graves and Hodge (1943); Hammerton (1976)), there has been little attempt at formally trying to capture the features of such scale in a quantitative manner. One approach has been to investigate the conditions of the external world that trigger the use of the different quantifiers: Subjects, presented with sets of a various number of target and non-target (visual) items, are asked either to pick, among a predetermined list, the quantifier that best fits the scene or to rate the appropriateness of a list of scene-quantifier associations. Studies of this sort are only very few, and they are hard to compare as they each investigate different sets of quantifiers, as well as slightly different aspects of the stimuli (some analyze the effect of the number of targets, e.g. Newstead and Coventry (2000), some the number of both targets and non-targets, e.g. Coventry et al. (2005, 2010), some the proportion of targets in the scene, e.g. Oaksford et al. (2002), often taking into account perceptual factors like the size of the items, their spatial arrangements or their category, e.g. Newstead and Coventry (2000); Coventry et al. (2010)), though without investigating the potential interactions across all the possible variables. Moreover, the experimental design of all these studies lacks cases where the various effects can be disentangled, for example visual scenes with a small number of targets corresponding to a high proportion (e.g., 3 targets out of 4 total objects).

Although with some inconsistencies, the results of these studies overall suggest that quantifiers are evaluated by taking into account the number of both targets and non-targets such that, given a fixed number of non-targets, scenarios with increasing targets are associated with quantifiers implying 'larger' quantities. A notable exception is that, when the targets are very few, the number of non-targets seems not to play a role (Coventry et al., 2005). This indirectly suggests that quantifiers might be represented on an internal scale based on proportions which behaves somewhat differently for small sets. What these studies lack, however, is a quantitative characterization of the laws subtending the relation between quantifiers and perceptual stimulation and thus a

[1] These works typically employ a verification task: Given a scene depicting a variable proportion of target and non-target dots and a sentence embedding a quantified expression, participants are asked to quickly verify the semantic truth value of the sentence. What these studies showed is that errors and reaction times are typically affected by perceptual difficulty in observance to Weber's law.

thorough description of the internal scale.

Another complementary approach that psychologists have used to infer the structure of mental representations is that of directly asking subjects to compare words pairwise and to rate, on a given scale, their semantic similarity in a purely linguistic context (with no direct relation to concrete objects/sets). This way, the potential confounds due to the constraints imposed by perception are eliminated. In this approach, the analysis of the global pattern of rated distances across words can then be used to reconstruct the internal geometry of the representational space of those words (using Multi-Dimensional Scaling, e.g. Arnold (1971); Steyvers et al. (2004)). To our knowledge, this approach has been applied to the domain of quantifiers only by Holyoak and Glass (1978), who experimented with a set of five items. Studies of this sort would be crucial for complementing the studies that explore quantifiers in grounded conditions. In particular, the comparison across the grounded and abstract use of quantifiers is useful to approach the question of to what extent the mental representations of quantifiers (and, more generally, of symbols) are, or are not, constrained by the way we perceptually elaborate the objects or objects features to which the symbols are typically used to refer to.

While the abstract view of semantics predicts that symbols are mainly organized according to purely linguistic variables (frequency of use, frequency of association in the lexicon, antinomy, etc.), the grounded cognition view predicts that symbols are mentally represented in a way that at least partially reflects (or is isomorphic to) the way we perceive the world through our senses. This should be reflected both in how subjects use quantifiers to describe perceptual scenes, and in purely abstract contexts when they evaluate quantifiers among each other. This approach has been taken for example in the number domain, where several pieces of data indicate that the internal representation of number symbols (words or Arabic digits, denoting cardinals) appears as governed by the same representational constraints that govern the perception of numerosities in concrete sets, namely on an internal scale which appears overall logarithmically compressed (see Piazza and Eger (2016), for a recent review). This is the case both when number symbols are compared among each other and when they are used to describe perceptual scenes (e.g. Izard and Dehaene (2008)).

The aim of this chapter is to export this approach to study the mental space of quantifiers, its main dimensions, and its internal geometry, and to contrast the predictions from the abstract cognition and the grounded cognition comparing grounded-perceptual and abstract tasks: Using a common list of quantifiers and two large groups of subjects, one

experiment investigates quantifiers in grounded conditions, asking subjects to describe visual scenes choosing the most appropriate quantifier (Experiment 1), and the other investigates quantifiers in a purely linguistic context, asking subjects to rate the similarity among quantifier word pairs (Experiment 2).

4.2 Methods

Two experiments were administered to native-Italian participants and employed the same set of 9 Italian quantifiers. The quantifiers used were *nessuno* ('none'), *quasi nessuno* ('almost none'), *la minor parte* ('the smaller part'), *pochi* ('few'), *alcuni* ('some'), *molti* ('many'), *la maggior parte* ('most'), *quasi tutti* ('almost all'), *tutti* ('all'). For sake of clarity, English translations will be used from now on throughout the chapter. The selection of the quantifiers was aimed at experimenting with a fairly comprehensive set, including logical-Aristotelian ('none', 'some', 'all'), proportional ('the smaller part', 'most'), and a range of other common quantifiers ('few', 'many', 'almost none', 'almost all'). Moreover, an equal number of low-magnitude ('none', 'almost none', 'few', 'the smaller part') and high-magnitude quantifiers ('many', 'most', 'almost all', 'all') was ensured. Note that we did not consider 'some' as belonging *a priori* to one or the other group.

4.2.1 Grounded Task: Quantifiers Used in Perception

Thirty native-Italian participants (21 females, 9 males) with normal or corrected-to-normal vision carried out the task of evaluating 340 synthetic visual scenes containing two categories of objects: Animals and artifacts. The total number of objects in the scene ranged from 3 to 20 (see section 4.2.1 for a detailed description of the visual stimuli), and the number of items in each of the two categories varied from 0 to 20. The experiment was implemented in Matlab using the Psychtoolbox-3 package. All participants performed the experiment in a quiet, dimly lit room at the CIMeC Psychophysic lab (Rovereto, Italy) using the same desktop computer, same monitor (size 23.6', resolution 1920x1080 pixels), and same mouse, and sitting at a distance of approximately 50cm from the screen. Eighteen participants requested and obtained university credits for their participation.

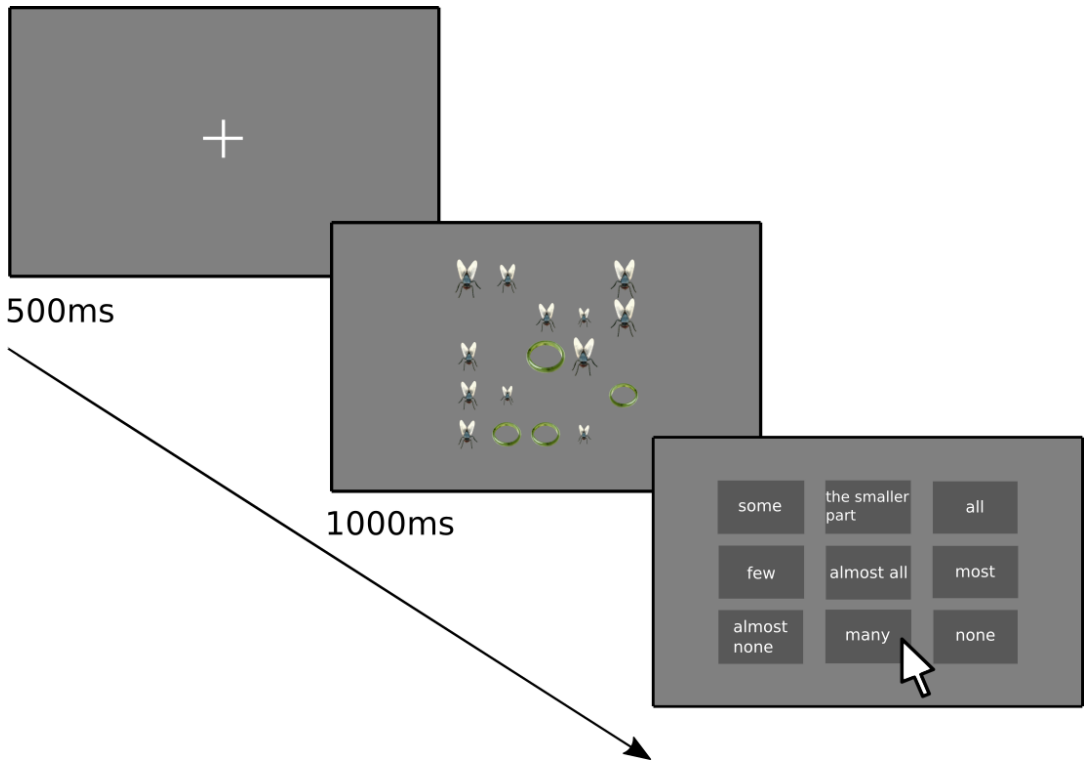

Figure 4.1: Schematic representation of the experiment. After a fixation cross of 500ms, a trial is presented for $1,000$ms. Then the participant is asked to click on the quantifier that better describes the scene.

Before starting, two instruction pages describing the task were displayed. Participants were asked to be as accurate and fast as possible. The task consisted of attending the visual scene and to select the quantifier which better answered the question: 'How many of the objects are animals?'. Particular focus was put on the fact that the quantifier had to be chosen *always* with respect to the set of *animals* (target set). This choice was aimed at diminishing the chance of errors merely due to wrong associations between the question and the target set. By fixing the set of animals as the target set, in fact, participants should be more focused on the quantification task *per se*. Importantly, the 9 quantifiers were never presented in any kind of order during the instructions.

After reading the instructions and having clarified any possible doubt with the experimenter, a training session was provided to get familiar with the task. The training session comprised of 5 trials which were not included in the 340 test stimuli. The procedure was the same as the test session (see Figure 4.1 for a schematic representation of the experiment): A white fixation cross was presented for 500ms in the center of a grey background screen; afterwards, a visual scene was displayed for $1,000$ms followed by the 9 quantifiers presented in a 3*3-cell grid centered in the middle of the screen. The

cells were well-spaced to prevent unwanted clicks, and highlighted by a darker shade of grey. Importantly, quantifiers were presented at each trial in a randomized position to avoid any familiarization effects. The task was to click on the chosen quantifier in the shortest possible time. After the response, a fixation cross appeared for 500ms followed by the next stimulus. After the first 5 training trials, a display was presented offering the possibility to train for extra 5 trials, different from the previous ones and also not included in the test set. Participants were asked to choose between training more or moving to the test session.

Before starting the test session, an instruction page was presented to specify that the experiment comprised of 10 blocks of 34 stimuli each. Subjects were reminded of the task. After left-clicking the mouse, participants started the first block of the experiment. At the end of each block, participants were allowed to take a self-paced pause. On each trial we recorded the chosen quantifier, its position on the grid, and the time taken to give the response. For each trial we also recorded a number of perceptual features describing the visual scene, such as the cardinality of animals and artifacts, their size (small, medium, large), and the ratio between animals and artifacts.

Responses by all participants were retained. 15 participants were in the age range 18-23, 11 in the range 24-29, 4 in the range 30-36. Seventeen requested and obtained university credits for their participation.

Materials

The visual scenes used in the experiment consisted in multiple colored pictures of animals (hence, targets) and artifacts (hence, non-targets) displayed on the top of a grey background (see Figure 4.2). Scenes differed on the total number of items displayed, that could vary from 3 to 20. Across scenes, the number of targets and non-targets varied such that different targets:non-targets *ratios* were equally represented. Crucially, each ratio corresponded to a fixed proportion of targets with respect to the total number of objects (i.e., targets+non-targets) in the scene. For example, ratio 1:3 corresponded to 25% of targets (see Figure 4.2). We used 17 ratios, each presented 20 times during the experiment, out of which 8 were 'positive' (targets $> 50\%$), 8 'negative' (targets $< 50\%$) and 1 'parity' (targets $= 50\%$). Because each ratio could be generated by different combinations of cardinalities (e.g., ratio 1:4 could result from the combination of 1 target and 4 non-targets, as well as 2 targets and 8 non-targets, etc.), for each ratio we

Figure 4.2: One visual scene used in the experiment, representing a targets:non-targets ratio of 1:3 (i.e. 25% of total items are targets).

presented all possible combinations of cardinalities. For any possible combination, a fixed number of visual scenes was built.

Visual scenes were generated with an inhouse Matlab script using the following pipeline: Two pictures, one depicting a target (e.g. an instance of a hedgehog) and one depicting a non-target (e.g. an istance of a basketball) were randomly chosen from a sample of the database by Kiani et al. (2007) including 100 instances of targets and 145 instances of non-targets. The sample was previously obtained by manually selecting pictures depicting whole items (not just parts) and whose color, orientation, and shape were not deceptive (for example, we discarded pictures depicting butterfly-shaped pasta as their target/non-target categorization could have been problematic). The target and the non-target pictures were randomly inserted by the script onto a 5*5-cell virtual grid. In order to inject some variability, each picture was randomly assigned to one orientation on the vertical axis (right or left) and one size (large, medium, small size, corresponding to approximately 5.3°, 3.4°, and 2.3° of visual angle). None of the scenes contained objects that were all the same size. As for the orientation, its effect is less measurable since it depends on the visual properties of the object (see, e.g., the different effects on the hedgehog and the basketball in Figure 4.2). However, this is not an issue since we are not interested in formally investigating the role of object orientation in the task. In total, 340 visual scenes were included in the experiment, together with additional 10 trials for training.

4.2.2 Abstract Task: Semantic Similarity Judgements

Thirty-three native-Italian participants (10 males, 22 females, 1 n.d.) completed this task. In an online survey powered by Google Forms, they were presented with pairs of quantifiers (e.g., 'almost none' and 'none'), and asked to rate their semantic similarity using a 7-point Likert-like scale, where 1 meant 'highly dissimilar' and 7 'highly similar'. Before starting the task, participants were presented with an instruction page where the terminology was briefly explained and the task exemplified. They were instructed that, in cases of difficulties in assessing the degree of semantic similarity between two quantifiers, they could adopt the strategy of mentally placing them into a default sentence (e.g., '*Few/Many* students have had high marks"), and judging the semantic similarity of the two resulting sentences. In order not to bias participants, only two trivial examples were provided in the instructions, namely 'all-none'=1, and 'some-some'=7. Moreover, given the constrained number of combinations, i.e. 9*9=81, no trial items were included. Each participant was asked to judge all 81 possible combinations in a randomized order of presentation. Each quantifier pair was rated twice by each participant, once in one order (i.e. 'all-none') and once in the opposite order (i.e. 'none-all'). To avoid any priming or repetition bias, we ensured that the two versions of the same pair never occurred in a row. Even though no time limits were set, participants were asked to provide their judgements as accurately as possible in the shortest possible time.

One participant's responses were discarded due to the repeated choice of the judgement 1 (i.e. 'highly dissimilar') in 55 out of 81 cases (68%). Responses by thirty-two participants (9 males, 22 females, 1 n.d.) were retained. 13 participants were in the age range 18-23, 14 in the range 24-29, 3 in the range 30-36, 2 in the range 37-42. Fifteen requested and obtained university credits for their participation.

4.3 Analysis and Results

4.3.1 Grounded Task: Quantifiers Used in Perception

All 30 participants successfully completed the experiment and provided each 340 responses. In total, 10,200 datapoints were collected. To ensure the quality of the responses, we removed those datapoints for which the reaction times exceeded the average of 2.5 SD. We did not perform any other filtering of the data. In total, 257 responses

quantifier	(a) resp	(b) % targ	(c) n targ	(d) n non-targ	(e) n total
none	604	0.01 (0.09)	0.13 (1.01)	11.35 (5.04)	11.48 (4.93)
almost none	861	0.19 (0.13)	1.69 (1.95)	7.81 (4.67)	9.45 (5.12)
few	1241	0.26 (0.13)	2.92 (1.58)	9.63 (4.96)	12.55 (5.40)
the smaller part	1135	0.32 (0.13)	3.79 (2.01)	8.99 (4.56)	12.78 (5.26)
some	1396	0.44 (0.13)	4.97 (2.30)	6.82 (3.66)	11.79 (4.79)
many	770	0.64 (0.14)	8.75 (3.76)	4.89 (2.66)	13.65 (4.53)
most	2110	0.69 (0.13)	8.82 (4.21)	3.90 (2.30)	12.72 (5.03)
almost all	1222	0.80 (0.12)	9.38 (5.08)	2.24 (2.00)	11.62 (5.68)
all	604	0.99 (0.09)	11.31 (5.04)	0.15 (1.13)	11.47 (4.99)

Table 4.1: Descriptive statistics. Columns are sorted with respect to ascending proportion of targets (b), which also corresponds to ascending cardinality of targets (c). Values in brackets refer to SD.

were discarded, equal to 2.52% of total. All statistical analyses were performed in the R environment on the resulting sample. For each quantifier, in Table 4.1 we report the following descriptive statistics: (a) The total number of responses assigned, (b) the average proportion of targets out of total number of items, (c) the average number of targets, (d) the average number of non-targets, (e) the average total number of items. Note that quantifiers are sorted according to ascending (b), which also corresponds to ascending (c).

As can be seen in the table, 'most' is the most used quantifier with $2,110$ responses. Low-magnitude quantifiers ('none', 'almost none', 'few', 'the smaller part') are used $3,841$ times (38.6%), high-magnitude quantifiers ('all', 'almost all', 'many', 'most') $4,706$ times (47.3%). As far as both the proportion and the cardinality of targets are concerned, the quantifiers turn out to be ordered on the following scale: 'none', 'almost none', 'few', 'the smaller part', 'some', 'many', 'most', 'almost all', 'all'. By looking at the proportions defining each quantifier, an almost perfect mirroring can be observed between 'none-all' ($\sim 0\%$-100%), 'almost none-almost all' ($\sim 20\%$-80%), 'the smaller part-most' ($\sim 30\%$-70%). Such a pattern can be better observed in Figure 4.3, which shows the frequency distribution of responses across proportions of targets. As can be seen, the quantifiers involved in these pairs have similar 'peaks' and distributions, though different frequencies.

In order to explore the role of cardinality of the target items in the scene, we separated the trials where the target items fell within the range of extremely well enumerable cardinalities (i.e. the so called 'subitizing' range, corresponding to scenes containing up to 3 animals) from those containing more than 3 items. The distribution of responses

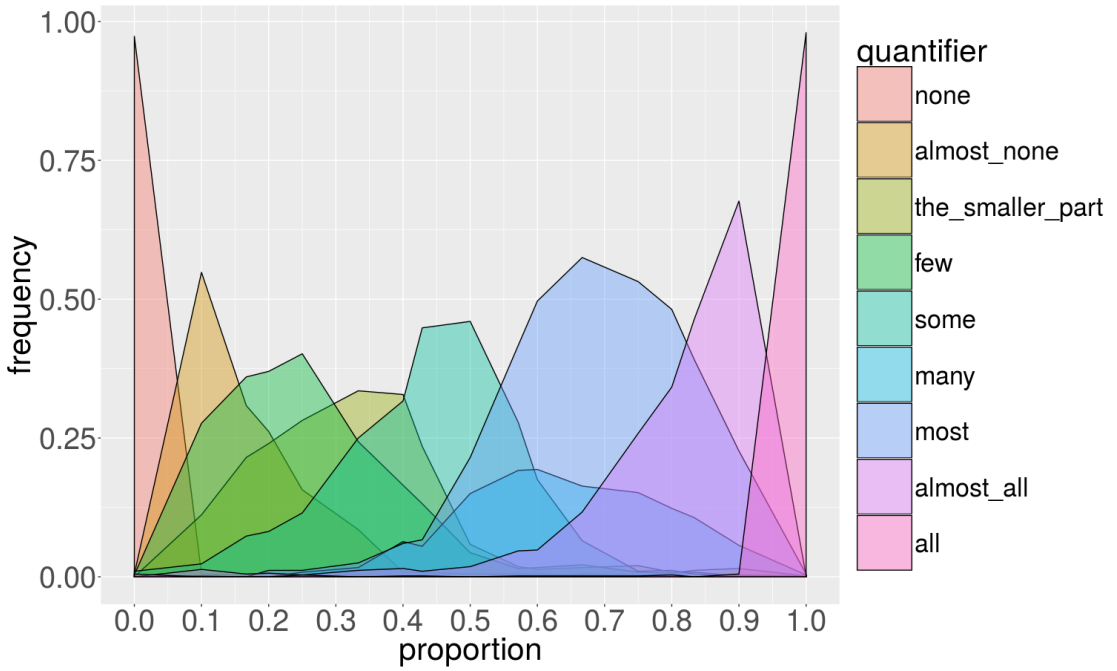

Figure 4.3: Density plot reporting the frequency distribution of responses for the 9 quantifiers (y-axis) against the proportion of targets in the scene (x-axis).

can be observed in Figure 4.4, which reports quantifiers frequency for scenes within the subitizing range (leftmost panel) and exceeding the subitizing range (rightmost panel). It should be noted that while in the former the whole range of quantifiers is used (though 'many' has an extremely low frequency), in the latter both 'none' and 'almost none' disappear, with an increasing use of quantifiers like 'most' and 'many'. It is worth mentioning that the choice of setting the subitizing threshold to 3 was aimed at making our results directly comparable to those reported by Coventry et al. (2005, 2010), who experimented with such setting.

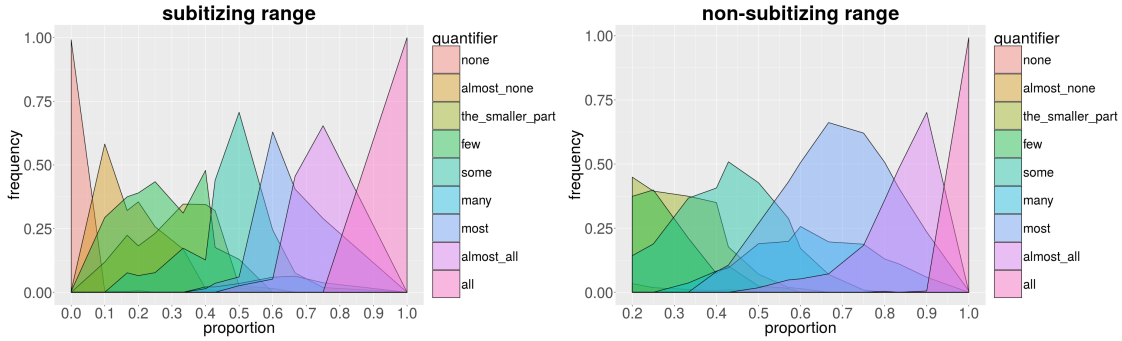

Figure 4.4: Density plots reporting frequency distribution of responses against proportion of targets for scenes whose number of targets is within the subitizing range (left) and exceeding it (right).

quantifier	AIC scores					
	(a) % targ	(b) n targ	(c) n non-targ	(d) sub/non-sub	(e) targ size	(f) non-targ size
none	**613.03**	756.31	3474.44	3113.78	–	3913.39
almost none	4353.51	**4230.42**	5591.74	4292.00	4686.18	5633.89
few	**5492.00**	6015.22	6486.62	6428.88	6987.15	7018.23
the smaller part	**5241.33**	5938.05	6109.82	6605.17	6540.34	6451.78
some	**5811.28**	6864.85	7342.64	7792.71	7608.18	7461.32
many	**4273.67**	4520.02	4834.66	4600.70	4909.47	5062.78
most	**6755.09**	8402.49	8741.28	8748.20	9330.23	9604.31
almost all	**5079.70**	6355.7	5692.78	6545.65	6762.34	6075.14
all	**482.37**	3323.29	732.50	3672.75	3568.47	–

Table 4.2: AIC scores for each of the models. **Bold** values (lowest) correspond to best models. Empty cells indicate cases for which the number of datapoints was too low to perform statistical analyses.

To more formally investigate which factors contribute in determining quantifiers meaning in grounded contexts, we performed statistical analyses on the collected data. Because our variables of interest are naturally highly correlated (crucially, proportion of targets and cardinality of both targets and non-targets), it was not possible to disentangle between the relative contribution of the two (or more) factors within the same logistic regression model aimed at predicting the choice of a given quantifier against all the others. We thus employed the 'one model, one predictor' strategy, according to which a number of separate models including only one predictor of interest (along with random factors) was performed for each quantifier. This way, the predictive power of each variable could be tested separately, and we could further evaluate the quality of each model relative to all other candidate models. Model selection was performed using Akaike Information Criterion (AIC), a measure based on information theory which allowed us to select the best model for a given set of data (Akaike, 1973). In particular, the lowest the AIC, the lowest the information loss compared with the 'true' model, namely the process that generated the data. We considered both raw AIC scores and AIC weights (Wagenmakers and Farrell, 2004).

Seven variables were used as predictors: (a) proportion of targets, (b) cardinality of targets, (c) cardinality of non-targets, (d) subitizing/non-subitizing range (dichotomic dummy variable), (e) average size of targets, (f) average size of non-targets[2]. In total, 52 models were tested. All models were mixed-effect logistic regressions (Baayen et al., 2008) with one fixed predictor (see above) and 3 random factorial variables, namely (1) participant, (2) experimental block, and (3) position of the quantifier in the response

[2]The average size of the targets was obtained by dividing their weigthed sum (each large target was multiplied by 1, medium ones by 0.75, small ones by 0.5) by the number of targets in the scene. The same criteria and procedure were used for non-targets. For intuitive reasons, scenes containing either 0 targets or 0 non-targets were excluded from this analysis.

quantifier	predictor	Estimate	z-value	p-value
none	proportion	424.78	19.36	.0001
almost none	n targets	82.86	9.66	.0001
few	proportion	-215.02	-22.41	.0001
the smaller part	proportion	-235.73	-25.98	.0001
some	proportion	-279.16	-35.69	.0001
many	proportion	-210.73	-6.31	.0001
most	proportion	-288.99	-29.79	.0001
almost all	proportion	-147.51	-13.67	.0001
all	proportion	462.95	18.66	.0001

Table 4.3: Estimate, z-value and p-value of the quadratic term for each of the best models.

grid. By including these random variables in the models, we ensured that significant effects were estimated for the whole set and not just for a sample of stimuli. That is, we ensured that the effects were not due to the variability among participants, blocks of stimuli, position of the quantifier word in the response grid. To better fit the data, all the models except (d) treated the predictor as a second-order polynomial variable. Logit models were performed using the function `lmer()` implemented in the package `lme4`.

To compare different models, raw AIC scores and AIC weights were used. Since, in all cases, AIC weights for the lowest-AIC model approximated 1 (i.e. the total weight of the models considered), Table 4.2 reports only AIC scores for all models. As can be seen, for 8 quantifiers out of 9, the best model (i.e. the one with the lowest information loss) turned out to be the one using proportion of targets (% targ). In one case, namely 'almost none', the best model was instead the one using cardinality of targets (n targ) as the predictor. The models based on all other predictors (cardinality of non-targets, subitizing/non-subitizing range, and either targets or non-targets average size) never emerged as the best ones for any quantifier.

It is worth stressing that AIC scores do not say anything about the absolute quality of the model, i.e. the testing of the null hypothesis. Once established the best models based on the AIC score, we could inspect them using the traditional null-hypothesis testing. For all best models, both the linear and the quadratic term of the polynomial variable turned out to be highly significant ($p<.0001$), meaning that each quantifier can be reliably predicted against the other quantifiers by means of the polynomial form of the given predictor. In Table 4.3, we report Estimate, z-value and p-value of the quadratic term (2nd order term) for each of the selected models.

Based on the well-reported effects due to subitizing, we analyzed separately the dat-

quantifier	predictor	AIC score	Estimate	z-value	p-value
none	n targets	328.2	158.15	11.41	.0001
almost none	n targets	2572.3	-136.47	-20.35	.0001
few	n targets	3541.3	-69.75	-13.84	.0001
the smaller part	proportion	2662.3	-110.61	-13.40	.0001
some	proportion	2057.6	-88.07	-12.69	.0001
many	proportion	256.9	-195.17	-4.38	.0001
most	proportion	733.8	-57.04	-4.74	.0001
almost all	proportion	629.2	8.97	13.81	.0001
all	proportion	57.8	247.04	2.72	.0064

Table 4.4: AIC score, estimate, z-value and p-value of the quadratic term (linear term for 'almost all') for each of the best models in the subitizing range.

apoints within the subitizing range, i.e. cardinality of targets up to 3 included. The intuition behind that is that when the target items are very easily enumerable (in the subitizing range), their absolute number might be a better predictor of the quantifier used by subjects than the proportion. To test this hypothesis, the same kind of analysis as above was performed on the split data ($3,771$ datapoints). For all quantifiers except 'almost all', the best models turned out to be the polynomial ones, whereas for 'almost all' the best model was the linear one. Table 4.4 reports AIC score, Estimate, z-value, and p-value of the quadratic term (linear term for 'almost all') for the best models in the subitizing range. As can be noticed, in the subitizing range the low-magnitude quantifiers 'none', 'almost none', and 'few' are better modeled by the absolute number of animals rather than by the proportion of targets. This suggests that the choice of these quantifiers in this range relies more on evaluating the set of targets on its own than comparing it against the set of non-targets.

Finally, we investigated whether the frequency of use of quantifiers in language is reflected in the distribution of responses observed in the experiment. The rationale is that, when choosing a quantifier from the various options, participants might be biased towards the most frequent words, irrespectively of the perceptual features of the visual stimulus. We extracted raw frequency values for each of the 9 Italian quantifiers at the lemma level from CORIS (Favretti et al., 2002) and we computed the Pearson's correlation (r) with the quantifier frequencies observed in the experiment. All the values were previously log-transformed. The correlation turned out to be very weak and not significant in the full dataset ($r(7) = -0.25$, p=0.52), in the subitizing range subset ($r(7) = -0.41$, p=0.27), and in the non-subitizing range subset ($r(7) = -0.04$, p=0.92). That is, participants are not affected by the linguistic frequency of the quantifier when picking it up from the list.

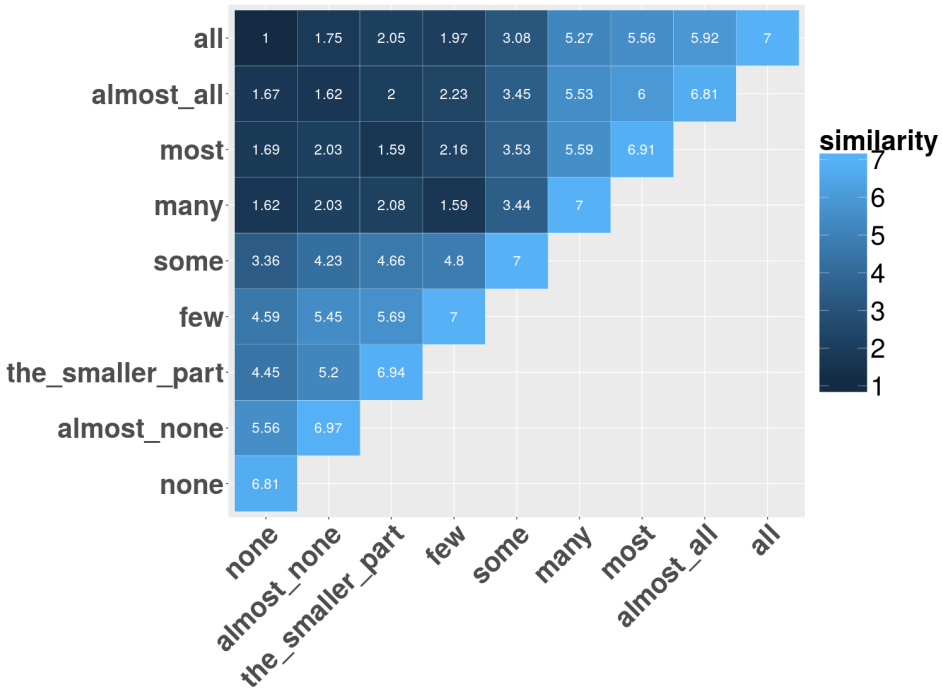

Figure 4.5: Heatmap reporting the average semantic similarity between quantifiers pairs. The lighter the blue, the more similar the pair.

4.3.2 Abstract Task: Semantic Similarity Judgements

The pattern of estimated similarities across quantifiers indicated that quantifiers are represented on an ordered but highly non-linear scale. A visualization of that can be observed in Figure 4.5, where a heatmap depicting the average semantic similarity between quantifier pairs is reported. Three interesting features can be appreciated: First, the ordered aspect of the internal scale can be seen by observing a roughly graded decrease in similarity as pairs move away from the diagonal. This indicates a rough 'distance effect', indexing an internal ordered scale. This distance effect appears stronger for low-magnitude quantifiers compared to high-magnitude ones. This can be appreciated qualitatively by inspecting Figure 4.6, where the bell functions peaking around the low-magnitude quantifiers ('few', 'the smaller part', 'almost none', 'none') appear sharper compared to those characterizing the high-magnitude quantifiers ('many', 'most', 'almost all', 'all').

Second, it appears that this graded effect is mostly confined in quantifiers that refer to similar magnitudes, and disappears for very distant quantifiers. Indeed, there seems to be a clear-cut distinction between low-magnitude and high-magnitude quantifiers. In this respect, 'some' turns out to be a 'hinge' between low- and high-magnitude quanti-

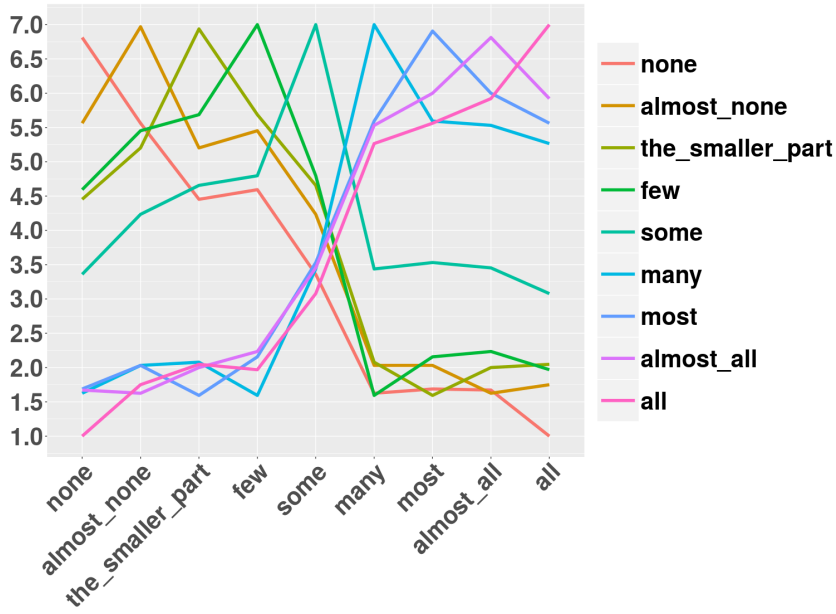

Figure 4.6: Line plot reporting the average semantic similarity between quantifiers.

fiers. It should be observed that none of the items are judged to be as extremely similar/dissimilar to it, with the lowest average similarity being equal to 3.08 ('all-some'), and the highest being equal to 4.8 ('few-some'). Though halfway between low- and high-magnitude quantifiers, however, 'some' results to be closer to the former than to the latter group. Finally, we observe a rather small but systematic linguistic 'antinomy effect': For each quantifier (with the exception of 'some') the most dissimilar item is represented not by the extreme on the other side of the scale, but by its linguistic *antonym*: The lowest similarity ratings are those among 'none-all', 'almost none-almost all', 'the smaller part-most', 'few-many' (this can be appreciated by the presence of an orthogonal diagonal to the main one in the similarity matrix).

To pool together the pattern of judgements and reconstruct the shape of the internal representation, we performed a metric Multi-Dimensional Scaling (MDS) analysis. Such technique is commonly used to visualize the degree of similarity between objects by placing them on a N-dimensional space where distances between them are preserved. Figure 4.7 shows the results of the analysis when taking into account two dimensions. By performing a goodness-of-fit analysis, it turned out that the first dimension only, depicted along the x-axis in the plot, accounts for 98.66% of the variance of the original data (R^2=0.9866, $F(1, 34)$=2496.81, p<.0001). As shown in Figure 4.7, such dimension clearly separates low-magnitude quantifiers from high-magnitude quantifiers, with 'some' somehow in between, though closer to the former block. By including the second dimension, the variance accounted for by the model increases to 98.80% (R^2=0.9880,

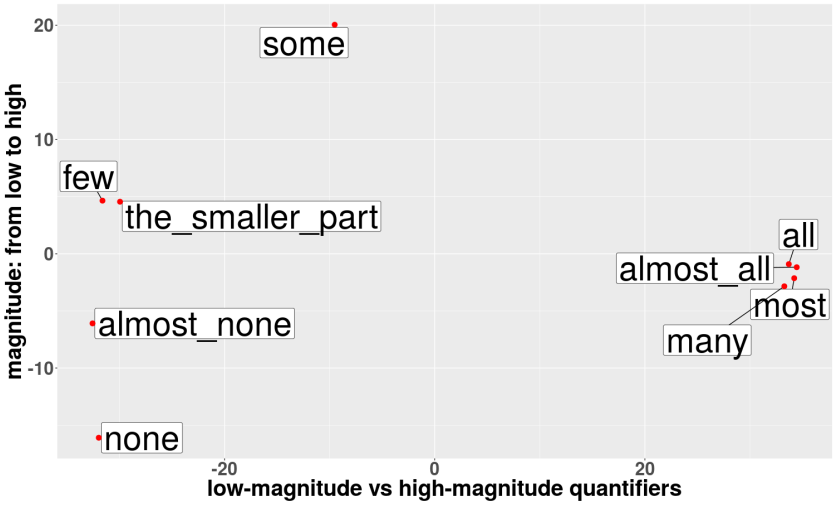

Figure 4.7: Plot reporting the absolute distance of quantifiers as resulting from a two-dimension metric MDS analysis.

$F(1, 34)$=2803.18, p$<$.0001), which is almost a perfect fit. Such dimension neatly represents magnitude: From low to high, along the y-axis. This analysis further confirms that low-magnitude quantifiers are better separated among them, indicating that they correspond to sharper representations. This allows their ordering on a scale to emerge very clearly, with 'none' being followed by 'almost none' that, in turn, is followed by 'few' and 'the smaller part' (which are not well separated among each other), and eventually by 'some'. On the contrary, high-magnitude quantifiers, while still being ordered along a scale, are extremely close to each other, indicating that their representations overlap greatly.

4.4 Discussion

4.4.1 Visually-Grounded Representation

In this chapter, I explored the use of quantifiers in both their visually-grounded and abstract representation. By asking participants to choose the quantifier that best represented the quantity of animals in a number of visual scenes, Experiment 1 was aimed at investigating the factors which contribute in determining the visually-grounded representation of such linguistic expressions. We showed that the proportion of targets is the best predictor for 8 quantifiers out of 9, with 'almost none' being better described by the cardinality of the target set. When zooming into the subitizing range, with cardinality

of animals up to 3, the absolute number of targets turned out to be the best predictor for 'none' and 'few' besides 'almost none', thus suggesting that when the information about precise number is available it becomes crucial for discriminating among low-magnitude quantifiers.

These findings are generally in line with previous studies investigating the appropriateness of quantifiers evaluated against visual scenes (Coventry et al., 2005, 2010). Using a different experimental design (evaluating the appropriateness of a number of quantifier-embedding sentences against a given visual scene), a different set of quantifiers ('a few', 'few', 'several', 'many', 'lots of'), and without constraining the exposure time to the scene, these works showed that the number of both targets and non-targets is predictive of the quantifier appropriateness. With cardinality of targets equal to 3 (their subitizing case), however, the use of quantifiers was no longer affected by the cardinality of the non-target objects. An exception was represented by 'few', which was affected by both (Coventry et al., 2010). On the one hand, our finding that proportion is overall the best predictor is not in contradiction with the effect of both number of targets and number of non-targets. Rather, we believe ours to be just a better measure to assess the contribution of both sets in determining quantifiers' use. On the other hand, the results we obtained in the subitizing range reinforce and better prove the increasingly important role of precise number in discriminating between low-magnitude quantifiers. In our study, interestingly, the only low-magnitude quantifier whose interpretation turned out to be best predicted by the proportion of targets also in the subitizing range was 'the smaller part', whose reading is intuitively more proportional compared to the others. Finally, it is worth stressing that our 340 visual scenes were balanced with respect to ratios, whereas the 36 used by Coventry et al. (2005, 2010) were balanced for target cardinality. Moreover, in the present work each ratio was represented by all possible combinations of cardinalities, whereas Coventry and colleagues experimented with ratios that were mostly depicted by just one combination. Finally, our subitizing range included four cardinalities, namely 0, 1, 2, and 3 – not just the number 3.

As far as the effect of object size is concerned, we found this factor not to be among the most predictive ones. This finding is in partial contrast with the results reported by Newstead and Coventry (2000), who showed a role of size in the task of evaluating quantifiers over scenes depicting dots placed in a container. In that study, both the dots and the container size were found to play a role: Low-magnitude quantifiers were found to be more appropriate when the dots were small and when the container was big. In our task, we solely investigated the size of the items, and found that this parameter was not

among the best predictors of quantifiers' use. This difference might be due to the different experimental settings: First, our scenes contain both target and non-target objects – not only targets. Second, we vary the size of the objects in a way that there are no scenes depicting, e.g., only small or large objects. Third, we employ a larger set of quantifiers, thus participants have more alternatives compared to the previous study. Moreover, contrary to us, Newstead and Coventry (2000) allowed subjects to explore the scenes for an infinite time, such that they might have used a different visuo-spatial strategy (namely, exact counting), and that might have influenced the enumeration process. Though we showed that object size is not among the most predictive factors of quantifier use, in our setting we could not rule out the possibility that participants relied on information regarding the area occupied by objects. To address this issue, the total number of pixels occupied in each scene by target and non-target objects should be controlled.

4.4.2 Abstract Representation

By asking participants to rate the degree of semantic similarity between quantifier pairs, Experiment 2 was aimed at testing whether these expressions are mentally ordered and, if so, which are the features of the resulting scale. We showed that, even without relying on any quantitative or contextual information, quantifiers do lie on an ordered scale, as resulting from a Multi-Dimensional Scaling Analysis (Kruskal and Wish, 1978). In particular, low-magnitude quantifiers ('none', 'almost none', 'few', 'the smaller part') turned out to be perceived as being fairly distant from each other, thus suggesting that their abstract semantic representation is well defined and nicely ordered on a scale. In contrast, high-magnitude quantifiers ('many', 'most', 'almost all', 'all') turned out to greatly overlap, though always along an ordered scale. Overall, these results suggest that the mental representation of quantifiers is ordered and highly non-linear, with small quantifiers better represented compared to large ones. This is highly reminiscent to the well-reported logarithmic scale inferred both from comparative judgements across numerical symbols and from the use of numerical symbols in perceptual quantification (Nieder and Miller, 2003; Dehaene, 2003; Dehaene et al., 2008).

It is worth stressing that, in doing this task, neither quantitative (numbers, proportions, etc.) nor explicit contextual (semantic) information was provided. That is, quantifiers were judged in isolation, solely on the basis of their bare semantic similarity, while in Holyoak and Glass (1978) participants were asked to rate dissimilarities between pair of sentences embedding different quantifiers. Another interesting finding was the ten-

dency to assign the lowest rating (i.e. lowest semantic similarity) to the direct antonym. For example, the most dissimilar word from 'few' was 'many', and not 'none'. While straightforward for the pair 'none-all', which also represent the two extreme endpoints of the scale, this finding is in principle not trivial in all the other cases. This finding falls off the prediction that quantifiers should solely lie on a quantitative scale (e.g. numerical or proportional) and suggests that, when asked to judge the semantic similarity of a word pair, speakers also take into account lexico-semantic features, such as information regarding the direct antonym (Miller and Fellbaum, 1991), as also reported by Hill et al. (2016b).

4.4.3 Mental Order

Finally, it should be mentioned that previous work has investigated the scalar nature of quantifiers from very different perspectives. With a set of 5 quantifiers and a task which was similar to ours, for example, Holyoak and Glass (1978) claimed that quantifiers can be described in terms of an unidimensional scale, essentially representing analog quantities. The authors, however, did not overtly exclude that information regarding other non-quantitative related semantic features might still be included in the memory representation of quantifiers. In contrast with the unidimensional nature of the quantifier scale was Routh (1994), whose results on a freesort task with 20 quantifiers suggested that several other components are in place beyond the quantity scale. Another study (Montalto et al., 2010) also adopted a similar paradigm where a number of Italian quantifiers (yet different from the list of quantifiers investigated in our study) were compared to each other on a magnitude scale: Given pairs of quantifiers subjects had to indicate if and which of the two indicated the largest amount. Differently from our experiment, however, subjects were given the possibility to indicate that the two quantifiers did not differ in the implicated amount. Results suggested that subjects lump quantifiers in two blocks, one comprising low and the other high-magnitude ones, with no hint of a continuous scale. However, there is the serious possibility that these results do not directly reflect the true mental scale but rather the degree of certainty, such that when prompted with uncertain decisions subjects indicated an absence of differentiation.

4.4.4 Impact of our Results on Foundational Theories

As for the theoretical implications of our work, our results provide evidence in support of some well-established assumptions on quantifiers. First, our findings show that quantifiers neither correspond to an exact number of entities nor to a fixed proportion (see section 2.3). This can be taken as an evidence in favor of their non-referential status, even in the new light shed by the integration of perception and quantifiers.

Second, our results do not shed new light on the proposal that 'few' and 'many' are not-extensional since, in our experiments, contextual factors were deliberately avoided. However, it is worth noticing that in Experiment 1 the meaning of 'few' is found to be ambiguous: It mostly depends on the number of targets in the subitizing range, on the proportion of targets in the whole data. This might be seen as an effect of a perceptual 'contextual' factor: 'Few' is more dependent on the perceptual context than are other quantifiers. However, the same effect was not observed for 'many'.

Third, our results are consistent with the literature on scalar implicatures (Grice, 1975) in the pragmatic use of quantifiers. In particular, both the ordering of quantifiers (from low- to high- magnitudes) and their narrow range of application observed in Experiment 1 suggest that, to some extent, speakers interpret such expressions as having an upper boundary which excludes the use of more informative options when these options are not true or uncertain (Horn, 1984). That is, participants choose the most informative quantifier 'all' (and not e.g. 'some', which would be logically true) when they are certain about its applicability (see also Degen and Tanenhaus (2015)). Similar implications can be drawn from Experiment 2, where the characteristics of the abstract representation might indicate that speakers have an internal representation of quantifier informative strength. Based on our findings, one possibility is that scalar implicatures are stronger for low-magnitude quantifiers (which turn out to be extremely well-defined and distinct from each other) than for high-magnitude ones (which are perceived to be very similar). We leave this issue for future research and refer the reader to Oaksford et al. (2002) for interesting results on the use of quantifiers as referring to different ranges of numerosities and their effect on informativeness.

Fourth, the results of Experiment 2 are in line with the position that the meaning of quantifiers is not only about amounts, numbers, or proportions. Indeed, similarity judgments provided by participants turned out to be dependent on lexico-semantic factors (e.g. antonymy) besides magnitude. This evidence is also in line with previous findings

showing an interplay between numerical and semantic information in the comprehension of quantifiers (see section 2.4).

Fifth, our results overall suggest that the meanings of quantifiers are at least partially tied to the representation of quantities. Though this is probably not enough to devise a general semantics for such expressions, we believe quantitative aspects to constitute the basis of quantifier meanings.

4.4.5 Final remarks

In sum, our results indicate that, in grounded contexts, quantifiers primarily represent *proportions* and not absolute cardinalities. They also show that quantifiers are mentally represented on a quantity scale which is well ordered and highly non-linear, bearing interesting similarities to the mental representation of both numerical quantities and continuous magnitudes. While our results cannot endorse one possibility over the other, they firmly support the view that quantifiers are mentally represented in a way that partially reflects the way we perceive quantities through our senses.

In the next chapter, I build on the evidence that numbers and quantifiers have different quantitative representations, and test whether two computational mechanisms are required to learn them from visual scenes.

Chapter 5

Quantifiers vs Cardinals: Two Computational Mechanisms

In this chapter, I focus on the computational operations underlying the use of cardinals (*one*, *two*, *three*, and *four*) and quantifiers (*no*, *few*, *most*, and *all*) when referring to objects that are grounded in visual scenes. Inspired by the evidence that, in humans, the two processes imply fairly different cognitive (see Chapter 4) and neural mechanisms, I propose that distinct models are required for learning the meaning of such expressions from images containing multiple objects. I show that a model capitalizing on a 'fuzzy' measure of similarity is effective for learning quantifiers, whereas the learning of cardinals is better accomplished when 'exact' information is provided.

5.1 Introduction

In everyday life, people can refer to quantities by using either cardinals (e.g. *one*, *two*, *three*) or natural language quantifiers (e.g. *few*, *most*, *all*). Although they share a number of syntactic, semantic and pragmatic properties (Hurewitz et al., 2006), and they are both learned in a fairly stable order of acquisition across languages (Wynn, 1992; Katsos et al., 2016), these quantity expressions underlie fairly different cognitive and neural mechanisms. First, they are handled differently by the language acquisition system, with children recognizing their disparate characteristics since early development, even before becoming 'full-counters' (Hurewitz et al., 2006; Sarnecka and Gelman, 2004; Barner et al., 2009). Second, while the neural processing of cardinals relies on the brain

Figure 5.1: How many pets are *dogs*? Three/Most. Image credits: cvalleyvet.com

region devoted to the representation of quantities, quantifiers rather elicit regions for general semantic processing (Wei et al., 2014). Intuitively, cardinals and quantifiers refer to quantities in a different way, with the former representing a mapping between a word and the exact cardinality of a set, the latter expressing a 'fuzzy' numerical concept denoting set relations or proportions of sets (Barner et al., 2009). As a consequence, speakers can reliably answer questions involving quantifiers even in contexts that preclude counting (Pietroski et al., 2009), as well as children lacking exact cardinality concepts can understand and appropriately use quantifiers in grounded contexts (Halberda et al., 2008; Barner et al., 2009). That is, knowledge about (large) precise numbers is neither necessary nor sufficient for learning the meaning of quantifiers.

Inspired by this evidence, the present study proposes two computational models for learning the meaning of cardinals and quantifiers from visual scenes. Our hypothesis is that learning cardinals requires taking into account the number of instances of the target object in the scene (e.g. number of *dogs* in Figure 5.1). Learning quantifiers, instead, would be better accomplished by a model capitalizing on a measure evaluating the 'fuzzy' amount of target objects in the scene (e.g. proportion of 'dogness' in Figure 5.1). In particular, we focus on those cases where both quantification strategies might be used, namely scenes containing target (*dogs*) and distractor objects (*cats*). Our approach is thus different from salient objects detection, where the distinction targets/distractors is missing (Borji et al., 2015; Zhang et al., 2015b, 2016). With respect to cardinals, our approach is similar to Seguí et al. (2015), who propose a model for counting people in natural scenes, and to more recent work aimed at counting either everyday objects in natural images (Chattopadhyay et al., 2017) or geometrical objects with attributes in synthetic scenes (Johnson et al., 2017). With respect to quantifiers,

our approach is similar to Sorodoc et al. (2016), who use quantifiers *no*, *some*, and *all* to quantify over sets of colored dots. Differently from ours, however, all these works tackle the issue as either a classification problem or a Visual Question Answering task, with less focus on learning the meaning representation of each cardinal/quantifier. To our knowledge, this is the first attempt to jointly investigate both mechanisms and to obtain the meaning representaton of each cardinal/quantifier as resulting from a language-to-vision mapping.

Based on their geometric intepretation, we propose to use **cosine** and **dot product** similarity between the target object and the scene as our functions for modeling quantifiers and cardinals, respectively. The former, ranging from -1 to 1, evaluates the similarity between two vectors with respect to their orientation and irrespectively of their magnitudes. That is, the more two vectors are *overall* similar, the closer they are. Ideally, cosine similarity between an image depicting a *dog* and a scene containing either 3 or 10 *dogs* without distractors (hence, 'all') should be equal to 1. Therefore, it would indicate that the the *proportion* of 'dogness' in the scene is highest. Dot product, on the other hand, is defined as the product of the cosine between two vectors and their Euclidean magnitudes. By taking into account the magnitudes, this measure ideally encodes information regarding the number of times a target object is repeated in the scene. In the above-mentioned example, indeed, dot product would be 3 and 10, respectively. In this simplified setting, thus, it would be equal to the *number* of 'dogs'.

Furthermore, we propose that the 'objective' meaning of each cardinal/quantifier can be learned by means of a cross-modal mapping (see Figure 5.4) between the linguistic representation of the target object and its quantity (either exact or fuzzy) in a visual scene. To test our hypotheses, we carry out a proof-of-concept on the synthetic datasets we describe in section 5.2. First, we explore our visual data by means of the two proposed similarity measures (section 5.3.1). Second, we learn the meaning representations of cardinals and quantifiers and evaluate them in the task of retrieving unseen combinations of targets/distractors (section 5.3.2). As hypothesized, the two quantification mechanisms turn out to be better accounted for by models capitalizing on the expected similarity measures.

5.2 Data

In order to test our hypothesis, we need a dataset of visual scenes which crucially include multiple objects. Moreover, some objects in the scene should be repeated, so that we might say, for instance, that out of 5 objects 'three'/'most' are *dogs*. Although a large number of image datasets are currently available (see Lin et al. (2014) among many others), no one fully satisfies these requirements. Typically, images depict one salient object and even when multiple salient objects are present, only a handful of cases contain both targets and distractors (Zhang et al., 2015b, 2016).

To bypass these issues, in the present work we experiment with synthetic visual scenes (hence, scenarios) that are made up by at most 9 images each representing one object. The choice of using a 'patchwork' of object-depicting images is motivated by the need of representing a reasonably large variability (e.g. 'few' refer to scenes containing 2 target objects out of 7 as well as 1/5, 4/9, etc.). This way, we avoid matching a quantifier always with the same number of target objects (except *no*, that is always represented by 0 targets), and allow cardinals to be represented by scenes with different numbers of distractors. At the same time, we get rid of any issues related to object localization.

We experiment with quantifiers (hence, Qs) *no*, *few*, *most*, and *all*, which we defined *a priori* by ratios 0%, 1-49%, 51-99% and 100%, respectively. Consistently with our goals, this arguably simplified setting does neither take into account pragmatic uses of Qs (i.e. we treat them as lying on an ordered scale) nor reflect possible overlappings. For these reasons, we avoid using quantifiers as *some* whose meaning overlaps with the meaning of many others. As far as cardinals (hence, Cs) are concerned, we experiment with scenarios in which the cardinality of the targets ranges from 1 to 4. Cs up to 4 are acquired by children incrementally at subsequent stages of their development, with higher numbers being learned upon this knowledge with the ability of counting (Barner et al., 2009). Also, Cs ranging from 1 to 3-4 are widely known to exhibit some peculiar properties (i.e. their exact number can be immediately and effortlessly grasped) due to which they are usually referred to as 'subitizing' range (Piazza et al., 2011; Railo et al., 2016).

5.2.1 Building the Scenarios

We use images from ImageNet (Deng et al., 2009). Starting from the full list of 203 concepts and corresponding images extracted by Cassani (2014), we discarded those concepts whose corresponding word had low/null frequency in the large corpus used in (Baroni et al., 2014). To get rid of issues related to concept identification, we used a single representation for each of the 188 selected concepts. Technically, we computed a centroid vector by averaging the 4096-dimension visual features of the corresponding images, which were extracted from the *fc7* of a CNN (Simonyan and Zisserman, 2014). We used the VGG-19 model pretrained on the ImageNet ILSVRC data (Russakovsky et al., 2015) implemented in the MatConvNet toolbox (Vedaldi and Lenc, 2015). Centroid vectors were reduced to 100-d via PCA and further normalized to length 1 before being used to build the scenarios. When building the scenarios, we put the constraint that distractors have to be different from each other. Moreover, only distractors whose visual cosine similarity with respect to the target is lower than the average are selected. For each scenario, target and distractor vectors are summed together. As a result, each scenario is represented by a 100-d vector.

We also experimented with scenarios where vectors are concatenated to obtain a 900-d vector (empty 'cells' are filled with 0s vectors) and further reduced to 100-d via PCA. Since the pattern of results in the only-vision evaluation (see section 5.3.1) turned out to be similar to the results obtained in the 'summed' setting, we will only focus on the 'summed' setting.

5.2.2 Datasets

We built one dataset for Cs and one for Qs, each containing 4512 scenarios.[1] We then split each of the two in one 3008-datapoint Training Dataset (Train) for training and validation and one 1504-datapoint Testing Dataset (Test) for testing. The two datasets were split according to their 'combinations', that is the mixture of targets and distractors in the scenario. As reported in Table 5.1, we kept 4 different combinations for each C/Q in Train and 2 in Test. Note that the numerator refers to the number of targets, the denominator to the total number of objects. The number of distractors is thus given by the difference between the two values. To illustrate, in Train-q 'few' is represented

[1] A visual representation of our scenarios is provided in the rightmost side of Figure 5.4, while Figure 5.1 is only intended to provide a more intuitive overview of the task.

Train-q				**Train-c**			
no	few	most	all	one	two	three	four
0/1	1/6	2/3	1/1	1/1	2/2	3/3	4/4
0/2	2/5	3/4	2/2	1/3	2/3	3/4	4/5
0/3	2/7	3/5	3/3	1/4	2/5	3/5	4/6
0/4	3/8	4/5	4/4	1/6	2/7	3/8	4/7
Test-q				**Test-c**			
no	few	most	all	one	two	three	four
0/5	1/7	4/6	5/5	1/2	2/4	3/7	4/8
0/8	4/9	6/8	9/9	1/7	2/9	3/9	4/9

Table 5.1: Combinations in Train and Test.

by scenarios 1/6, 2/5, 2/7, and 3/8, whereas in Test-q 'few' is represented by scenarios 1/7 and 4/9. The initial 4512 scenarios have been obtained by building a total of 24 different scenarios (6 combinations * 4 C/Q classes) for each of the 188 objects. A particular effort has been paid in making the datasets as balanced as possible. When designing the combinations for 'few' and 'most', for example, we controlled for the proportion of targets in the scene, in order to avoid making one of the two easier to learn. Also, combinations were thought to avoid biasing cardinals toward fixed proportions of targets/distractors.

5.3 Experiments

5.3.1 Only-Vision Evaluation

As a first step, we carry out a preliminary evaluation aimed at exploring our visual data. If our intuition about the information encoded by the two similarity measures is correct (see section 5.1), we should observe that cosine is more effective than dot product in distinguishing between different Qs, while the latter should be better than cosine for Cs. Moreover, Qs/Cs should lie on an ordered scale. To test our hypothesis, we compute cosine distances (i.e. $1-$cosine, to avoid negative values) and dot product similarity for each target-scenario pair in both Train and Test (e.g. *dog* vs 2/5 *dogs*). Figure 5.2 reports the distribution of Qs with respect to cosine (left) and Cs with respect to dot product (right) in Train. As can be seen from the boxplots, both Qs and Cs are ordered on a scale. In particular, cosine distance is highest in *no* scenarios (where the target is not present), lowest in *all* scenarios. For Cs, dot product is highest in *four* scenarios,

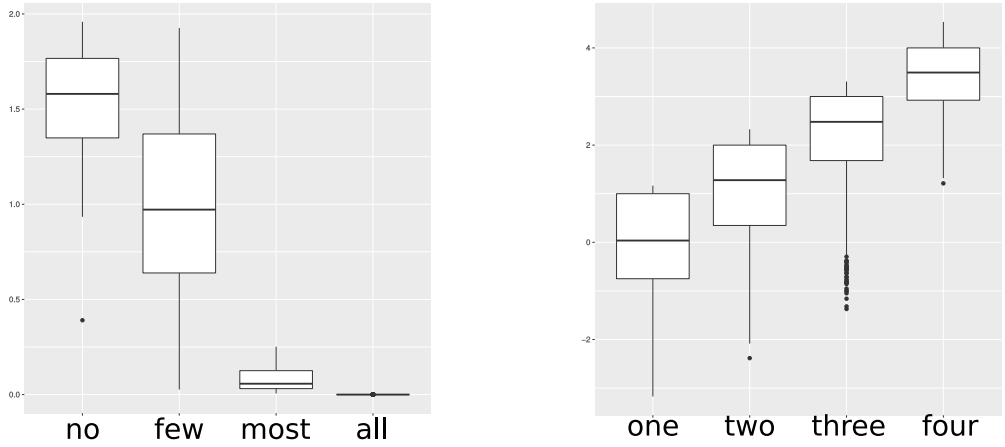

Figure 5.2: Left: Qs against cosine distance. Right: Cs against dot product.

lowest in *one* scenarios.

Our intuition is further confirmed by the results of a radial-kernel SVM classifier fed with either cosine or dot product similarities as predictors.[2] Qs are better predicted by cosine than dot product (78.6% vs 63.8%), whereas dot product is a better predictor of Cs than cosine (68.7% vs 44.7%). As shown in Figure 5.3, the ordered scale is indeed represented to a much lesser extent when Qs are plotted against dot product (left) and Cs against cosine (right). A similar pattern of SVM results and similar plots emerged when experimenting with Test.

5.3.2 Cross-Modal Mapping

Our core proposal is that the meaning of each C/Q can be learned by means of a cross-modal mapping between the linguistic representation of the target object (e.g. *dog*, *mug*, etc.) and a number of scenarios representing the target object in a given C/Q setting (e.g. 'two'/'few' *dogs*). In our approach, each word (e.g. *dog*) is represented by a 400-d embedding built with the `CBOW` architecture of `word2vec` (Mikolov et al., 2013) and the best-predictive parameters of Baroni et al. (2014) on a 2.8B tokens corpus. The original 400-d vectors are further reduced to 100-d via PCA before being fed into the model.

Figure 5.4 reports a single learning event of our proposed model. Each C/Q (e.g. *two*, *few*) is learned as a separate function that maps each of the 188 words representing our

[2]We experimented with linear, polynomial, and radial kernels. We only report results obtained with default radial kernel, that turned out to be the overall best model.

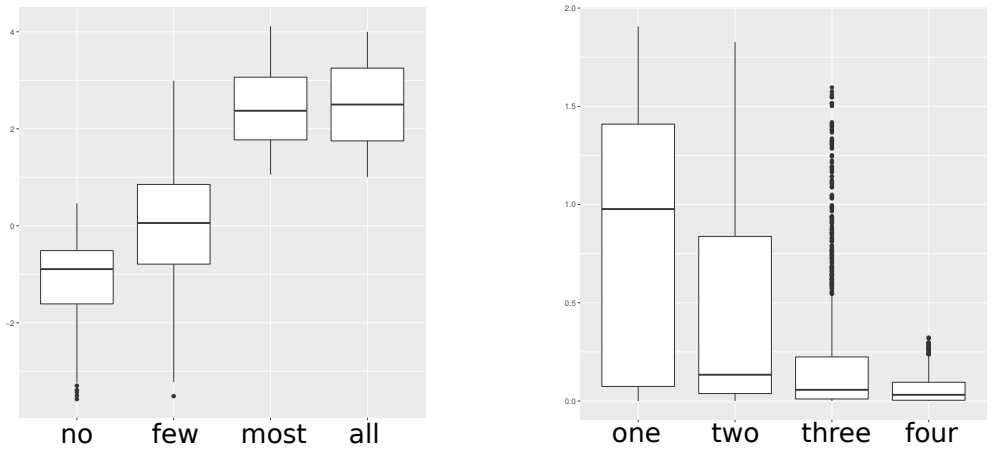

Figure 5.3: Left: Qs against dot product. Right: Cs against cosine distance.

selected concepts to its corresponding 4 scenarios in Train (see section 5.2.2). To illustrate, the meaning of *few* is learned by mapping each word into the 4 visual scenes where the amount of 'targetness' is less than 50% (see section 5.2), whereas *two* is learned by mapping each word to the scenarios where the number of targets is 2, and so on. This mapping, we conjecture, would mimic the multimodal mechanism by which children acquire the meaning of both Cs and Qs (see Halberda et al. (2008)). Once learned, the function representing each C/Q can be evaluated against scenarios containing an unseen mixture of (known) target objects and distractors. If it has encoded the correct meaning of the quantified expression, the function will retrieve the unseen scenarios containing the correct quantity (either exact or fuzzy) of target objects.

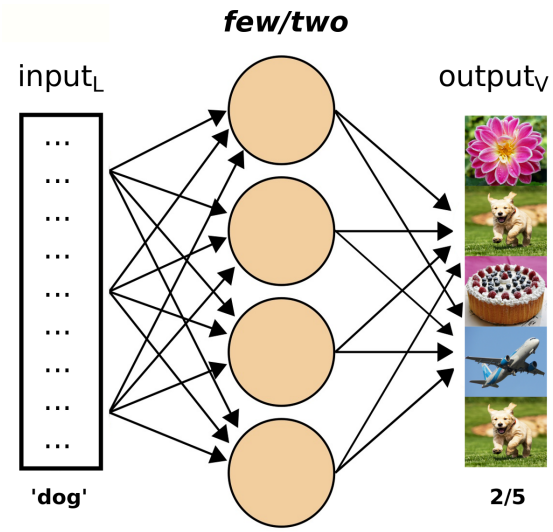

Figure 5.4: One learning event of our proposed cross-modal mapping. Cosine is used for quantifiers (*few*), dot product for cardinals (*two*).

	lin		**nn-cos**		**nn-dot**	
	mAP	*P2*	*mAP*	*P2*	*mAP*	*P2*
no	0.78	0.65	**0.87**	0.77	0.54	0.37
few	0.59	0.39	**0.68**	0.51	0.59	0.43
most	0.61	0.36	0.60	0.29	**0.62**	0.45
all	0.75	0.66	**1**	1	0.33	0.12
one	0.44	0.30	0.38	0.21	**0.61**	0.45
two	0.35	0.15	0.38	0.21	**0.57**	0.43
three	0.38	0.16	0.36	0.13	**0.56**	0.40
four	0.65	0.47	0.75	0.60	**0.76**	0.61

Table 5.2: *mAP* and *P2* for each model.

We experiment with three different models: linear (lin), cosine neural network (nn-cos), dot-product neural network (nn-dot). The first model is a simple linear mapping. The second is a single-layer neural network (activation function ReLU) that maximizes the cosine similarity between input (linguistic) and output vector (visual). The third is a similar neural network that approximates to 1 the dot product between input and output. We evaluate the mapping functions by means of a retrieval task aimed at picking up the correct scenarios from Test among the set of 8 scenarios built upon the same target object. Recall that in Test there are 2 combinations * 4 C/Q classes for each concept.

5.4 Results

As reported in Table 5.2, nn-cos is overall the best model for Qs, whereas nn-dot is the best model for Cs. In particular, mean average precision (*mAP*) is higher in nn-cos for 3 out of 4 Qs, with only *most* reaching slightly better mAP in Q nn-dot due to the high number of cases confounded with *all* by the Q nn-cos model (see Table 5.3). Conversely, both mAP and precision at top-2 positions (*P2*) for Cs are always higher in nn-dot compared to the other models. From a qualitative analysis of the results, it emerges that both the best-predictive models make 'plausible' errors, i.e. they confound Cs/Qs that are close to each other in the ordered scale. Table 5.3 reports the confusion matrices for the best performing models. Besides retrieving more cases of *all* instead of (correct) *most*, the Q nn-cos model often confounds *few* with *no*. Similarly, the C nn-dot model often confounds *three* with *four*, *one* with *two*, *two* with *three*, and so on. Overall, both models pick up very few or no responses that are on the opposite end of the 'scale', thus suggesting that the meaning representation they learn encodes, to a

	no	few	most	all		one	two	three	four
no	**288**	88	0	0	one	**168**	113	54	41
few	141	**191**	38	6	two	64	**136**	124	52
most	0	0	111	**265**	three	23	80	130	**145**
all	0	0	0	**376**	four	10	24	72	**272**

Table 5.3: Left: Q nn-cos, retrieved cases in top-2 positions. Right: same for C nn-dot.

certain extent, information about the ordered position of the quantified expressions.

5.5 Discussion

5.5.1 Two Mechanisms

In this chapter, I explored the computational mechanisms underlying the learning of cardinals and quantifiers from vision. Based on the evidence that these expressions are governed by different cognitive and neural mechanisms, I tested whether distinct operations are needed also on the computational level. In particular, I proposed that a model capitalizing on a 'precise' objective function (dot product) is required for the learning of cardinals, whereas quantifiers would be better modeled by a 'fuzzy' function (cosine). By means of a language-to-vision mapping, I showed the validity of such assumption: On the one hand, cardinals and quantifiers were shown to be better modeled by dot product and cosine, respectively; on the other, best-performing neural networks turned out to outperform linear models. This finding is in line with the evidence that, in grounded contexts, cardinals are described by the precise numerosity of the set, with quantifiers being rather represented by approximate proportional information (see Chapter 4).

5.5.2 One Expression, One Model

In this work, we focused on the objective functions needed to learn cardinals and quantifiers and we employed a 'one expression, one model' approach. That is, we modeled each cardinal (e.g. *one*) and quantifier (e.g. *few*) via a dedicated network rather than using a unique model for all. This setting was partially inspired by neuroscience work suggesting that, in human brain, each number would activate specific neurons, also known as 'number neurons' (Nieder, 2016). Moreover, this approach allowed us to better contrast the two versions of each model (three, if we include the linear one) and

gain a better understanding of the role of the objective function. However, a valuable and only partially competing approach would be to implement a single model for learning several cardinals or quantifiers at a time. Even further, one possibility would be to test a unique architecture in the task of modeling jointly cardinals and quantifiers, or quantifiers and other, more compatible quantity expressions (e.g. comparatives).

5.5.3 Limitations

Though we proved the validity of our intuition on the different learning mechanisms, both the visual scenes (fully synthetic) and the definition of quantifiers (fixed ranges of proportions) were arguably rather simplistic. In the next chapter, I overcome most of these limitations by both using the more complex visual data introduced in Chapter 4 and by computationally modeling the probabilities associated with the human choice of quantifiers in grounded contexts. Following the intuition described in section 5.5.2, I propose a multi-task learning architecture for jointly modeling quantifiers ('most'), comparatives ('more'), and proportions ('80%'). As for the visual scenes, they differ from those used in this chapter by several aspects: They depict a higher number of total objects (up to 20) and the size, orientation and spatial arrangement of the objects are randomly varied.

Chapter 6

A Multi-Task Model for Learning Quantity Expressions from Vision

In this chapter I study whether different quantification mechanisms (set comparison, vague quantification, and proportional estimation) can be jointly learned from visual scenes by a multi-task computational model. The motivation is that, in humans, these processes underlie the same cognitive, non-symbolic ability, which allows an automatic estimation and comparison of set magnitudes. I show that when information about lower-complexity tasks is available, the higher-level proportional task becomes more accurate than when performed in isolation. Moreover, the multi-task model is able to generalize to unseen combinations of target/non-target objects. Consistently with behavioral evidence showing the interference of absolute number in the proportional task, the multi-task model no longer works when asked to provide the number of target objects in the scene.

6.1 Introduction

Understanding and producing sentences like 'There are *more* cars than parking lots', '*Most* of the supporters wear blue t-shirts', '*Twenty percent* of the trees have been planted last year', or '*Seven* students passed the exam', is a fundamental competence which allows speakers to communicate information about quantities. Crucially, the type of information conveyed by these expressions, as well as their underlying cognitive

mechanisms, are not equivalent, as suggested by evidence from linguistics, language acquisition, and cognition.

First, comparatives ('more', 'less'), quantifiers ('some', 'most', 'all'), and proportions ('20%', 'two thirds') express a comparison or relation *between sets* (e.g., between the set of cars and the set of parking lots). Such relational information is rather coarse when expressed by comparatives and vague quantifiers, more precise when denoted by proportions. In contrast, numbers ('one', 'six', 'twenty-two') denote the exact, absolute cardinality of the items belonging to *one set* (e.g., the set of students who passed the exam).

Second, during language acquisition, these expressions are neither learned at the same time nor governed by the same rules. Recent evidence showed that children can understand comparatives at around 3.3 years (Odic et al., 2013; Bryant, 2017), with quantifiers being learned a few months later, at around 3.4-3.6 years (Hurewitz et al., 2006; Minai, 2006; Halberda et al., 2008). Crucially, knowing the meaning of numbers, an ability that starts not before the age of 3.5 years (Le Corre and Carey, 2007), is not required to understand and use these expressions. As for proportions, they are acquired significantly later, being fully mastered only at the age of 9 or 10 (Hartnett and Gelman, 1998; Moss and Case, 1999; Sophian, 2000).

Third, converging evidence from cognition and neuroscience supports the hypothesis that some important components of these expressions of quantity are grounded on a preverbal, non-symbolic system representing magnitudes (Piazza, 2010). This system, often referred to as Approximate Number System (ANS), is invariant to the sensory modality and almost universal in the animal domain, and consists in the ability of holistically extracting and comparing approximate numerosities (Piazza and Eger, 2016). In humans, it is present since the youngest age, with 6-month-old infants being able to automatically compare sets and combine them by means of proto-arithmetical operations (Xu and Spelke, 2000; McCrink and Wynn, 2004). Since it obeys Weber's law, according to which highly differing sets (e.g. 2:8) are easier to discriminate than highly similar sets (e.g. 7:8), ANS has been recently claimed to be a *ratio-based* mechanism (Sidney et al., 2017; Matthews et al., 2016). In support of this, behavioral findings indicate that, in non-symbolic contexts (e.g. visual scenes), proportional values are extracted holistically, i.e. without relying on the pre-computed cardinalities of the sets (Fabbri et al., 2012; Yang et al., 2015). Indeed, people are fairly accurate in providing the proportion of targets in a scene, even in high-speed settings (Healey et al., 1996;

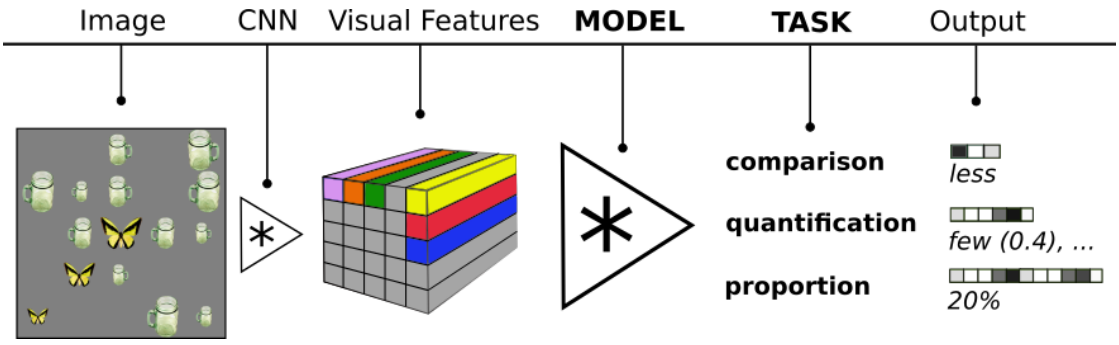

Figure 6.1: Toy representation of the quantification tasks and corresponding outputs explored in the chapter. Note that quantification always refers to animals (target set).

Treisman, 2006). Similarly, in briefly-presented scenes, the interpretation of quantifiers is shown to be best described by proportional information (see Pezzelle et al. (2018) or Chapter 4).

Altogether, this suggests that performing (1) set comparison, (2) vague quantification, and (3) proportional estimation, which all rely on information regarding relations among sets, underlies increasingly-complex steps of the same mechanism. Notably, such complexity would range from 'more/less' judgements to proportional estimation, as suggested by the increasing precision of ANS through years (Halberda and Feigenson, 2008), the reported boundary role of 'half' in early proportional reasoning (Spinillo and Bryant, 1991), and the different age of acquisition of the corresponding linguistic expressions. Finally, the ratio-based operation underlying these task would be different from (and possibly conflicting with) that of estimating the absolute numerosity of one set. Indeed, absolute numbers are found to interfere with the access to proportions (Fabbri et al., 2012).

Inspired by this converging evidence, the present work proposes a computational framework to explore various quantification tasks in the visual domain (see Figure 6.1). In particular, we investigate whether ratio-based quantification tasks can be modeled by a single, multi-task learning neural network. Given a synthetic scene depicting animals (in our setting, the 'target' objects) and artifacts ('non-target'), our model is designed to jointly perform all the tasks by means of an architecture that reflects their increasing complexity.[1] To perform proportional estimation (the most complex), the model builds on the representations learned to perform vague quantification and, in turn, set comparison (the least complex). We show that the multi-task model achieves both higher accuracy and higher generalization power compared to the one-task models. In con-

[1]Data and code can be found at github.com/sandropezzelle/multitask-quant

trast, we prove that introducing the absolute number task in the loop is not beneficial and indeed hurts the performance.

Our main contribution lies in the novel application and evaluation of a multi-task learning architecture on the task of jointly modeling 3 different quantification operations. On the one hand, our results confirm the interdependency of the mechanisms underlying the tasks of set comparison, vague quantification, and proportional estimation. On the other, we provide further evidence on the effectiveness of these computational architectures.

6.2 Related Work

6.2.1 Quantities in Language & Vision

In recent years, the task of extracting quantity information from visual scenes has been tackled via Visual Question Answering (VQA). Given a real image and a natural language question, a VQA computational model is asked to understand the image, the linguistic query, and their interaction to provide the correct answer. So-called *count* questions, i.e. 'How many *Xs* have the property *Y*?', are very frequent and have been shown to be particularly challenging for any model (Antol et al., 2015; Malinowski et al., 2015; Ren et al., 2015; Fukui et al., 2016). The difficulty of the task has been further confirmed by the similarly poor performance achieved even on the 'diagnostic' datasets, which include synthetic visual scenes depicting geometric shapes (Johnson et al., 2017; Suhr et al., 2017).

Using Convolutional Neural Networks (CNN), a number of works in Computer Vision (CV) have proposed specific architectures for counting digits (Seguí et al., 2015), people in the crowd (Zhang et al., 2015a), and penguins (Arteta et al., 2016). With a more cognitive flavor, Chattopadhyay et al. (2017) employed a 'divide-and-conquer' strategy to split the image into subparts and count the objects in each subpart by mimicking the 'subitizing' mechanism (i.e. numerosities up to 3-4 can be rapidly and accurately appreciated). Inspired by the same cognitive ability is Zhang et al. (2015b), who trained a CNN to detect and count the salient objects in the image. Except Suhr et al. (2017), who evaluated models against various types of quantity expressions (including existential quantifiers), these works were just focused on the absolute number.

More akin to our work is Stoianov and Zorzi (2012), who showed that hierarchical gen-

erative models learn ANS as a statistical property of (synthetic) images. Their networks were tested on the task of set comparison ('more/less') and obtained 93% accuracy. A few studies specifically focused on the learning of quantifiers. Sorodoc et al. (2016) proposed a model to assign the correct quantifier to synthetic scenes of colored dots, whereas Sorodoc et al. (2018) operationalized the same task in a VQA fashion, using real images and object-property queries (e.g. 'How many *dogs* are *black*?'). Overall, the results of these studies showed that vague quantification can be learned by neural networks, though the performance is much lower when using real images and complex queries. Finally, Pezzelle et al. (2017) (Chapter 5) investigated the difference between the learning of cardinals and quantifiers from visual scenes, showing that they require two distinct computational operations. To our knowledge, this is the first attempt to jointly investigate the whole range of quantification mechanisms. Moreover, we are the first to exploit a multi-task learning paradigm for exploring the interactions between set comparison, vague quantification, and proportions.

6.2.2 Multi-Task Learning

Multi-Task Learning (MTL) has been shown to be very effective for a wide range of applications in machine learning (for an overview, see Ruder (2017)). The core idea is that different and yet related tasks can be jointly learned by a multi-purpose model rather than by separate and highly fine-tuned models. Since they share representations between related (or 'auxiliary') tasks, multi-task models are more robust and generalize better than single-task models. Successful applications of MTL have been proposed in CV to improve object classification (Girshick, 2015), face detection and rotation (Zhang et al., 2014; Yim et al., 2015), and to jointly perform a number of tasks as object detection, semantic segmentation, etc. (Misra et al., 2016; Li and Hoiem, 2016). Though, recently, a few studies applied MTL techniques to either count or estimate the number of objects in a scene (Sun et al., 2017; Sindagi and Patel, 2017), to our knowledge none of them were devoted to the learning of various quantification mechanisms.

In the field of natural language processing (NLP), MTL turned out to be beneficial for machine translation (Luong et al., 2016) and for a range of tasks such as chunking, tagging, semantic role labelling, etc. (Collobert et al., 2011; Søgaard and Goldberg, 2016; Bingel and Søgaard, 2017). In particular, Søgaard and Goldberg (2016) showed the benefits of keeping low-level tasks at the lower layers of the network, a setting which enables higher-level tasks to make a better use of the shared representations.

Since this finding was also in line with previous evidence suggesting a natural order among different tasks (Shen and Sarkar, 2005), further work proposed MTL models in which several increasingly-complex tasks are hierarchically ordered (Hashimoto et al., 2017). The intuition behind this architecture, referred to as 'joint many-task model' in the source paper (Hashimoto et al., 2017), as well as its technical implementation, constitute the building blocks of the model proposed in the present study.

6.3 Tasks and Dataset

6.3.1 Tasks

Given a visual scene depicting a number of animals (targets) and artifacts (non-targets), we explore the following tasks, represented in Figure 6.1:

(a) set comparison (hence, **setComp**), i.e. judging whether the targets are 'more', 'same', 'less' than non-targets;

(b) vague quantification (hence, **vagueQ**), i.e. predicting the probability to use each of the 9 quantifiers ('none', 'almost none', 'few', 'the smaller part', 'some', 'many', 'most', 'almost all', 'all') to refer to the target set;

(c) proportional estimation (hence, **propTarg**), i.e. predicting the proportion of targets choosing among 17 ratios, ranging from 0 to 100%.

Tasks (a) and (c) are operationalized as classification problems and evaluated through accuracy. That is, only one answer out of 3 and 17, respectively, is considered as correct. Given the vague status of quantifiers, whose meanings are 'fuzzy' and overlapping, task (b) is evaluated by means of Pearson's correlation (r) between the predicted and the ground-truth probability vector (cf. section 6.3.2), for each datapoint.[2] The overall r is obtained by averaging these scores. It is worth mentioning that we could either evaluate (b) in terms of a classification task or operationalize (a) and (c) in terms of a correlation with human responses. The former evaluation is straightforward and can be easily carried out by picking the quantifier with the highest probability. The latter,

[2]We also experimented with Mean Average Error and dot product and found the same patterns of results (not reported).

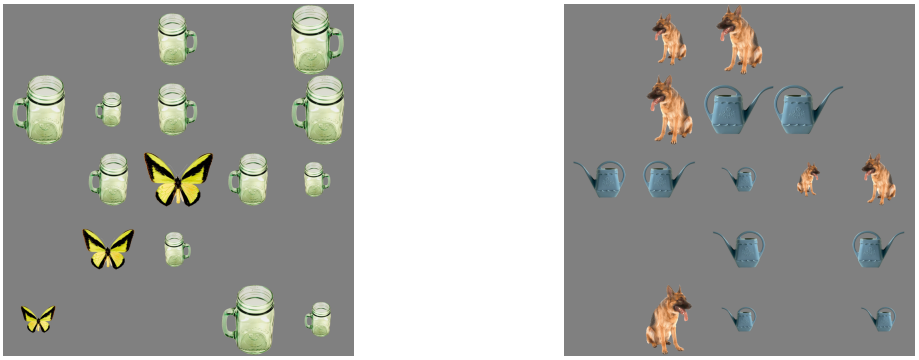

Figure 6.2: Two scenes included in our dataset. The letfmost one depicts a ratio 1:4 (3 animals, 12 artifacts, 15 total items), the rightmost one a ratio 2:3 (6, 9, 15).

in contrast, implies relying on behavioral data assessing the degree of overlap between ground-truth classes and speakers' choice. Though interesting, such evaluation is less crucial given the discrete, non-overlapping nature of the classes in tasks (a) and (c).

The tasks are explored by means of a MTL network that jointly performs the three quantification operations (see section 6.4.2). The intuition is that solving the lower-level tasks would be beneficial for tackling the higher-level ones. In particular, providing a proportional estimation ('80%') after performing vagueQ ('most') and setComp ('more') should lead to a higher accuracy in the highest-level task, which represents a further step in complexity compared to the previous ones. Moreover, lower-level tasks might be boosted in accuracy by the higher-level ones, since the latter include all the operations that are needed to carry out the former. In addition to the MTL model, we test a number of 'one-task' networks specifically designed to solve one task at a time (see section 6.4.1).

6.3.2 Dataset

We built a large dataset of synthetic visual scenes depicting a variable number of animals and artifacts on the top of a neutral, grey background (see Figure 6.2). In doing so, we employed the same methodology and materials used in Chapter 4, where the use of quantifiers in *grounded* contexts was explored by asking participants to select the most suitable quantifier for a given scene. Since the category of animals was always treated as the 'target', and that of artifacts as the 'non-target', we will henceforth use this terminology throughout the chapter. The scenes were automatically generated by an in-house script using the following pipeline: (a) Two natural images, one depict-

	train	val	test	**total**
no. datapoints	11.9K	1.7K	3.4K	**17K**
% datapoints	70%	10%	20%	**100%**

Table 6.1: Number and partitioning of the datapoints.

ing a target object (e.g. a butterfly) and one depicting a non-target (e.g. a mug) were randomly picked up from a sample of the dataset by Kiani et al. (2007). The sample was obtained in Chapter 4, where we manually selected pictures depicting whole items (not just parts) and whose color, orientation and shape were not deceptive. In total, 100 unique instances of animals and 145 unique instances of artifacts were included; (b) The proportion of targets in the scene (e.g. 20%) was chosen by selecting one among 17 pre-defined *ratios* between targets:non-targets (e.g. 1:4, 'four non-targets to one target'). Out of 17 ratios, 8 were positive (targets $>$ 50%), 8 negative (targets $<$ 50%), and 1 equal (targets = 50%); (c) The absolute number of targets/non-targets was chosen to equally represent the various combinations available for a given ratio (e.g., for ratio 1:4: 1-4, 2-8, 3-12, 4-16), with the constraint of having a number of total objects in the scene (targets+non-targets) ranging from 3 to 20. In total, 97 combinations were represented in the dataset, with an average of 5.7 combinations/ratio (min 2, max 18); (d) To inject some variability, the instances of target/non-target objects were randomly resized according to one of three possible sizes (i.e. medium, big, and small) and flipped on the vertical axis before being randomly inserted onto a 5*5-cell virtual grid. As reported in Table 6.1, 17K scenes balanced per ratio (1K scenes/ratio) were generated and further split into train (70%), validation (10%), and test (20%).

Ground-truth classes for the tasks of setComp and propTarg were automatically assigned to each scene while generating the data. For vagueQ, we took the probability distributions obtained on a dataset of 340 scenes (see Chapter 4) and we applied them to our datapoints, which were built in the exact same way. These probability distributions had been collected by asking participants to select, from a list of 9 quantifiers (reported in section 6.3.1), the most suitable one to describe the target objects in a visual scene presented for 1 second. In particular, they were computed against the proportion of targets in the scene, which in that study was shown to be the overall best predictor for quantifiers. To illustrate, given a scene containing 20% of targets (cf. leftmost panel in Figure 6.2), the probability of choosing 'few' (ranging from 0 to 1) is 0.38, 'almost none' 0.27, 'the smaller part' 0.25, etc. It is worth mentioning that, for scenes containing either 100% or 0% targets the probability of choosing 'all' and 'none', respectively, is around 1. In all other cases, the distribution of probabilities is fuzzier and reflects the

largely overlapping use of quantifiers, as in the example above. On average, the probability of the most-chosen quantifier across ratios is 0.53. Though this number cannot be seen as a genuine inter-annotator agreement score, it suggests that, on average, there is one quantifier which is preferred over the others.

6.4 Models

In this section, we describe the various models implemented to perform the tasks. For each model, several settings and parameters were evaluated by means of a thorough ablation analysis. Based on a number of factors like performance, speed, and stability of the networks, we opted for using ReLU nonlinear activation at all hidden layers and the simple and effective Stochastic Gradient Descent (SGD) as optimizer (lr = 0.01). We run each model for 100 epochs and saved weights and parameters of the epoch with the lowest validation loss. The best model was then used to obtain the predictions in the test set. All models were implemented using Keras.[3]

6.4.1 One-Task Models

We implemented separate models to tackle one task at a time. For each task, in particular, both a network using 'frozen' (i.e. pretrained) visual features and one computing the visual features in an 'end-to-end' fashion were tested.

One-Task-Frozen These models are simple, 2-layer (ReLU) Multi-Layer Perceptron (MLP) networks that take as input a 2048-d frozen representation of the scene and output a vector containing softmax probability values. The frozen representation of the scene had been previously extracted using the state-of-art Inception v3 CNN (Szegedy et al., 2016) pretrained on ImageNet (Deng et al., 2009). In particular, the network is fed with the average of the features computed by the last Convolutional layer, which has size 25*2048.

One-Task-End2end These models are MLP networks that take as input the 203*203-pixel image and compute the visual features by means of the embedded Inception v3

[3] https://keras.io/

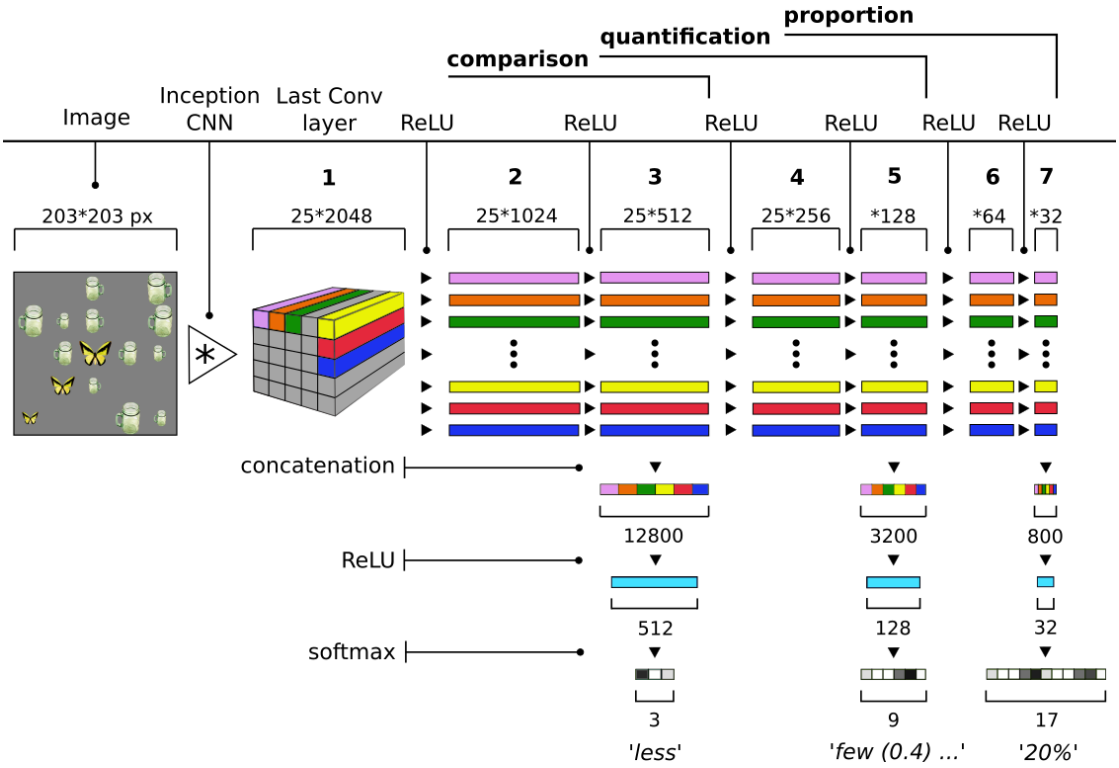

Figure 6.3: Architecture of the `multi-task-prop` model jointly performing (a) set comparison, (b) vague quantification, and (c) proportional estimation. Given a 203*203-pixel image as input, the model extracts a 25*2048 representation from the last Convolutional layer of the Inception v3. Subsequently, the vectors are reduced twice via ReLU hidden layers to 1024 and 512 dimensions. The 512-d vectors are concatenated and reduced, then a softmax layer is applied to output a 3-d vector with probability distributions for task (a). The same structure (i.e., 2 hidden layers, concatenation, reduction, and softmax) is repeated for tasks (b) and (c). All the tasks are trained with cross-entropy. To evaluate tasks (a) and (c), in testing, we extract the highest-probability class and compute **accuracy**, whereas task (b) is evaluated via **Pearson's correlation** against the 9-d ground-truth probability vector.

module, which outputs 25*2048-d vectors (the grey and colored box in Figure 6.1). Subsequently, the 25 feature vectors are reduced twice via ReLU hidden layers, then concatenated, reduced (ReLU), and fed into a softmax layer to obtain the probability values.

6.4.2 Multi-Task Model

The `multi-task-prop` model performs 3 tasks at the *same time* with an architecture that reproduces in its *order* the conjectured complexity (see Figure 6.3 and its caption for technical details). The model has a core structure, represented by layers 1-5 in the

model	setComp	vagueQ	propTarg	nTarg
	accuracy	*Pearson r*	*accuracy*	*accuracy*
chance/majority	0.470	0.320	0.058	0.132
one-task-frozen	0.783	0.622	0.210	0.312
one-task-end2end	0.902	0.964	0.659	**0.966**
multi-task-prop	**0.995**	**0.982**	**0.918**	–
multi-task-number	0.854	0.807	–	0.478

Table 6.2: Performance of the models in the tasks of set comparison (setComp), vague quantification (vagueQ), proportional estimation (propTarg), and absolute number of targets (nTarg). Values in **bold** are the highest.

figure, which is *shared* across tasks and trained with multiple outputs. In particular, (a) layers 1, 2, and 3 are trained using information regarding the output of all 3 tasks. That is, these layers are updated three times by as many backpropagation passes: One on the top of setComp output, the second on the top of vagueQ output, the third on the top of propTarg output; (b) layers 4 and 5 are affected by information regarding the output of vagueQ and propTarg, and thus updated twice; (c) layers 6 and 7 are updated once, on the top of the output of propTarg. Importantly, the three lower layers in Figure 6.3 (concatenation, ReLU, softmax) are not shared between the tasks, but specialized to output each a specific prediction. As can be noted, the order of the tasks reflects their complexity, since the last task in the pipeline has 2 more layers than the preceding one and 4 more than the first one.

6.5 Results

Table 6.2 reports the performance of each model in the various tasks (note that the lowest row and the rightmost column report results described in section 6.6.1). In setComp, all the models are neatly above chance/majority level (0.47). The `one-task-end2end` model achieves a remarkable 0.90 acc., which is more than 10% better compared to the simple `one-task-frozen` model (0.78). The same pattern of results can be observed for vagueQ, where the Pearson's correlation (r) between the ground-truth and the predicted probability vector is around 0.96, that is more than 30% over the simpler model (0.62). This gap increases even more in propTarg, where the accuracy of the frozen model is more than 40 points below the one achieved by the `one-task-end2end` model (0.21 against 0.66). These results firmly indicate that, on the one hand, the frozen representation of the visual scene encodes little information about the propor-

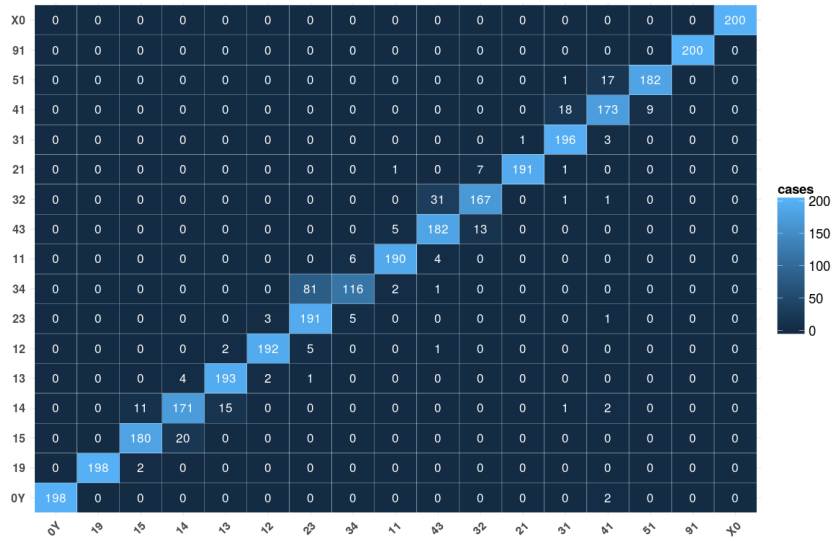

Figure 6.4: PropTarg. Heatmap reporting the errors made by the `multi-task-prop` model. Note that labels refer to *ratios*, i.e. 14 stands for ratio 1:4 (20% targets).

tion of targets (likely due to the the different task for which they were pretrained, i.e. object classification). On the other hand, computing the visual features in an end-to-end fashion leads to a significant improvement, suggesting that the network learns to pay attention to features that are helpful for specific tasks.

The most interesting results, however, are those achieved by the multi-task model, which turns out to be the best in all the tasks. As reported in Table 6.2, sharing the weights between the various tasks is especially beneficial for propTarg, where the accuracy reaches 0.92, that is, more than 25 points over the end-to-end, one-task model. An almost perfect performance of the model in this task can be observed in Figure 6.4, which reports the confusion matrix with the errors made by the model. As can be seen, the few errors are between 'touching' classes, e.g. between ratio 3:4 (43% of targets) and ratio 2:3 (40%). Since these classes differ by a very small percentage, we gain indirect evidence that the model is learning some kind of proportional information rather than trivial associations between scenes and orthogonal classes.

To further explore this point, one way is to inspect the last layer of the proportional task (i.e. the 32-d turquoise vector in Figure 6.3). If the vectors contain information regarding the proportion of targets, we should expect scenes depicting the same proportion to have a similar representation. Also, scenes with similar proportions (e.g. 40% and 43%) would be closer to each other than are scenes with different proportions (e.g. 25% and 75%). Figure 6.5 depicts the results of a two-dimensional PCA analysis performed

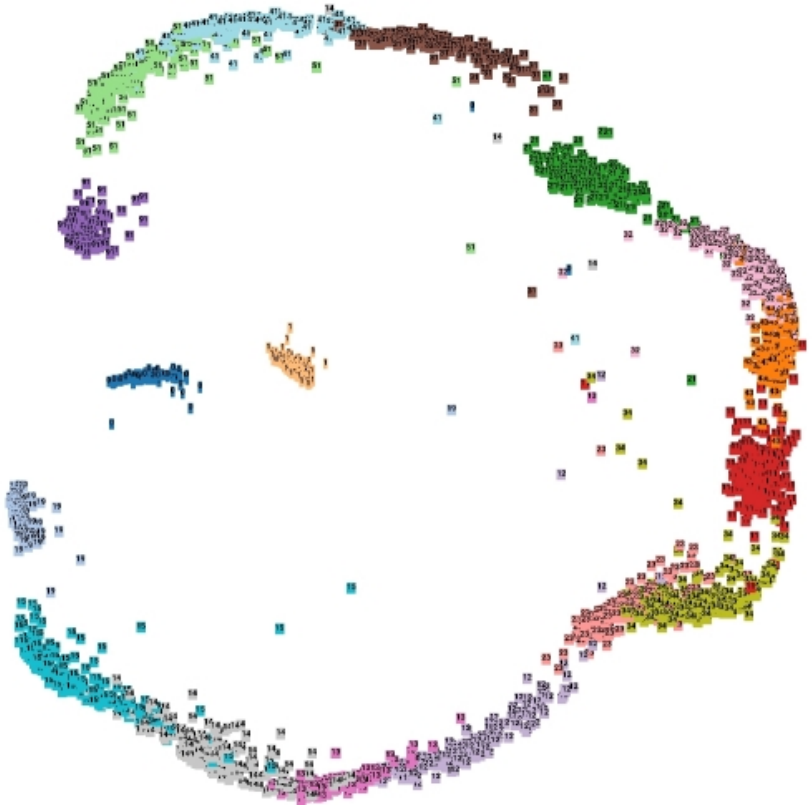

Figure 6.5: PCA visualization of the last layer (before softmax) of the proportional task in the MTL model.

on the vectors of the last layer of the proportional task (the 32-d vectors).[4] As can be noted, scenes depicting the same proportion clearly cluster together, thus indicating that using these representations in a retrieval task would lead to a very high precision. Crucially, the clusters are perfectly ordered with respect to proportion. Starting from the purple cluster on the left side (90%) and proceeding clockwise, we find 83% (green), 80% (turquoise), 75% (brown), and so on, until reaching 10% (light blue). Proportions 0% (blue) and 100% (yellow) are neatly separated from the other clusters, being at the extremes of the 'clock'.

An improvement in the results can be also observed for setComp and vaqueQ, where the model achieves 0.99 acc. and 0.98 *r*, respectively. Figure 6.6 reports, for each quantifier, the probability values predicted by the model against the ground-truth ones. As can be seen, the red lines (model) approximate very closely the green ones (humans). In the following section, we perform further experiments to provide a deeper evaluation of the results.

[4] We used https://projector.tensorflow.org/

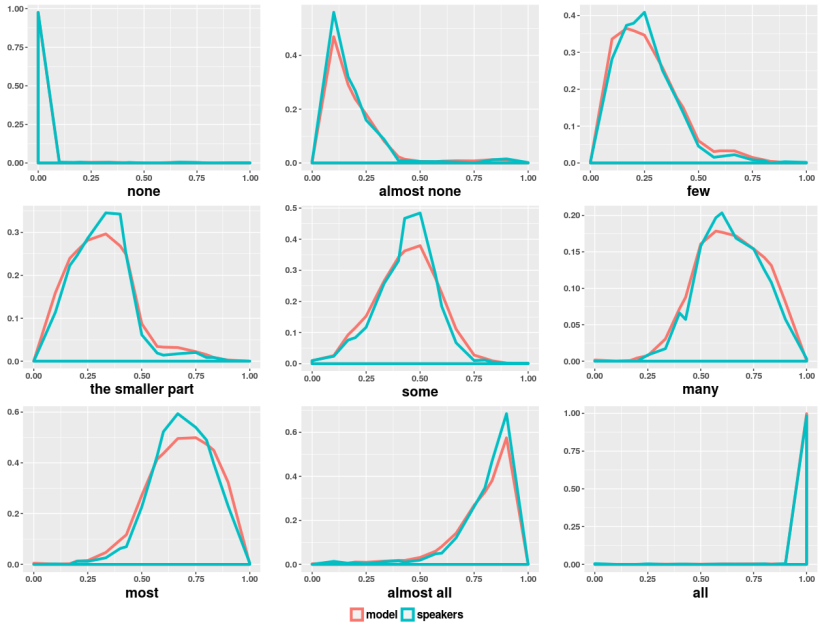

Figure 6.6: VagueQ. Probability values predicted by the `multi-task-prop` model against ground-truth probability distributions for each quantifier.

6.6 In-Depth Evaluation

6.6.1 Absolute Numbers in the Loop

As discussed in section 6.1, the cognitive operation underlying setComp, vagueQ, and propTarg is different compared to that of estimating the absolute number of objects included in one set. To investigate whether such dissociation emerges at the computational level, we tested a modified version of our proposed multi-task model where propTarg task has been replaced with nTarg, namely the task of predicting the absolute number of targets. One-task models were also tested to evaluate the difficulty of the task when performed in isolation. Since the number of targets in the scenes ranges from 0 to 20, nTarg is evaluated as a 21-class classification task (majority class 0.13).

As reported in Table 6.2, the accuracy achieved by the `one-task-end2end` model is extremely high, i.e. around 0.97. This suggests that, when learned in isolation, the task is fairly easy, but only if the features are computed *within* the model. In fact, using frozen features results in a quite low accuracy, namely 0.31. This pattern of results is even more interesting if compared against the results of the `multi-task-number` model. When included in the multi-task pipeline, in fact, nTarg has a huge, 50-point accuracy drop (0.48). Moreover, both setComp and vagueQ turn out to be significantly

hurt by the highest-level task, and experience a drop of around 14 and 17 points compared to the `one-task-end2end` model, respectively. These findings seem to corroborate the incompatibility of the operations needed for solving the tasks.

6.6.2 Reversing the Architecture

Previous work exploring MTL suggested that defining a hierarchy of increasingly complex tasks is beneficial for jointly learning related tasks (see section 6.2.2). In the present work, the order of the tasks was inspired by cognitive and linguistic abilities (see section 6.1). Though cognitively implausible, it might still be the case that the model is able to learn even when reversing the order of the tasks, i.e. from the conjectured highest-level to the lowest-level one. To shed light on this issue, we tested the `multi-task-prop` model after reversing its architecture. That is, propTarg is now the first task, followed by vagueQ, and setComp.

In contrast with the pattern of results obtained by the original pipeline, no benefits are observed for this version of MTL model compared to one-task networks. In particular, both vagueQ (0.32 *r*) and propTarg (0.08 acc.) performance are around chance level, with setComp reaching just 0.65 acc., i.e. 25 point lower than the `one-task-end2end` model. The pipeline of increasing complexity motivated theoretically is thus confirmed at the computational level.

6.6.3 Does MTL Generalize?

As discussed in section 6.2.2, MTL is usually claimed to allow a higher generalization power. To investigate whether our proposed `multi-task-prop` model genuinely learns to quantify from visual scenes, and not just associations between patterns and classes, we tested it with unseen combinations of targets/non-targets. The motivation is that, even in the most challenging propTarg task, the model might learn to match a given combination, e.g. 3:12, to a given proportion, i.e. 20%. If this is the case, the model would solve the task by learning "just" to assign a class to each of the 97 possible combinations included in the dataset. If it learns a more abstract representation of the proportion of targets depicted in the scene, in contrast, it should be able to generalize to unseen combinations.

model	setComp	vagueQ	propTarg
	accuracy	*Pearson r*	*accuracy*
chance/majority	0.470	0.320	0.058
one-task-frozen	0.763	0.548	0.068
one-task-end2end	0.793	0.922	0.059
multi-task-prop	**0.943**	**0.960**	**0.539**

Table 6.3: Unseen dataset. Performance of the models in each task. Values in **bold** are the highest.

We built an additional dataset using the exact same pipeline described in section 6.3.2. This time, however, we randomly selected one combination per ratio (17 combinations in total) to be used only for validation and testing. The remaining 80 combinations were used for training. A balanced number of datapoints for each combination were generated in val/test, whereas datapoints in training set were balanced with respect to ratios, by randomly selecting scenes among the remaining combinations. The *unseen* dataset included around 14K datapoints (80% train, 10% val, 10% test).

Table 6.3 reports the results of the models on the unseen dataset. Starting from setComp, we note a similar and fairly high accuracy achieved by the two one-task models (0.76 and 0.79, respectively). In vagueQ, in contrast, the `one-task-end2end` model neatly outperforms the simpler model (0.92 vs. 0.55 *r*). Finally, in propTarg both models are at chance level, with an accuracy that is lower than 0.07. Overall, this pattern of results suggests that propTarg is an extremely hard task for the separate models, which are not able to generalize to unseen combinations. The `multi-task-prop` model, in contrast, shows a fairly high generalization power. In particular, it achieves 0.54 acc. in propTarg, that is, almost 10 times chance level.

The overall good performance in predicting the correct proportion can be appreciated in Figure 6.7, where the errors are represented by means of a heatmap. The error analysis reveals that end-of-the-scale proportions (0% and 100%) are the easiest, followed by proportions 75% (3:1), 67% (2:1), 50% (1:1), and 60% (3:2). More in general, negative ratios (targets $< 50\%$) are mispredicted to a much greater extent than are positive ones. Moreover, the model shows a bias toward some proportions, that the model seems 'to see everywhere'. However, the fact that the errors are found among the adjacent ratios (similar proportions) seems to be a convincing evidence that the model learns representations encoding genuine proportional information. Finally, it is worth mentioning that in setComp and vagueQ the model achieves very high results, 0.94 acc. and 0.96 *r*, respectively.

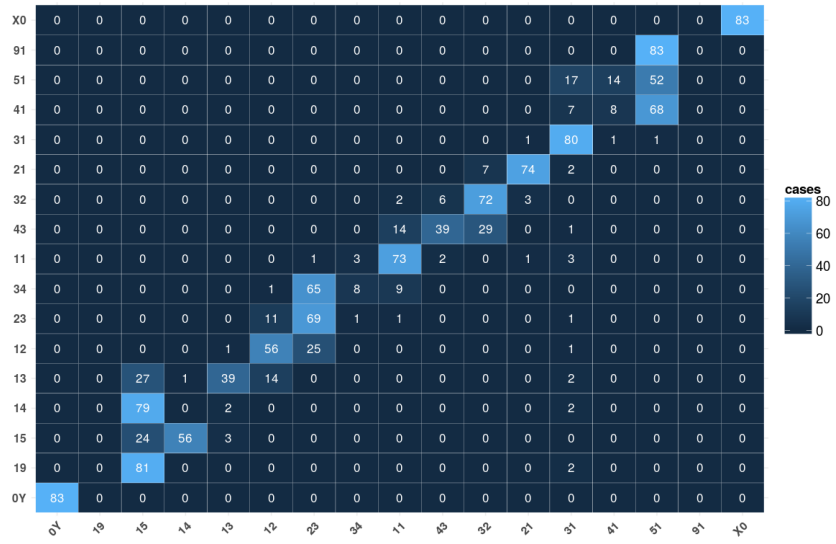

Figure 6.7: PropTarg. Heatmap with the errors made by the `multi-task-prop` model in the unseen dataset.

6.7 Discussion

6.7.1 Ratio-Based Mechanisms

In this chapter, I investigated whether *ratio-based* quantification mechanisms, expressed in language by comparatives, quantifiers, and proportions, can be computationally modeled in vision exploiting Multi-Task Learning (MTL). I proved that sharing a common core turns out to boost the performance in all the tasks, supporting evidence from linguistics, language acquisition, and cognition. Moreover, I reported analyses indicating both the increasing complexity of the tasks and the high generalization power of MTL.

6.7.2 Quantifiers vs Numbers

As far as numbers are concerned, our results clearly show that learning the precise cardinality of one set requires a different, competing mechanism compared to the one needed for quantifiers. This, one the one hand, is in line with behavioral evidence showing the interference of precise number to the access to proportional information (Fabbri et al., 2012). On the other hand, these findings are in line with those reported in Chapter 5, where cardinals and quantifiers were found to require different computational functions.

Chapter 7

Conclusion

Quantifiers are mysterious creatures. On the one hand – the word itself leaves no doubt – they are used *to quantify*, that is, to express the quantity of something. On the other hand, these expressions are *vague*, that is, they have "a single but nonspecific meaning" (Tuggy, 1993, p. 168). As discussed in Chapter 1, quantifiers can be used in similar contexts as numbers or proportions, but the information they convey can be either purely quantitative or something more/different than quantities (see Chapter 2). Their intriguing status has fascinated theorists since Aristotle (see Bonevac (2012)), and a myriad of issues related to their use and comprehension, meaning and formalization have been explored by many perspectives (Chapter 2).

In this thesis, I focused on vague, frequently-used quantifiers ('none', 'few', 'almost all', 'many', 'all', etc.) from a novel, cognitively-inspired computational perspective. On the the cognitive level, I carried out several behavioral studies with human speakers (Chapter 3 and 4). On the computational level, I exploited recent advances in computational linguistics and computer vision to either compare state-of-the-art networks with human performance (Chapter 3) or model speakers' use of quantifiers in grounded contexts (Chapter 5 and 6).

In Chapter 3 I explored the role of linguistic context in modulating the choice of quantifiers in discourse. I showed that a broader context helps speakers in predicting the missing quantifier, whereas state-of-the-art neural language models are hurt by more context. Though the task turned out to be challenging, both humans and the models were able to grasp the magnitude of the missing quantifier. I considered this finding as an evidence in favor of an ordering (i.e., a scale) among quantifiers. With precisely the

aim of investigating the mental scale of quantifiers, in Chapter 4 I proposed two behavioral studies: One exploring the abstract representation of quantifier words, the other focusing on the use of quantifiers in grounded contexts. In both settings, the representation of quantifiers turned out to resemble that of numbers and continuous quantities, thus supporting the intuition that some important components of the meaning of these expressions are *quantitative*. When used to describe visual scenes, moreover, quantifiers turned out to be better described by proportions rather than numbers. Along these lines, in Chapter 5 I investigated the nature of the computational mechanisms underlying the learning of quantifiers and numbers from their use in multimodal contexts (language and vision). I showed that two different operations are required, in line with previous evidence. Building on all the previous findings, in Chapter 6 I proposed that comparatives, quantifiers and proportions might be governed by the same, relation-based mechanism. I showed that a multi-task neural network jointly learning the meaning of these expressions from visual scenes outperforms the models learning one task at a time. Also, consistently with previously-obtained results, I showed that numbers require a radically different operation.

These results lead to several additional questions. For example, can the computational architectures proposed in Chapter 5 and 6 be successfully applied to datasets of real scenes? Though we lack an empirical answer to this question, the encouraging results obtained by Sorodoc et al. (2018) in a Visual Question Answering (VQA) tasks involving quantifiers and real images seem to suggest that, in principle, moving to real scenes should be perfectly possible. However, we might obtain lower results due to an imperfect multi-object recognition, or because of the natural bias that is present in real images. Another question concerns the applicability of our computational methods to other modalities than vision. For example, is the pipeline of increasing complexity found in Chapter 6 specific to vision (non-symbolic level), or is it shared across modalities, *in primis* language? Since linguistic expressions of quantity are grounded on a non-symbolic system, we might expect that a model trained on one modality can be applied to another, at least to some extent. Even further, jointly learning representations from both modalities might represent an even more natural, human-like way to learn and refer to quantities. Finally, some issues remain open on the cognitive and linguistic level. As suggested by the results of the abstract task in Chapter 4 (the one involving quantifier words), the mental representation of quantifiers would be tied to quantity information. This holds, at least to some extent, when quantifiers are used in a linguistic context (Chapter 3). However, some other components were repeatedly noticed to

come into play in the linguistic use of quantifiers, such as the lexical-semantic effect of *antonymy*. To illustrate, 'few' was judged to be the most dissimilar item from 'many', though – if we put them on a scale – the most distant one should be 'all'. This issue, together with the impact of experimenting with a larger set of quantifiers including lower-frequency expressions, deserves to be investigated in future work.

It is worth mentioning that several intuitions and methodologies presented in this thesis can be applied to other domains than quantifiers. For example, an intuitively valuable application could involve *gradable* adjectives (GAs) like 'minuscule', 'small', 'big', 'very big', 'huge', etc. These expressions share a number of commonalities with quantifiers: They have a partially overlapping distribution, they can have antonyms ('many'-'few' and 'tall'-'short'), they can be gradable by degree adverbs ('very many books' and 'very big book') and by inflection for comparative and superlative degrees ('few', 'fewer', 'fewest' and 'tall', 'taller', 'tallest'). Moreover, both quantifiers and GAs are usually claimed to lie on ordered scales and have flexible, context-sensitive meanings. Finally, they are both learned by children in grounded contexts, by having experience of many instances uttered in real-life (Halberda et al., 2008; Barner and Snedeker, 2008). To mention some possible directions, humans and models could be tested in the task of guessing a missing GA from texts (Chapter 3), as well as a Multi-Task Learning approach (Chapter 6) could be applied to the learning of GAs from vision.

Overall, this thesis contributes to the theoretical debate on quantifiers and proves the validity of using a multi-perspective, multi-modal approach to the study of complex, high-level linguistic expressions. Being semantically *vague* but frequently used in real-life situations, quantifiers represent a particularly interesting case where language, perception, and human cognition are irretrievably intertwined. However, quantifiers are just one of the countless linguistic phenomena that might be investigated using a similar approach. I hope my thesis can be of inspiration for future work in this direction.

Bibliography

Hirotugu Akaike. 1973. Information Theory and an extension of the Maximum Likelihood Principle. In *2nd International Symposium on Information Theory, Budapest: Akademiai Kaido*. pages 267–281.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. In *Proceedings of NAACL-HLT*. pages 1545–1554.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*. pages 2425–2433.

Jack B Arnold. 1971. A multidimensional scaling study of semantic distance. *Journal of Experimental Psychology* 90(2):349.

Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. 2016. Counting in the wild. In *European Conference on Computer Vision*. Springer, pages 483–498.

Rolf Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language* 59(4):390–412.

David Barner, Amanda Libenson, Pierina Cheung, and Mayu Takasaki. 2009. Cross-linguistic relations between quantifiers and numerals in language acquisition: Evidence from Japanese. *Journal of experimental child psychology* 103(4):421–440.

David Barner and Jesse Snedeker. 2008. Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. *Child development* 79(3):594–608.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 23–32.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 238–247.

Jon Barwise and Robin Cooper. 1981. Generalized Quantifiers and Natural Language. *Linguistics and Philosophy* 4(2):159–219.

Bernard M Bass, Wayne F Cascio, and Edward J O'connor. 1974. Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology* 59(3):313.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. *EACL 2017* page 164.

Daniel Bonevac. 2012. A History of Quantification. *Logic: A History of its Central Concepts* 11.

Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. 2015. Salient object detection: A benchmark. *IEEE Transactions on Image Processing* 24(12):5706–5722.

Richard Breheny, Napoleon Katsos, and John Williams. 2006. Are generalised scalar implicatures generated by default? an on-line investigation into the role of context in generating pragmatic inferences. *Cognition* 100(3):434–463.

Peter Bryant. 2017. *Perception and understanding in young children: An experimental approach*, volume 4. Routledge.

Matthew Capetola. 2013. Towards universal quantification in distributional semantic space. In *Joint Symposium on Semantic Processing (JSSP2013)*. Citeseer, pages 75–79.

Giovanni Cassani. 2014. *Distributional Semantics for Child Directed Speech: A multimodal approach*. Master's thesis, University of Trento.

Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra, and Devi Parikh. 2017. Counting everyday objects in everyday scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zewei Chu, Hai Wang, Kevin Gimpel, and David McAllester. 2016. Broad context language modeling as reading comprehension. *arXiv preprint arXiv:1610.08431* .

Robin Clark. 2011. Generalized quantifiers and number sense. *Philosophy Compass* 6(9):611–621.

Robin Clark and Murray Grossman. 2007. Number sense and quantifier interpretation. *Topoi* 26(1):51–62.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Kenny R Coventry, Angelo Cangelosi, Stephen Newstead, Alison Bacon, and Rohanna Rajapakse. 2005. Grounding natural language quantifiers in visual attention. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum Associates*.

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.

Judith Degen and Noah Goodman. 2014. Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 36.

Judith Degen and Michael K Tanenhaus. 2011. Making inferences: the case of scalar implicature processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 33.

Judith Degen and Michael K Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive science* 39(4):667–710.

Judith Degen and Michael K Tanenhaus. 2016. Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive science* 40(1):172–201.

Stanislas Dehaene. 2003. The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in cognitive sciences* 7(4):145–147.

Stanislas Dehaene and Jean-Pierre Changeux. 1993. Development of elementary numerical abilities: A neuronal model. *Journal of cognitive neuroscience* 5(4):390–407.

Stanislas Dehaene, Véronique Izard, Elizabeth Spelke, and Pierre Pica. 2008. Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science* 320(5880):1217–1220.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pages 248–255.

Isabelle Deschamps, Galit Agmon, Yonatan Loewenstein, and Yosef Grodzinsky. 2015. The processing of polar quantifiers, and numerosity perception. *Cognition* 143:115–128.

Sara Fabbri, Sara Caviola, Joey Tang, Marco Zorzi, and Brian Butterworth. 2012. The role of numerosity in processing nonsymbolic proportions. *The Quarterly Journal of Experimental Psychology* 65(12):2435–2446.

Rema Rossini Favretti, Fabio Tamburini, and Cristiana De Santis. 2002. CORIS/CODIS: A corpus of written italian based on a defined and a dynamic model. *A rainbow of corpora: Corpus linguistics and the languages of the world* pages 27–38.

Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. 2004. Core systems of number. *Trends in cognitive sciences* 8(7):307–314.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* .

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing*. ACL, pages 457–468.

Reza Ghaeini, Sadid A Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Fern, and Oladimeji Farri. 2018. DR-BiLSTM: Dependent Reading Bidirectional LSTM for Natural Language Inference. In *Proceedings of*

the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). volume 1, pages 1460–1469.

Ross Girshick. 2015. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*. pages 1440–1448.

Robert Graves and Alan Hodge. 1943. The reader over your shoulder.

Herbert Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics*, Academic Press, New York, volume 3, pages 41–58.

Justin Halberda and Lisa Feigenson. 2008. Developmental change in the acuity of the "Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental psychology* 44(5):1457.

Justin Halberda, Len Taing, and Jeffrey Lidz. 2008. The development of 'most' comprehension and its potential dependence on counting ability in preschoolers. *Language Learning and Development* 4(2):99–121.

Max Hammerton. 1976. How much is a large part? *Applied ergonomics* 7(1):10–12.

Zellig S Harris. 1954. Distributional structure. *Word* 10(2-3):146–162.

Patrice Hartnett and Rochel Gelman. 1998. Early understandings of numbers: Paths or barriers to the construction of new understandings? *Learning and instruction* 8(4):341–374.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Copenhagen, Denmark, pages 446–456.

Christopher G Healey, Kellogg S Booth, and James T Enns. 1996. High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 3(2):107–135.

Stefan Heim, Katrin Amunts, Dan Drai, Simon B Eickhoff, Sarah Hautvast, and Yosef Grodzinsky. 2012. The language–number interface in the brain: a complex parametric study of quantifiers and quantities. *Frontiers in evolutionary Neuroscience* 4.

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.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*. pages 1693–1701.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016a. The Goldilocks Principle: Reading Children's books with explicit memory representations. In *ICLR 2016*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2016b. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* .

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Keith J Holyoak and Arnold L Glass. 1978. Recognition confusions among quantifiers. *Journal of verbal learning and verbal behavior* 17(3):249–264.

Laurence Horn. 1972. *On the semantic properties of the logical operators in English*. Ph.D. thesis, University of California, Los Angeles.

Laurence Horn. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, form, and use in context: Linguistic applications* pages 11–42.

Yi Ting Huang and Jesse Snedeker. 2009. Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive psychology* 58(3):376–415.

Felicia Hurewitz, Anna Papafragou, Lila Gleitman, and Rochel Gelman. 2006. Asymmetries in the acquisition of numbers and quantifiers. *Language learning and development* 2(2):77–96.

Véronique Izard and Stanislas Dehaene. 2008. Calibrating the mental number line. *Cognition* 106(3):1221–1247.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional

language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pages 1988–1997.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759* .

Roi Cohen Kadosh, Jan Lammertyn, and Veronique Izard. 2008. Are numbers special? An overview of chronometric, neuroimaging, developmental and comparative studies of magnitude representation. *Progress in neurobiology* 84(2):132–147.

Napoleon Katsos, Chris Cummins, Maria-José Ezeizabarrena, Anna Gavarró, Jelena Kuvač Kraljević, Gordana Hrzica, Kleanthes K Grohmann, Athina Skordi, Kristine Jensen de López, Lone Sundahl, et al. 2016. Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences* 113(33):9244–9249.

Edward L Keenan and Denis Paperno. 2012. *Handbook of quantifiers in natural language*, volume 90. Springer.

Edward L Keenan and Jonathan Stavi. 1986. A semantic characterization of natural language determiners. *Linguistics and philosophy* 9(3):253–326.

Roozbeh Kiani, Hossein Esteky, Koorosh Mirpour, and Keiji Tanaka. 2007. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of neurophysiology* 97(6):4296–4309.

Manfred Krifka. 1995. The semantics and pragmatics of polarity items. *Linguistic analysis* 25(3-4):209–257.

Joseph B Kruskal and Myron Wish. 1978. Quantitative applications in the social sciences: Multidimensional scaling (vol. 11). Beverly Hills.

Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104(2):211.

Mathieu Le Corre and Susan Carey. 2007. One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition* 105(2):395–438.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521(7553):436.

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Mike Lewis and Mark Steedman. 2013. Combined distributional and logical semantics. *Transactions of the Association of Computational Linguistics* 1:179–192.

Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. In *European Conference on Computer Vision*. Springer, pages 614–629.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014*, Springer, pages 740–755.

Per Lindström. 1966. First Order Predicate Logic with Generalized Quantifiers. *Theoria* 32(3):186–195.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*. pages 1–9.

Percival G Matthews, Mark Rose Lewis, and Edward M Hubbard. 2016. Individual differences in nonsymbolic ratio processing predict symbolic math performance. *Psychological science* 27(2):191–202.

Gareth McCray and Tineke Brunfaut. 2018. Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing* 35(1):51–73.

Koleen McCrink and Karen Wynn. 2004. Large-number addition and subtraction by 9-month-old infants. *Psychological Science* 15(11):776–781.

Corey T McMillan, Robin Clark, Peachie Moore, Christian Devita, and Murray Grossman. 2005. Neural basis for generalized quantifier comprehension. *Neuropsychologia* 43(12):1729–1737.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

George A Miller and Christiane Fellbaum. 1991. Semantic networks of English. *Cognition* 41(1):197–229.

Utako Minai. 2006. *Everyone knows, therefore every child knows: An investigation of logico-semantic competence in child language*. Ph.D. thesis, University of Maryland.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 3994–4003.

Richard Montague. 1973. The proper treatment of quantification in ordinary English. In *Philosophy, language, and artificial intelligence*, Springer, pages 141–162.

Ruggero Montalto, Angeliek Van Hout, and Petra Hendriks. 2010. Comparing children's and adults' interpretation of Italian indefinite quantifiers. *Linguistics in Amsterdam* 3(2):1–19.

Brianna Morgan, Rachel G Gross, Robin Clark, Michael Dreyfuss, Ashley Boller, Emily Camp, Tsao-Wei Liang, Brian Avants, Corey T McMillan, and Murray Grossman. 2011. Some is not enough: Quantifier comprehension in corticobasal syndrome and behavioral variant frontotemporal dementia. *Neuropsychologia* 49(13):3532–3541.

Joan Moss and Robbie Case. 1999. Developing children's understanding of the rational numbers: A new model and an experimental curriculum. *Journal for research in mathematics education* pages 122–147.

Andrzej Mostowski. 1957. On a generalization of quantifiers. *Fundamenta mathematicae* 44:12–36.

Linda M Moxey and Anthony J Sanford. 1993a. *Communicating Quantities. A psychological perspective*. Lawrence Erlbaum Associates Publishers.

Linda M Moxey and Anthony J Sanford. 1993b. Prior expectation and the interpretation of natural language quantifiers. *European Journal of Cognitive Psychology* 5(1):73–91.

Linda M Moxey and Anthony J Sanford. 2000. Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology* 14(3):237–255.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The RepEval 2017 Shared Task: Multi-Genre Natural Language Inference with Sentence Representations. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. pages 1–10.

Stephen E Newstead and Janet M Collis. 1987. Context and the interpretation of quantifiers of frequency. *Ergonomics* 30(10):1447–1462.

Stephen E Newstead and Kenny R Coventry. 2000. The role of context and functionality in the interpretation of quantifiers. *European Journal of Cognitive Psychology* 12(2):243–259.

Stephen E Newstead, Paul Pollard, and D Riezebos. 1987. The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics* 18(3):178–182.

Andreas Nieder. 2016. The neuronal code for number. *Nature Reviews Neuroscience* .

Andreas Nieder and Earl K Miller. 2003. Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron* 37(1):149–157.

Rick Nouwen. 2010. What's in a quantifier? In Martin Everaert, Tom Lentz, Hannah de Mulder, Øystein Nilsen, and Arjen Zondervan, editors, *The Linguistic Enterprise. Linguistik Aktuell 150*, John Benjamins Publishing Company, pages 235–256.

Mike Oaksford, Lisa Roberts, and Nick Chater. 2002. Relative informativeness of quantifiers used in syllogistic reasoning. *Memory & cognition* 30(1):138–149.

Darko Odic, Paul Pietroski, Tim Hunter, Jeffrey Lidz, and Justin Halberda. 2013. Young children's understanding of "more" and discrimination of number and surface area. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39(2):451.

John W Oller. 1973. Cloze tests of second language proficiency and what they measure. *Language learning* 23(1):105–118.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. *arXiv preprint arXiv:1608.05457* .

Anna Papafragou and Naomi Schwarz. 2006. Most wanted. *Language Acquisition* 13(3):207–251.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of ACL 2016*.

Barbara H Partee. 1989. Many quantifiers. In *Proceedings of the 5th Eastern States Conference on Linguistics*. volume 5, pages 383–402.

Barbara H Partee. 2008. *Compositionality in Formal Semantics*, Blackwell Pusblishing Ltd, chapter Many Quantifiers.

Kevin B Paterson, Ruth Filik, and Linda M Moxey. 2009. Quantifiers and discourse processing. *Language and Linguistics Compass* 3(6):1390–1402.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543.

Stanley Peters and Dag Westerståhl. 2006. *Quantifiers in Language and Logic*. Clarendon Press, Oxford.

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

Sandro Pezzelle, Raffaella Bernardi, and Manuela Piazza. 2018. Probing the mental representation of quantifiers. *Cognition* 181:117–126.

Sandro Pezzelle, Marco Marelli, and Raffaella Bernardi. 2017. Be Precise or Fuzzy: Learning the Meaning of Cardinals and Quantifiers from Vision. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 337–342.

Manuela Piazza. 2010. Neurocognitive start-up tools for symbolic number representations. *Trends in cognitive sciences* 14(12):542–551.

Manuela Piazza and Evelyn Eger. 2016. Neural foundations and functional specificity of number representations. *Neuropsychologia* 83:257–273.

Manuela Piazza, Antonia Fumarola, Alessandro Chinello, and David Melcher. 2011. Subitizing reflects visuo-spatial object individuation capacity. *Cognition* 121(1):147–153.

Paul Pietroski, Jeffrey Lidz, Tim Hunter, and Justin Halberda. 2009. The meaning of 'most': Semantics, numerosity and psychology. *Mind & Language* 24(5):554–585.

Colin Raffel and Daniel P. W. Ellis. 2016. Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems. In *International Conference of Learning Representations*.

Henry Railo, Veli-Matti Karhu, Jeremy Mast, Henri Pesonen, and Mika Koivisto. 2016. Rapid and accurate processing of multiple objects in briefly presented scenes. *Journal of vision* 16(3):8–8.

Rohana K Rajapakse, Angelo Cangelosi, Kenny R Coventry, Steve Newstead, and Alison Bacon. 2005. Grounding linguistic quantifiers in perception: Experiments on numerosity judgments. In *2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*.

Earl F Rankin and Susan Thomas. 1980. Contextual constraints and the construct validity of the cloze procedure. *Perspectives on reading: Research and instruction* pages 47–55.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in neural information processing systems*. pages 2953–2961.

Susannah K Revkin, Manuela Piazza, Véronique Izard, Laurent Cohen, and Stanislas Dehaene. 2008. Does subitizing reflect numerical estimation? *Psychological science* 19(6):607–614.

David A Routh. 1994. On representations of quantifiers. *Journal of Semantics* 11(3):199–214.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* .

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Barbara W Sarnecka and Susan A Gelman. 2004. Six does not just mean a lot: Preschoolers see number words as specific. *Cognition* 92(3):329–352.

Anthea Schöller and Michael Franke. 2017. Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of few & many. *Linguistics Vanguard* 3(1).

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Santi Seguí, Oriol Pujol, and Jordi Vitria. 2015. Learning to count with deep object features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pages 90–96.

Hong Shen and Anoop Sarkar. 2005. Voting between multiple data representations for text chunking. In *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, pages 389–400.

Sailee Shikhare, Stefan Heim, Elise Klein, Stefan Huber, and Klaus Willmes. 2015. Processing of numerical and proportional quantifiers. *Cognitive science* 39(7):1504–1536.

Pooja G Sidney, Clarissa A Thompson, Percival G Matthews, and Edward M Hubbard. 2017. From continuous magnitudes to symbolic numbers: The centrality of ratio. *Behavioral and Brain Sciences* 40.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Vishwanath A Sindagi and Vishal M Patel. 2017. CNN-Based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*. IEEE, pages 1–6.

Frank Smith. 1971. *Understanding reading: A psycholinguistic analysis of reading and learning to read.*. Holt, Rinehart & Winston.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. volume 2, pages 231–235.

Stephanie Solt. 2009. *The semantics of adjectives of quantity*. City University of New York.

Stephanie Solt. 2016. On Measurement and Quantification: The Case of most and more than half. *Language* 92:65–100.

Catherine Sophian. 2000. Perceptions of proportionality in young children: matching spatial ratios. *Cognition* 75(2):145 – 170.

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.

Ionut Sorodoc, Sandro Pezzelle, Aurélie Herbelot, Mariella Dimiccoli, and Raffaella Bernardi. 2018. Learning quantification from images: A structured neural architecture. *Natural Language Engineering* page 130.

Alina G Spinillo and Peter Bryant. 1991. Children's proportional judgments: The importance of half. *Child Development* 62(3):427–440.

Mark Steyvers, Richard M Shiffrin, and Douglas L Nelson. 2004. Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer* pages 237–249.

Ivilin Stoianov and Marco Zorzi. 2012. Emergence of a 'visual number sense' in hierarchical generative models. *Nature neuroscience* 15(2):194–196.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *55th Annual Meeting of the Association for Computational Linguistics, ACL*.

Maojin Sun, Yan Wang, Teng Li, Jing Lv, and Jun Wu. 2017. Vehicle counting in crowded scenes with multi-channel and multi-task convolutional neural networks. *Journal of Visual Communication and Image Representation* 49:412–419.

Anna Szabolcsi. 2010. *Quantification (Research Surveys in Linguistics)*. Cambridge University Press.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 2818–2826.

Jakub Szymanik. 2016. *Quantifiers and Cognition. Logical and Computational Perspectives*. Studies in Linguistics and Philosophy. Springer.

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

Anne Treisman. 2006. How the deployment of attention determines what we see. *Visual Cognition* 14(4-8):411–443. PMID: 17387378.

Vanessa Troiani, Jonathan E Peelle, Robin Clark, and Murray Grossman. 2009. Is it logical to count on quantifiers? Dissociable neural networks underlying numerical and logical quantifiers. *Neuropsychologia* 47(1):104–111.

David Tuggy. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics (includes Cognitive Linguistic Bibliography)* 4(3):273–290.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37:141–188.

Johannes F A K van Benthem. 1986. *Essays in logical semantics*. Springer.

Kees Van Deemter. 2012. *Not exactly: In praise of vagueness*. Oxford University Press.

Bob van Tiel, Michael Franke, Rasmus Båth, and Uli Sauerland. in preparation. Speaking of quantifiers. *Zentrum für Allgemeine Sprachwissenschaft* .

Andrea Vedaldi and Karel Lenc. 2015. *MatConvNet – Convolutional Neural Networks for MATLAB*. Proceeding of the ACM Int. Conf. on Multimedia.

Eric-Jan Wagenmakers and Simon Farrell. 2004. AIC model selection using Akaike weights. *Psychonomic bulletin & review* 11(1):192–196.

Wei Wei, Chuansheng Chen, Tao Yang, Han Zhang, and Xinlin Zhou. 2014. Dissociated neural correlates of quantity processing of quantifiers, numbers, and numerosities. *Human brain mapping* 35(2):444–454.

Dag Westerståhl. 1985. Determiners and context sets. *Generalized quantifiers in natural language* 1:45–71.

Karen Wynn. 1992. Children's acquisition of the number words and the counting system. *Cognitive psychology* 24(2):220–251.

Fei Xu and Elizabeth S Spelke. 2000. Large number discrimination in 6-month-old infants. *Cognition* 74(1):B1–B11.

Ying Yang, Qingfen Hu, Di Wu, and Shuqi Yang. 2015. Childrens and adults automatic processing of proportion in a Stroop-like task. *International Journal of Behavioral Development* 39(2):97–104.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of NAACL-HLT 2016*. pages 1480–1489.

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

Ilker Yildirim, Judith Degen, Michael K Tanenhaus, and T Florian Jaeger. 2016. Talker-specificity and adaptation in quantifier interpretation. *Journal of memory and language* 87:128–143.

Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. 2015. Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 676–684.

Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. 2015a. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 833–841.

Jianming Zhang, Shuga Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price, and Radomír Měch. 2016. Salient Object Subitizing. *arXiv preprint arXiv:1607.07525* .

Jianming Zhang, Shugao Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. 2015b. Salient object subitizing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 4045–4054.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*. Springer, pages 94–108.