Linguistic issues behind visual question answering

Bernardi, R.; Pezzelle, S.

DOI
10.1111/lnc3.12417

Publication date
2021

Document Version
Final published version

Published in
Language and Linguistics Compass

Received: 1 December 2020 | Revised: 7 May 2021 | Accepted: 10 May 2021

DOI: 10.1111/lnc3.12417

ARTICLE

WILEY

Linguistic issues behind visual question answering

Raffaella Bernardi[1] | **Sandro Pezzelle**[2]

[1]CIMeC and DISI, University of Trento, Trento, Italy

[2]ILLC, University of Amsterdam, Amsterdam, The Netherlands

Correspondence
Raffaella Bernardi, CIMeC and DISI, University of Trento, Rovereto, Trento, Italy.
Email: raffaella.bernardi@unitn.it

Funding information
European Research Council, Grant/Award Number: 819455

Abstract

Answering a question that is *grounded* in an image is a crucial ability that requires understanding the question, the visual context, and their interaction at many linguistic levels: among others, semantics, syntax and pragmatics. As such, visually-grounded questions have long been of interest to theoretical linguists and cognitive scientists. Moreover, they have inspired the first attempts to computationally model natural language understanding, where pioneering systems were faced with the highly challenging task—still unsolved—of jointly dealing with syntax, semantics and inference whilst understanding a visual context. Boosted by impressive advancements in machine learning, the task of answering visually-grounded questions has experienced a renewed interest in recent years, to the point of becoming a research sub-field at the intersection of computational linguistics and computer vision. In this paper, we review current approaches to the problem which encompass the development of datasets, models and frameworks. We conduct our investigation from the perspective of the theoretical linguists; we extract from pioneering computational linguistic work a list of *desiderata* that we use to review current computational achievements. We acknowledge that impressive progress has been made to reconcile the engineering with the

theoretical view. At the same time, we claim that further research is needed to get to a unified approach which jointly encompasses all the underlying linguistic problems. We conclude the paper by sharing our own desiderata for the future.

1 | INTRODUCTION

Anyone interested in studying language has to deal with a core aspect of it: 'meaning'. If one wants to understand how meaning is acquired by children or how it can be interpreted by a computer, the question of how it can be represented arises and, with it, sooner or later the importance of *grounding* meaning representations into the visual context pops up. From Quine (1960) to Barsalou (2008), convincing arguments have been made to highlight the importance of developing models able to understand what a word (a symbol) refers to, namely models that account for the symbol grounding problem (Harnad, 1990; Searle, 1980).

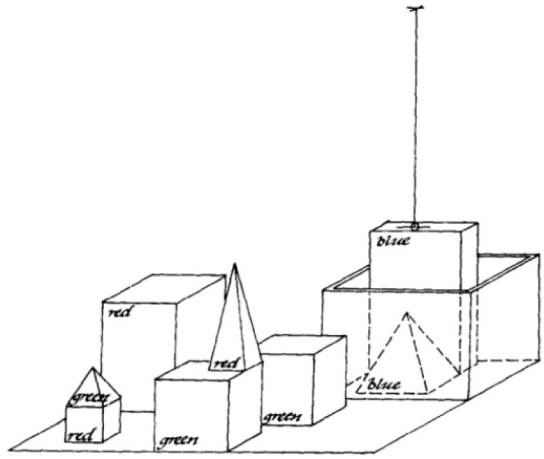
> '*What is the representation of a zebra? It is just the symbol string "horse & stripes". But because "horse" and "stripes" are grounded in their respective iconic and categorical representations, "zebra" inherits the grounding, through its grounded symbolic representation. In principle, someone who had never seen a zebra (but had seen and learned to identify horses and stripes) could identify a zebra on first acquaintance armed with this symbolic representation alone (plus the nonsymbolic – iconic and categorical – representations of horses and stripes that ground it)*' (Harnad, 1990, p. 343).

Through the years, various proposals have been made to tackle this challenge. Based on different frameworks, they link computational models of language with computational models of vision (Baroni, 2016; Bisk et al., 2020; Jackendoff, 1987; Landauer & Dumais, 1997; Silberer et al., 2017).

Another well-established core claim about language shared across disciplines is that language is a process of communication (Carpenter et al., 1998; Fazly et al., 2010; Winograd, 1972).

> '*[S]ocial processes [...] make language acquisition possible by creating a shared referential framework within which the child may experientially ground the language used by adults*' (Carpenter et al., 1998, p. 24).

A crucial role in interaction is played by question-answer exchanges. Questions have attracted the interest of theoretical linguists who have studied, for instance, how their syntactic structure helps build their meaning (for an overview see, e.g., Borsley & Müller, 2019), or formal semanticists who have studied the role played by the answer to build the truth functional meaning of the question (for an overview, see Groenendijk & Stokhof, 1997; Wisniewski, 2015). More recently, empirical studies have been carried out on how the comprehension of questions emerges, showing that children learn to understand *wh-* questions before learning information oriented polar questions. The motivation has the root in the assumption that question understanding emerges as a consequence of interactive learning (Moradlou et al., 2021).



Is there anything which is bigger than every pyramid but is not as wide as the thing that supports it?

FIGURE 1 One visually-grounded question in Winograd (1972). To answer it, the system has to handle reasoning abilities and deal with language ambiguity, vagueness, negation and pragmatics, viz., the list of desiderata we extract from Winograd's detailed dialogue sample

The importance of combining these two main challenges—modelling of symbolic grounding and communication exchanges—was acknowledged by one of the first computational systems about natural language understanding, which focused on visually-grounded dialogues. Winograd (1972) introduced a system that 'answers questions, executes commands, and accepts information in an interactive English dialog' (Winograd, 1972, p. 1). Crucially, such questions are about a visual scene, illustrated in Figure 1: it contains a table on which there are several boxes and pyramids; a person gives instructions, related to such scene, to a robot which has to execute them (e.g., 'pick up a big red block'). The step of obtaining a representation of the visual input was put aside—the system was fed with a pre-compiled symbolic representation of the scene—with the focus being on language understanding. Winograd provided a detailed dialogue sample to discuss the various functionalities such a system must simultaneously deal with at various linguistics levels, namely syntax, semantics and inference. The system had to be able to deal with questions containing an anaphoric expression; to draw inference beyond the question necessary to give the answer; to ask to clarify ambiguity. It was expected to be able to understand when it did not understand the question; when it did not know the answer; when the question was non-sense. Furthermore, questions were grounded into the scene as well as in the language context, hence they had to be interpreted and answered based on the previous dialogue history. For instance, to answer the question given as example in Figure 1, the system had to reason on relations between sets of objects ('anything *bigger than* every pyramid'), interpret negation ('but is *not* as wide') and resolve the anaphora ('that supports *it*').

From Winograd's sample dialogue, we extract a list of main linguistic phenomena (ambiguity, vagueness, negation) and skills (reasoning and pragmatic-based interpretation) that we believe a multimodal system should be able to model. We will refer to this list as our *desiderata*, that we use to review recent achievements in a specific subtask tackled by Winograd's system, namely answering visually grounded questions.

Thanks to efforts within computational linguistics and computer vision, Visual Question Answering (VQA) has become a widely studied task, and important progress has been made on

the development of computational multimodal models. VQA has been treated both as a downstream task and as a pre-training task to effectively encode multimodal input and trasfer it to other multimodal tasks. In this paper, we will discuss both uses of it. By reviewing how current models handle each of our desiderata, we will highlight where further research is needed to turn VQA from an in-lab exercise to a real-life application, and point to what we think is feasible to achieve in the short and medium term.

2 | THE RECENT REVIVAL OF VQA

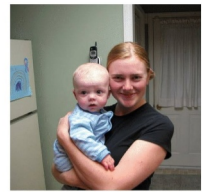
In the last years, there has been a steep increase of interest in the task of answering visually grounded questions. This revival was motivated by the development of models to assist visually impaired people (Bigham et al., 2010) or the attempt to establish a Turing Test based on visual information (Malinowski & Fritz, 2014). This pioneering work was immediately followed by a vigorous worldwide effort aimed at building new datasets and models (Antol et al., 2015; Gao et al., 2015; Geman et al., 2015; Goyal et al., 2016, 2017; Malinowski et al., 2015; M. Ren, Kiros et al., 2015; Yu et al., 2015). This effort has been exhaustively summarized in various surveys (Kafle & Kanan, 2017b; Manmadhan & Kovoor, 2020; Srivastava et al., 2021; Wu et al., 2017), as well as tutorials (Kordjamshidi et al., 2020; Teney et al., 2017).[1] In particular, Srivastava et al. (2021) nicely sketch the timeline of the major breakthroughs in VQA in the last five years, whilst Wu et al. (2017) provide interesting connections with structured knowledge base and an in-depth description of the question/answer pairs present in VQA datasets. Finally, Kafle and Kanan (2017b) discuss shortcomings of current VQA datasets.

2.1 | The VQA task

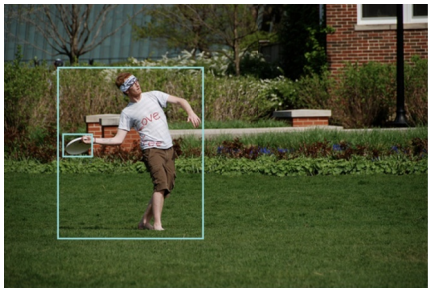
Since 2015, the VQA challenge is organised yearly. Thanks to it, progress in the field can be constantly monitored.[2] The original dataset (VQA v1.0) consisted of images taken from the Microsoft Common Objects in Context (MS-COCO) dataset (T. Y. Lin et al., 2014) and questions collected from human annotators via crowdsourcing. As we will discuss later, the baseline model relied on coarse multimodal representations obtained by performing simple operations on the language and visual representations. The original dataset was shown to contain heavy biases that models could easily exploit to perform the task (B. Zhou et al., 2015). Since then, quite some attention has been paid to the language bias issue. In particular, a new dataset has been released (VQA v2.0; Goyal et al., 2017) in which each question is paired with very similar images. Figure 2 illustrates the difference compared to the previous version of the VQA dataset, a change that requires *finer-grained representations* as it was advocated by, for example, Shekhar et al. (2017) and J. Wang et al. (2018).

Since the original work on VQA, a careful analysis of model results has been carried out to go beyond an evaluation solely based on task success (see, for instance, Agrawal et al., 2016; X. Lin & Parikh, 2015; Zhu et al., 2016). Driven by the goal of gaining a deep understanding of the multimodal behaviour, Agrawal et al. (2018) reorganised the VQA dataset to assess the robustness of models when exposed to different question biases at test time compared to what is seen during training. From these analyses, it turned out that questions involving *reasoning about relations between objects*, such as, for instance, those involving role labelling and spatial relations, are the hardest to be answered. To help making progress on questions involving role

VQA v1.0: What is the mustache made of?



VQA v2.0: Where is the child sitting?



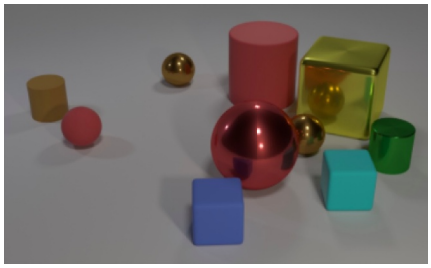
Visual Genome: What is the man holding?

GQA: Are the napkin and the cup the same color?

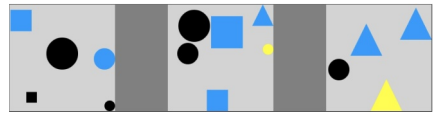
FIGURE 2 Datasets of natural images: The task of answering a question about an image has been promoted by the release of datasets containing an image and a question about it, such as VQA v1.0 (Antol et al., 2015). By controlling for the multimodal data points, models have been pushed to build *finer-grained representations* (see VQA v2.0; Goyal et al., 2017). The release of densely annotated datasets, such as Visual Genome (Krishna et al., 2017), made it possible to tackle the challenge of building multimodal representations of *relations between objects*. This paved the way to resources, such as GQA (Hudson & Manning, 2019), which include compositional questions involving such relations

labelling, Yatskar et al. (2016) released ImSitu, a dataset containing annotation about actions, roles and objects. In parallel, datasets such as Visual7W (Zhu et al., 2016), TDIUC (Kafle & Kanan, 2017a), Visual Genome (Krishna et al., 2017) and GQA (Hudson & Manning, 2019) have been developed to test the visual reasoning and compositionality abilities of models. Figure 2 (bottom) illustrates the Visual Genome and GQA datasets by means of example. The main novelty of Visual Genome is the high-density annotation of its images and the fine-grained alignment between images and language descriptions. The GQA dataset was carefully designed building on such annotation (Hudson & Manning, 2019). By adopting a *diagnostic* approach, they paired natural images of Visual Genome with automatically-generated questions to enable a fine-grained diagnosis for different question types. Furthermore, they introduced new metrics aimed to evaluate models with respect to consistency, plausibility and grounding. Finally, both for VQA and GQA new out-of-domain test sets have been proposed to allow a more reliable evaluation of the models (Gokhale et al., 2020a; Kervadec et al., 2021).

The diagnostic approach has been undertaken also in other work proposing datasets of synthetic images coupled with either templated (Andreas et al., 2016; Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017; Kuhnle & Copestake, 2017; Sorodoc et al., 2016; Zhang



CLEVR: Are there an equal number of large things and metal spheres?



NLVR: There is a box with 2 triangles of same color nearly touching each other

FIGURE 3 Datasets of synthetic images: Benchmarks like CLEVR (Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017) and NLVR (Suhr et al., 2017) require models capture the relations between the objects depicted in a synthetic scene. In CLEVR, relations involve objects depicted in a single scene; in NLVR, they span over three *boxes*. Language is synthetically generated in CLEVR and crowd sourced in NLVR

et al., 2016) or crowd sourced (Suhr et al., 2017) language. Figure 3 illustrates samples from CLEVR (Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017) and NLVR (Suhr et al., 2017). CLEVR images are paired with questions generated through functional programs; the data points are carefully designed to test the skills a model needs to master to answer attribute, existential and counting questions, as well as questions based on comparisons and spatial relationships. NLVR (Suhr et al., 2017) is based, instead, on a verification task: models have to answer whether a given sentence is true or false within the given visual context. This is the same setting of NLVR2 (Suhr et al., 2019), which uses natural images. In Section 3, we will come back to the role of diagnostic datasets to evaluate the reasoning abilities of multimodal models.

From the very beginning of this VQA revival, attention has been paid also to questions that require information available in a Knowledge Base to be answered (Wang et al., 2017a). This line of research has been pursued by several studies, particularly thanks to the introduction of the Fact-based VQA dataset (Wang et al., 2017b). The VQA task is now taking new directions, such as embodied approaches where an agent has to navigate an environment and answer questions about it (H. Chen et al., 2019; Das et al., 2018); video VQA, where the answer has to be found in videos rather than in static images (Lei et al., 2018, 2020); answering questions about diagrams and charts (Ebrahimi Kahou et al., 2017; Kafle et al., 2018); text VQA, which involves recognizing and interpreting textual content in images (Biten et al., 2019; Han et al., 2020); answering questions about medical images (see, Abacha et al., 2020); and many others.

2.2 | Multimodal representations

Research on the interplay between language and vision has benefited from the comparable representations developed and used by the computational linguistics and computer vision communities in the last decade or so. On the language side, the distributional semantics approach (Firth, 1957; Harris, 1954) has become the most popular view of natural language semantics: A word is represented by a *vector* (also called *word embedding*) which encodes the contexts in which it occurs (Landauer & Dumais, 1997). In traditional approaches this vector is

static, that is, not dependent on the various senses of a word (see, Mikolov et al., 2013; Pennington et al., 2014), whilst last-generation neural network models, such as Transformers, are able to produce *contextualized* representations whilst processing a linguistic string (see, Devlin et al., 2019). Similarly, a whole image (or each of the objects in it) is represented by a vector computed by a deep neural network, which learns such representation in an end-to-end fashion, viz. starting from the image's pixels whilst being trained on an object classification task (He et al., 2016; S. Ren, He, et al., 2015; Simonyan & Zisserman, 2015).

The availability of word embeddings and visual vectors has facilitated the fertilization between the two communities, that has been further boosted by the availability of multimodal baselines and state-of-the-art models. Earlier approaches obtained multimodal representations by concatenating the linguistic and visual vectors (Bruni et al., 2014) or by taking their inner product (Antol et al., 2015).[3] We are currently experiencing the boom of Transformer-based *universal multimodal encoders* pretrained on several multimodal tasks, and aimed at obtaining task-agnostic multimodal representations (Y.-C. Chen et al., 2020; Li et al., 2019; Lu et al., 2019; Su et al., 2019; H. Tan & Bansal, 2019; L. Zhou et al., 2020).

2.3 | VQA models

2.3.1 | Early models

The most popular VQA baseline model by Antol et al. (2015) learns the word embeddings through the VQA task itself; starting from one-hot-encodings, it builds word embeddings that are incrementally composed by an LSTM (Long Short Term Memory; Hochreiter & Schmidhuber, 1997) to obtain the question representation; for the images, it uses VGGNet image embeddings (Simonyan & Zisserman, 2015), which are further processed by a linear transformation to match the LSTM encoding of the question. These features are then combined using element-wise operations to a common multimodal feature; this is given as input to a softmax classifier to obtain the probability distribution among the candidate answers, and select the one with the highest probability. Building on this early VQA baseline, a plethora of models have been proposed. Since exhaustive overview papers are already available (Kafle & Kanan, 2017b; Manmadhan & Kovoor, 2020; Srivastava et al., 2021; Wu et al., 2017), here we do not review all the approaches and models that have been proposed. Instead, we highlight and explain the major milestones that have been achieved and that we can relate to our desiderata listed in Section 1.

2.3.2 | Attention-based models

The first crucial enhancement has been the use of *attention mechanisms* which have led to build *fine-grained* representation of the multimodal input. One modality guides the interpretation of the other so to give more weight to salient regions of the image or to relevant words of the question (Yang et al., 2016). The promising results obtained with the introduction of these reweighting methods led researchers to propose more complex mechanisms like hierarchical co-attention (Lu et al., 2016), or the combination of bottom-up and top-down mechanisms, an approach that has dominated the scene since its introduction by Anderson et al. (2018). A detailed analysis of the effect of its various parameters is given by Teney et al. (2018). The main

advancement brought by the bottom-up top-down approach lies in the use of attention to focus on the objects in the scene that are most salient to answer the question—rather than on generic (important) regions of the image. This is made possible by the use of Faster R-CNN (S. Ren, He, et al., 2015), which proposes several candidate bounding boxes—each containing one object—to the network. The object identification phase allows the model to exploit bottom-up information regarding objects instead of starting from scratch from the understanding of the entire scene, and informs the top-down component which selects the relevant objects to perform the task.

2.3.3 | Neural module networks

This family of models treats a question as a collection of predefined subproblems (e.g., counting, localization, conjunction, etc.), each handled by a dedicated module. Whilst NMN (Andreas et al., 2016) requires a parser to process the sentence into its components, N2NMN (Hu et al., 2017) does not require any external supervision. Building on these approaches, hybrid methods which combine symbolic and neural components have been recently promoted. Johnson, Hariharan, Van Der Maaten, Hoffman, et al. (2017) claim that models based only on neural representations unavoidably learn dataset biases instead of the visual reasoning skills needed to properly perform complex tasks such as VQA. Hence, they propose a model that represents a question like a program and answers the question by composing functions from a function dictionary. The model learns compositional reasoning from a small amount of the ground-truth programs used in CLEVR (Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017) to generate the questions. The model is shown to generalize to novel questions by composing modules in ways that are not seen during training. This hybrid approach has been pushed forward by Yi et al. (2018), who propose a neural-symbolic VQA approach that disentangles reasoning from visual perception and language understanding, and by Mao et al. (2019), who add a neuro-symbolic concept learner. The hybrid approach does not fall into the bias traps and is easily interpretable—which makes it potentially different from 'black-box' neural networks models.

2.3.4 | From labs to real-life applications

A crucial challenge all these models have to face is the ability to generalize the knowledge learned to unseen data, which can be achieved only if the model is able to compositionally build the multimodal representations, a must for any model of human intelligence (Lake et al., 2017). Since neural-based VQA models have been shown to produce inconsistent answers to questions that are either similar or mutually exclusive, approaches to mitigate this behaviour have been recently proposed (Ray et al., 2019; Selvaraju et al., 2020). Interestingly, Gandhi and Lake (2020) showed that whilst children are driven by the mutual exclusivity assumption in their learning process, neural networks are not, and set this as an open challenge.

All the work we have reviewed so far has paved the way toward incorporating challenging linguistic phenomena into the VQA framework and benchmarks. However, none of them jointly account for the whole range of phenomena encountered in real-life question answering scenarios. Once we move from labs to real-life applications, indeed, additional challenges emerge both at the visual and language level. Models are required to master a variety of language phenomena, such as language ambiguities, pragmatic aspects and context dependence,

negation, entailment, mutual exclusivity and all the reasoning skills subtending them. Some of these extra challenges are present in goal-oriented datasets such as VizWiz (Gurari et al., 2018), which contains pictures taken by visually-impaired people with their mobile phone, the questions they ask about these pictures, and the corresponding answers provided by human assistants via crowdsourcing.

In the following, we highlight what has been achieved (and what has not) of our list of desiderata extracted from Winograd's dialogue sample. By so doing, we also emphasize what we believe deserves further attention from the language and vision community.

3 | REVISITING THE WISHES FROM THE PAST

As we mentioned above, Winograd looked at the challenges for a visually grounded interactive system solely from the perspective of the language modality (the images were assigned pre-compiled symbolic representations). On the other hand, most of the work carried out recently on VQA has been driven by the computer vision community. We are now in the fortunate position to promote a joint view on how the long-standing theoretical questions about grounded language understanding are addressed by computational models. Hence, in what follows, we review where the current visually-grounded research stands with respect to the *desiderata* we extracted from Winograd's dialogue sample.

3.1 | Reasoning

Winograd called for a system that is able to infer from the visual scene the answer to a question of the type: '*Is there anything which is bigger than every pyramid but is not as wide as the thing that supports it?*'. As seen in Section 2, in the recent past the reasoning skills of multimodal models have been studied both by controlling the reasoning steps that a system has to perform to answer VQA questions and by building datasets that are specifically designed for testing these abilities.

As we have mentioned above, several diagnostic datasets have been released with the aim to assess model abilities to reason over a question grounded in a visual context (hence, *visual reasoning*; Andreas et al., 2016; Johnson, Hariharan, van der Maaten, Fei-Fei, et al., 2017; Kuhnle & Copestake, 2017; Suhr et al., 2017, 2019). These works brought a shift from *non-relational* questions, which require reasoning about the attributes of one particular object instance, to *relational* questions (Santoro et al., 2017), which instead require to genuinely reason over the relations between multiple objects depicted in the image. From the computer vision perspective, solving non-relational questions implies locating an object in an image, that is, paying *attention* to the region of the image which 'contains' the object. Relational reasoning problems, in contrast, require models to pay attention on multiple objects in the visual scene, to identify their attributes (colour, size, category, etc.), and to perform a higher-level reasoning step over this perceptual knowledge. If one views language from a denotational semantic perspective, it becomes clear that the move from non-relational to relational skills is also crucial to master language phenomena of increasing complexity, for which yet another step is necessary, namely to deal with questions involving *relations between sets of objects*. For instance, to properly represent quantifiers a model has to identify the sets of relevant objects; similarly, gradable adjectives require a comparison of the set of entities against which they are

NLVR2: The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

GTE: P: A family in front of a chimney and H: A family trying to get warm

FIGURE 4 Reasoning: NLVR2 (Suhr et al., 2019) evaluates the high-level reasoning skill of models: a model has to say whether the given sentence is true or false with respect to the two images; GTE (Vu et al., 2018) evaluates their ability to ground textual entailment: the model has to choose whether the two sentences (a premise P and a hypothesis H) are in an entailment, contradictory or neutral relation, given the image

interpreted; negation of, for example, a noun points to the alternative sets of the negated noun (the set of other candidate objects), etc. Answering to questions involving these expressions is therefore a higher-level problem as compared to first-level relations and non-relational questions described above.

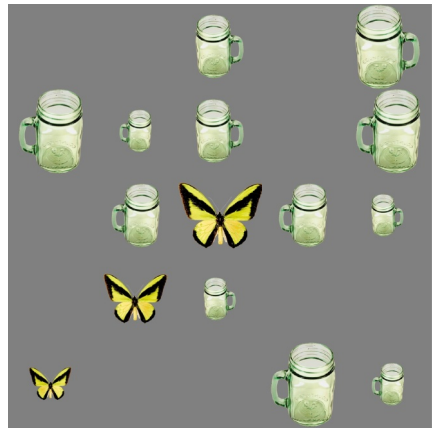
Recently, the reasoning skills of multimodal models have been tested by means of either probing tasks involving high-level reasoning or grounded textual entailment (see Figure 4). In the recent NLVR2 dataset (Suhr et al., 2019), a visual scene comprising two natural images is coupled with a crowdsourced statement describing some relation between the entities depicted in these two images. In order to verify whether the statement is true for that scene, models are required to deal with complex linguistic phenomena such as quantification, negation, coreference and syntactic ambiguity resolution. Whilst handling these phenomena is straightforward for humans (who achieve a virtually perfect accuracy in the task), current state-of-the-art models are shown to struggle with them. Indeed, the gap with human performance is around -20% in this dataset (see Suhr et al., 2019; Zheng et al., 2020). This reveals that a full understanding of complex language phenomena is an ability required for models to deal with real-life multimodal questions.

The reasoning skills of multimodal models have been studied also by directly investigating how they perform on the entailment task. To test these abilities, Vu et al. (2018) proposed a dataset of grounded textual entailment: a model has to say whether two given sentences (a premise and a hypothesis) are in an entailment, contradictory or neutral relation with respect to a given image; whereas Xie et al. (2019) released a visual entailment dataset where models are asked to check whether a given image entails a given text (Figure 5).

3.2 | Language ambiguity

'*Put the blue pyramid on the block in the box*' is one of the instructions the Winograd's system is challenged to handle. The instruction is syntactically ambiguous, but the visual context disambiguates it. Current multimodal models have been evaluated on the ability to acquire and use such 'disambiguation' skills. For example, Christie et al. (2016) addressed the issue of prepositional phrase attachment resolution by training the system to pick, among the possible

Language ambiguity: Sam approached the chair with a bag



Vagueness: Few of the objects are animals

Negation: Are they in a restaurant and are they all not boys?

Pragmatics: A red double-decker bus

FIGURE 5 Benchmarks requiring models to compute higher-level relations, that is, between sets of objects. In Berzak et al. (2015) (top left panel), visual information from videos is used to disambiguate sentences that are ambiguous at many levels, for example, syntactic. Pezzelle et al. (2018) focus on quantifiers and challenge models to learn their vague, context-dependent interpretation from visual scenes (top right). In Gokhale et al. (2020b), a new dataset and computational method is proposed to tackle negation (bottom left). Finally, Cohn-Gordon et al. (2018) force Image Captioning models to be pragmatically informative, that is, to produce captions that are discriminative (bottom right)

interpretations produced by a vision and language model, the one that is consistent between the two modalities.

Along with syntactic ambiguities, Winograd's system is faced with questions and instructions that are ambiguous at the semantic and discourse level since they involve anaphora resolution, for example, '*Is it supported?*' or '*Put a small one into the green cube*'. Berzak et al. (2015) studied ambiguity at the syntactic, semantic and discourse level, and introduced a novel dataset of ambiguous sentences coupled with short videos. Overall, their multimodal model was shown to be able to perform the disambiguation task, with model performance being higher for syntactic compared to either semantic or discourse ambiguities.

Syntax has been shown to be useful also to disambiguate referring expressions and properly locate (ground), in an image, the object to which the expression refers (Cirik et al., 2018). Here, a syntactic analysis of the input referring expression was used to inform the structure of a

computation graph. Moreover, some other work (see Shutova et al., 2016) focused on a special type of semantic ambiguity, metaphors and proposed the task of visually-grounded metaphor detection. Given an adjective-noun phrase such as 'black hole', the task is to understand whether the phrase represents a metaphor or not. Once again, visual information was shown to be useful for the task.

3.3 | Vagueness

'*Is at least one of them narrower than the one which I told you to pick up?*'. To answer this and similar questions, Winograd's system is required to understand quantifiers (*at least one*) and gradable adjectives (here, the comparative form *narrower*). These expressions can be *vague*, that is, their interpretation can depend on the context in which they are used. For example, the applicability of words like *most* or *big* in a certain context depends on the properties of the set of objects that are relevant for their interpretation. Moreover, their interpretation can be *borderline* and therefore differ across human speakers.

Whilst numbers represent a well-known challenging problem in VQA (Acharya et al., 2019; Chattopadhyay et al., 2017), the presence of quantifiers in standard VQA datasets is limited. Though quantification is present in some visual reasoning benchmarks, such as ShapeWorld (Kuhnle & Copestake, 2017), NLVR (Suhr et al., 2017) and NLVR2 (Suhr et al., 2019), these approaches only include numerical or logical quantifiers, for example, *at least two* or *more than half*. In contrast, quantifiers such as *few* or *most*—whose intepretation largely depends on the (visual) context in which they are uttered—are absent. A strand of work has focused on quantifiers combining formal semantics and cognitive science to propose models to perform grounded quantification in a human-like manner (Sorodoc et al., 2018); to assign the correct quantifier to a visual scene (Sorodoc et al., 2016); and to model the use of quantifiers jointly with numbers and proportions (Pezzelle et al., 2017, 2018).

Gradable adjectives have long been studied by formal semanticists interested in understanding how word meaning changes depending on the context in which the word is uttered (Kennedy, 2007; Partee, 1995). However, in standard VQA benchmarks, these expressions are treated as static rather than context-dependent attributes; alternatively, they are present only in their comparative or superlative forms (Kuhnle & Copestake, 2017; Suhr et al., 2017). Recently, Pezzelle and Fernández (2019b) released a novel dataset of synthetically generated images and statements containing the gradable adjectives *big* and *small*, and showed that state-of-the-art visual reasoning models can, to some extent, learn the function underlying their use. However, models were shown to be unable to learn an *abstract* representation of such words that can be compositionally applied to unseen objects (see also Pezzelle & Fernández, 2019a).

3.4 | Negation

Winograd's system should also be able to handle negation in order to answer questions like '*How many blocks are not in the box?*'. Kruszewski et al. (2016) argue that conversational negation does not create the complement set, but rather the alternative set. If we look at this claim from the perspective of visually-grounded negation, this means that its interpretation requires looking at the set of alternative entities in the scene, or even understanding that the reference is not in the image (hence, it is not visually grounded). Nordemeyer and Frank (2014)

show that processing negation can be easier for humans if a visual context creates pragmatic expectation that motivates its use. However, it is unknown whether this holds for multimodal models. van Miltenburg et al. (2016) provide a preliminary corpus study on the use of negation in image captioning (IC) and points the implication to IC models. Suzuki et al. (2019) propose a logic-based visual inference system and evaluate it on the retrieval of images from text including logical operators (negation, quantifiers and numerals). More recently, some interest has been paid in the computer vision community to logical skills of VQA models, particularly negation. Gokhale et al. (2020b), for example, showed that state-of-the-art models struggle to handle such phenomenon, and proposed a method and dataset to tackle this problem. Greco et al. (2021) show that multimodal universal encoders have difficulty in interpreting negatively answered questions.

3.5 | Pragmatics

Winograd's system is also required to use referring expressions that are pragmatically discriminative based on the context in which they are used; for example, *the big red block* if there are other blocks and none else are both big and red. In the language and vision community, pragmatic aspects have been taken into account in the task of IC, where approaches building on Bayesian frameworks have been proposed to generate descriptions that contrastively refer to one but not another (similar) image (Achlioptas et al., 2019; Andreas & Klein, 2016; Cohn-Gordon et al., 2018; Monroe et al., 2017). Similar approaches have been proposed for *zero-shot* referring expression generation (Zarrieß & Schlangen, 2019).

Some recent work investigated the use and interpretation of colour terms in grounded communication contexts. Monroe et al. (2016) focused on the generation of compositional colour descriptions, whilst Monroe et al. (2017) presented a novel corpus of colour descriptions from reference games, and showed that an agent equipped with both a neural listener and speaker component interprets colour descriptions better than the listener alone. More recently, Schüz and Zarrieß (2020) focused on predicting objects' colours and showed that combining categorical with perceptual, entity-based information is the best-performing approach.

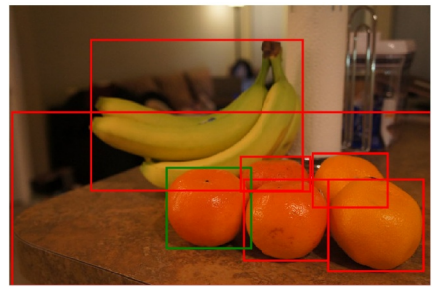
4 | OPEN CHALLENGES AND FUTURE DIRECTIONS

We conclude the survey by touching upon new challenges that we see deserve further attention and could be addressed in the near future.

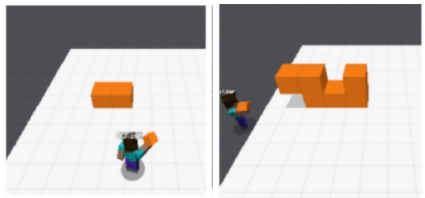
4.1 | Further challenges from computational linguistics

As mentioned above, Windograd's system was designed to ground questions into a visual scene but also based on the dialogue history. The move from a question-answering system to a QA system able to answer follow-up questions (FUQs) has been undertaken by the QA community early on. It was shown to be an interesting case-study in between QA, information retrieval and dialogue systems: users are given the chance to refine their query/question based on the linguistic answer they received (Webb & Webber, 2009). Follow-up visual questions have been

VisDial: 1: What color is the mug? White and red. 2: Are there any pictures on it? No, something is there that I can't tell what it is. 3: Is the mug and cat on a table? Yes, they are.



GuessWhat?!: 1: Is it a fruit? Yes. 2: Is it the orange? Yes. 3: One of them I suppose? Yes. 4: Is it to our right? No. 5: In the middle? No. 6: The last single one? Yes.



Minecraft: B: (puts down 1 orange). A: so it will look like a *v*. B: (puts down 4 orange).

FIGURE 6 Interactive VQA. In VisDial (Das, Kottur, Gupta, et al., 2017), the model has to ground a follow-up question into the linguistic and visual context to answer it. GuessWhat?! (de Vries et al., 2017) requires the model to generate a sequence of Y/N-questions to gather information about the target object to be guessed. Finally, in Minecraft (Jayannavar et al., 2020), a full interaction in the Winograd-style is required

studied for instance in F. Tan et al. (2019), where the system has to retrieve the correct image by receiving a sequence of questions asked by a user.

Multimodal models have been evaluated also on visual dialogue tasks, in which the agent has to answer a FUQ by grounding it on the dialogue history and on the image the question is about. The most popular dataset, VisDial (Das, Kottur, Gupta, et al., 2017), has been used for a yearly organised challenge: a model (the Oracle) has to answer a question about an image given either a caption or a caption together with a sequence of question-answer pairs about the image (see Figure 6). Agarwal et al. (2020) shows that only 11% of the samples in the VisDial dataset need the previous context to be correctly answered. Hence, this research line requires further effort on the collection of datasets containing more challenging dialogue phenomena. Since universal multimodal encoders are available, checking their grounding skills on the relatively small datasets including dialogue history may be a first interesting step.

When opening the box of interaction, the next challenge that pops up immediately is question generation. Task-oriented visual games (de Vries et al., 2017; Das, Kottur, Moura, et al., 2017; Haber et al., 2019; Ilinykh et al., 2019) are a good way to measure the progress in such direction. Figure 6 illustrates the simple dialogues of GuessWhat?! game (de Vries et al., 2017). Task success is taken to be a measure of how well the model has been able to ask

informative questions. However, as shown by some studies (Mazuecos et al., 2020; Shekhar et al., 2019; Testoni, Shekhar, et al., 2019), task-success does not relate to the quality of the dialogue nor to the informativeness of the question generated. More work is needed to develop conversational multimodal models that are able to generate pragmatically sound utterances. Crucially, the community lacks datasets to evaluate such skills. An interesting project that could represent an important contribution towards this aim involves two agents playing the Minecraft visual game (Jayannavar et al., 2020).

Most multimodal conversational models exploit the encoder-decoder architecture (Sutskever et al., 2014): an encoder receives the embeddings of both modalities, it combines them, and uses its hidden state to condition the decoder module to generate the (follow-up) question. We hope that universal decoders that are able to transfer their knowledge to new tasks will be developed. In line with the general claim advocated by Linzen (2020), we hope to see carefully-designed visual dialogue datasets that are useful to give exact diagnoses of the communication skills achieved/not yet achieved by the conversational systems.

4.2 | Further challenges from computer vision

The VQA task has been further extended to QA about videos. The largest-scale Video-QA dataset currently available is TVQA (Lei et al., 2018, 2020), which contains questions about popular TV shows. Besides visual grounding, VQA models are also challenged to deal with the audio modality in the Audio-Visual Scene-aware Dialogues dataset (AVSD; Hori et al., 2019). Finally, the fervent activities we are experiencing these days on interactive QA over images and videos will certainly create a boost towards the interesting goal of developing models that are able to 'predict future events' (Huang et al., 2016; Walker et al., 2014). Humans highly rely on their prediction skills when interpreting a new input, integrating their perceptual signal with prior knowledge. We hope that more awareness of cognitive and neuroscience findings towards the combination of bottom-up (perceptual) and top-down (prior) knowledge will help shaping new multimodal models (Schüz & Zarrieß, 2020; Suglia et al., 2020; Testoni, Pezzelle et al., 2019).

5 | CONCLUSION

Reviewing the literature on VQA is, by itself, a stimulating activity since progress is tangible and fast. We hope that this paper will contribute to promote further work and collaboration between experts in the language and vision community, which we have shown to be crucial for the development of fully fledged multimodal models. We agree with the call for more research on contextual language learning promoted by Bisk et al. (2020), and for the importance of developing a vision and language decathlon benchmark to measure holistic progress advocated by Kafle et al. (2019) (for a first step towards this goal, see Parcalabescu et al., 2020). We furthermore call for more awareness of neuroscience findings on how human brain processes these two modalities and on settings in which the two modalities convey complementary, rather than aligned, information (Pezzelle et al., 2020).

ACKNOWLEDGEMENTS

Sandro is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455 awarded to

Raquel Fernández). We gratefully thank Yoav Artzi, Evan Defrancesco, Stella Frank and Kushal Kafle for their useful feedback on earlier versions of the paper.

ORCID

Raffaella Bernardi https://orcid.org/0000-0002-3423-1208
Sandro Pezzelle https://orcid.org/0000-0002-3969-7445

ENDNOTES

[1] An updated list of papers can be found here: https://github.com/jokieleung/awesome-visual-question-answering
[2] https://visualqa.org/
[3] See Elliott et al. (2016) for an overview.

REFERENCES

Abacha, A. B., Datla, V. V., Hasan, S. A., Demner-Fushman, D., & Müller, H. (2020). Overview of the VQA-Med task at imageCLEF 2020: Visual question answering and generation in the medical domain. *CLEF 2020 Working Notes*, 22–25.

Acharya, M., Kafle, K., & Kanan, C. (2019). TallyQA: Answering complex counting questions *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, pp. 8076–8084).

Achlioptas, P., Fan, J., Hawkins, R., Goodman, N., & Guibas, L. J. (2019). ShapeGlot: Learning language for shape differentiation. In *Proceedings of the IEEE international conference on computer vision* (pp. 8938–8947).

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1955–1960).

Agrawal, A., Batra, D., Parikh, D., & Kembhavi, A. (2018). Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4971–4980).

Agarwal, S., Bui, T., Lee, J.-Y., Konstas, I., & Rieser, V. (2020). History for visual dialog: Do we really need it? In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8182–8197). Association for Computational Linguistics. Online.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6077–6086).

Andreas, J., & Klein, D. (2016). Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1173–1182).

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016). Neural module networks. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 39–48).

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425–2433).

Baroni, M. (2016). Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, *10*(1), 3–13.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–45 .

Berzak, Y., Barbu, A., Harari, D., Katz, B., & Ullman, S. (2015). Do you see what I mean? Visual resolution of linguistic ambiguities. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1477–1487).

Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., & Yeh, T. (2010). VizWiz: Nearly real-time answers to visual questions. In *ACM user interface software and technology symposium (UIST)*.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience grounds language. In *Proceedings of the 2020*

conference on empirical methods in natural language processing (EMNLP) (pp. 8718–8735). Association for Computational Linguistics. Online.

Biten, A. F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., & Karatzas, D. (2019). Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4291–4301).

Borsley, R. D., & Müller, S. (2019). *Head-driven phrase structure grammar: The handbook*, chapter HPSG and Minimalism. Language Science Press.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47.

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63*(4), 174 .

Chattopadhyay, P., Vedantam, R., Selvaraju, R. R., Batra, D., & Parikh, D. (2017). Counting everyday objects in everyday scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1135–1144).

Chen, H., Suhr, A., Misra, D., Snavely, N., & Artzi, Y. (2019). Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 12530–12539).

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). UNITER: Universal image-text representation learning. In *European conference on computer vision* (pp. 104–120). Springer.

Christie, G., Laddha, A., Agrawal, A., Antol, S., Goyal, Y., Kochersberger, K., & Batra, D. (2016). Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1493–1503). Association for Computational Linguistics.

Cirik, V., Berg-Kirkpatrick, T., & Morency, L.-P. (2018). Using syntax to ground referring expressions in natural images *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32).

Cohn-Gordon, R., Goodman, N., & Potts, C. (2018). Pragmatically informative image captioning with character-level inference *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies* (Vol. 2, pp. 439–443). (Short Papers).

Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018). Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 2054–2063).

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., & Batra, D. (2017a). Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 326–335).

Das, A., Kottur, S., Moura, J. M., Lee, S., & Batra, D. (2017b). Learning cooperative visual dialog agents with deep reinforcement learning. In *2017 IEEE international conference on computer vision* (pp. 2951–2960).

de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. C. (2017). GuessWhat?! Visual object discovery through multi-modal dialogue. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 5503–5512).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. (Long and Short Papers).

Ebrahimi Kahou, S., Atkinson, A., Michalski, V., Kadar, A., Trischler, A., & Bengio, Y. (2017). FigureQA: An annotated figure dataset for visual reasoning. In *Visually grounded interaction and language workshop, NIPS 2017*.

Elliott, D., Kiela, D., & Lazaridou, A. (2016). Multimodal learning and reasoning. In *Proceedings of the 54th annual meeting of the association for computational linguistics: Tutorial abstracts*. Association for Computational Linguistics.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*, 1017–1063.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. Oxford University Press.

Gandhi, K., & Lake, B. M. (2020). Mutual exclusivity as a challenge for deep neural networks. *Advances in Neural Information Processing Systems*, *33*.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. (2015). Are you talking to a machine? Dataset and methods for multilingual image question. *Advances in Neural Information Processing Systems*, *28*, 2296–2304.

Geman, D., Geman, S., Hallonquist, N., & Younes, L. (2015). Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(12), 3618–3623.

Gokhale, T., Banerjee, P., Baral, C., & Yang, Y. (2020a). MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 878–892). Association for Computational Linguistics. Online.

Gokhale, T., Banerjee, P., Baral, C., & Yang, Y. (2020b). VQA-LOL: Visual question answering under the lens of logic. In *European conference on computer vision* (pp. 379–396). Springer.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6904–6913).

Goyal, Y., Mohapatra, A., Parikh, D., & Batra, D. (2016). Towards transparent AI systems: Interpreting visual question answering models. In *Proceedings of ICML visualization workshop*.

Greco, C., Testoni, A., & Bernardi, R. (2021). "Yes" and "No": Visually grounded polar answers. In *Proceedings of visually grounded interaction and language (ViGIL)*. NAACL.

Groenendijk, J., & Stokhof, M. (1997). *Handbook of logic and linguistics*, chapter Questions. North Holland.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., & Bigham, J. P. (2018). VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3608–3617).

Haber, J., Baumgärtner, T., Takmaz, E., Gelderloos, L., Bruni, E., & Fernández, R. (2019). The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1895–1910).

Han, W., Huang, H., & Han, T. (2020). Finding the evidence: Localization-aware answer prediction for text visual question answering. In *Proceedings of the 28th international conference on computational linguistics* (pp. 3118–3131).

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346.

Harris, Z. (1954). Distributional structure. *Word*, *10*(2-3), 1456–1162.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., Marks, T. K., Cartillier, V., Lopes, R. G., Das, A., Essa, I., Batra, D., & Parikh, D. (2019). End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2352–2356). IEEE.

Hu, R., Andreas, J., Rohrbach, M., Darrell, T., & Saenko, K. (2017). Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 804–813).

Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., & Mitchell, M. (2016). Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 1233–1239). Association for Computational Linguistics.

Hudson, D. A., & Manning, C. D. (2019). GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6700–6709).

Ilinykh, N., Zarrieß, S., & Schlangen, D. (2019). Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th international conference on natural language generation* (pp. 152–157).

Jackendoff, R. (1987). On beyond zebra: The relation of linguistic and visual information. *Cognition*, *26*, 89–114.

Jayannavar, P., Narayan-Chen, A., & Hockenmaier, J. (2020). Learning to execute instructions in a Minecraft dialogue. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2589–2602). Association for Computational Linguistics. Online.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017a). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2901–2910).

Johnson, J., Hariharan, B., Van Der Maaten, L., Hoffman, J., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017b). Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE international conference on computer vision* (pp. 2989–2998).

Kafle, K., & Kanan, C. (2017a). An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision* (pp. 1965–1973).

Kafle, K., & Kanan, C. (2017b). Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, *163*, 3–20.

Kafle, K., Price, B., Cohen, S., & Kanan, C. (2018). DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5648–5656).

Kafle, K., Shrestha, R., & Kanan, C. (2019). Challenges and prospects in vision and language research. *Frontiers in Artificial Intelligence*, *2*, 28.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics & Philosophy*, *30*(1), 1–45.

Kervadec, C., Antipov, G., Baccouche, M., & Wolf, C. (2021). Roses are red, violets are blue...but should VQA expect them to? In *Proceedings of CVPR*.

Kordjamshidi, P., Pustejovsky, J., & Moens, M.-F. (2020). EMNLP 2020 tutorial: Representation, learning and reasoning on spatial language for down-stream NLP tasks.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., & Li, F.-F. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32–73.

Kruszewski, G., Paperno, D., Bernardi, R., & Baroni, M. (2016). There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics*, *42*(4), 637–660.

Kuhnle, A., & Copestake, A. (2017). ShapeWorld: A new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lei, J., Yu, L., Bansal, M., & Berg, T. (2018). TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1369–1379). Association for Computational Linguistics.

Lei, J., Yu, L., Berg, T., & Bansal, M. (2020). TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8211–8225). Association for Computational Linguistics. Online.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Proceedings of ECCV (European conference on computer vision)* (pp. 740–755).

Lin, X., & Parikh, D. (2015). Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2984–2993).

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5210–5217). Association for Computational Linguistics. Online.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th international conference on neural information processing systems* (pp. 289–297).

Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input *Proceedings of the 27th international conference on neural information processing systems* (Vol. 1, pp. 1682–1690).

Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision* (pp. 1–9).

Manmadhan, S., & Kovoor, B. C. (2020). Visual question answering: A state-of-the-art review. *Artificial Intelligence Review*, *53*(8), 5705–5745.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Proceedings of ICRL*.

Mazuecos, M., Testoni, A., Bernardi, R., & Benotti, L. (2020). On the role of effective and referring questions in GuessWhat?! In *Proceedings of the first workshop on advances in language and vision research* (pp. 19–25). Association for Computational Linguistics. Online.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Monroe, W., Goodman, N., & Potts, C. (2016). Learning to generate compositional color descriptions. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2243–2248).

Monroe, W., Hawkins, R. X. D., Goodman, N. D., & Potts, C. (2017). Colors in context: A pragmatic neural model for grounded language understanding. *TACL*, *5*, 325–338.

Moradlou, S., Zheng, X., Tian, Y., & Ginzburg, J. (2021). Wh-Questions are understood before polar-questions: Evidence from English, German, and Chinese. *Journal of Child Language*, *48*(1), 157–183.

Nordemeyer, A., & Frank, M. (2014). A pragmatic account of the processing of negative sentences *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Parcalabescu, L., Gatt, A., Frank, A., & Calixto, I. (2020). Seeing past words: Testing the cross-modal capabilities of pretrained V&L models. *arXiv preprint arXiv:2012.12352*.

Partee, B. (1995). Lexical semantics and compositionality. *An invitation to cognitive science: Language*, *1*, 311–360.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Pezzelle, S., & Fernández, R. (2019a). Big generalizations with small data: Exploring the role of training samples in learning adjectives of size. In *Proceedings of the beyond vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)* (pp. 18–23).

Pezzelle, S., & Fernández, R. (2019b). Is the Red Square Big? MALeViC: Modeling Adjectives Leveraging Visual Contexts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 2858–2869).

Pezzelle, S., Greco, C., Gandolfi, G., Gualdoni, E., & Bernardi, R. (2020). Be different to be better! A benchmark to leverage the complementarity of language and vision. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 2751–2767). Association for Computational Linguistics. Online.

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* (pp. 337–342). Association for Computational Linguistics.

Pezzelle, S., Sorodoc, I., & 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 (Vol. 1, pp. 419–430). (Long Papers)

Quine, W. V. O. (1960). *Word and Object*. MIT Press.

Ray, A., Sikka, K., Divakaran, A., Lee, S., & Burachas, G. (2019). Sunny and dark outside?! Improving answer consistency in VQA through entailed question generation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5860–5865). Association for Computational Linguistics.

Ren, M., Kiros, R., & Zemel, R. S. (2015a). Exploring models and data for image question answering *Proceedings of the 28th international conference on neural information processing systems* (Vol. 2, pp. 2953–2961).

Ren, S., He, K., Girshick, R., & Sun, J. (2015b). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. In *Advances in neural information processing systems* (pp. 4967–4976).

Schüz, S., & Zarrieß, S. (2020). Knowledge supports visual language grounding: A case study on colour terms. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6536–6542).

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*, 417–424.

Selvaraju, R. R., Tendulkar, P., Parikh, D., Horvitz, E., Ribeiro, M. T., Nushi, B., & Kamar, E. (2020). SQuINTing at VQA models: Interrogating VQA models with sub-questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Shekhar, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., & Bernardi, R. (2017). Foil it! Find one mismatch between image and language caption *Proceedings of the 55th annual meeting of the association for computational linguistics* (Vol. 1, pp. 255–265). Long Papers

Shekhar, R., Venkatesh, A., Baumgärtner, T., Bruni, E., Plank, B., Bernardi, R., & Fernández, R. (2019). Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 2578–2587). (Long and Short Papers)

Shutova, E., Kiela, D., & Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 160–170).

Silberer, C., Ferrari, V., & Lapata, M. (2017). Visually grounded meaning representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2284–2297.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR*.

Sorodoc, I., Lazaridou, A., Boleda, G., Herbelot, A., Pezzelle, S., & Bernardi, R. (2016). "Look, some green circles!": Learning to quantify from images. In *Proceedings of the 5th Workshop on Vision and Language* (pp. 75–79). Association for Computational Linguistics.

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

Srivastava, Y., Murali, V., Dubey, S. R., & Mukherjee, S. (2021). Visual question answering using deep learning: A survey and performance analysis. In S. Singh, P. Roy, B. Raman, & P. Nagabhushan (Eds.), *Computer vision and image processing. CVIP 2020, volume 1377 of communications in computer and information science*. Springer.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2019). VL-BERT: Pre-training of generic visual-linguistic representations. In *International conference on learning representations*.

Suglia, A., Vergari, A., Konstas, I., Bisk, Y., Bastianelli, E., Vanzo, A., & Lemon, O. (2020). Imagining grounded conceptual representations from perceptual information in situated guessing games. In *Proceedings of the 28th international conference on computational linguistics* (pp. 1090–1102). International Committee on Computational Linguistics.

Suhr, A., Lewis, M., Yeh, J., & Artzi, Y. (2017). A corpus of natural language for visual reasoning. In *Proceedings of the annual meeting of the association for computational linguistics* (pp. 217–223). Association for Computational Linguistics.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., & Artzi, Y. (2019). A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 6418–6428). Association for Computational Linguistics.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).

Suzuki, R., Yanaka, H., Yoshikawa, M., Mineshima, K., & Bekki, D. (2019). Multimodal logical inference system for visual-textual entailment. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop* (pp. 386–392). Association for Computational Linguistics.

Tan, F., Cascante-Bonilla, P., Guo, X., Wu, H., Feng, S., & Ordonez, V. (2019). Drill-down: Interactive retrieval of complex scenes using natural language queries. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 2651–2661). Curran Associates, Inc.

Tan, H., & Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5100–5111). Association for Computational Linguistics.

Teney, D., Anderson, P., He, X., & Van Den Hengel, A. (2018). Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4223–4232).

Teney, D., Wu, Q., & van den Hengel, A. (2017). Visual question answering: A tutorial. *IEEE Signal Processing Magazine*, *34*(6), 63–75.

Testoni, A., Pezzelle, S., & Bernardi, R. (2019a). Quantifiers in a multimodal world: Hallucinating vision with language and sound. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 105–116). Association for Computational Linguistics.

Testoni, A., Shekhar, R., Fernández, R., & Bernardi, R. (2019b). The devil is in the detail: A magnifying glass for the GuessWhich visual dialogue game. In *Proceedings of the 23rd SemDial workshop on the semantics and pragmatics of dialogue (LondonLogue)* (pp. 15–24).

van Miltenburg, E., Morante, R., & Elliott, D. (2016). Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th workshop on vision and language* (pp. 54–59). Association for Computational Linguistics.

Vu, H. T., Greco, C., Erofeeva, A., Jafaritazehjan, S., Linders, G., Tanti, M., Testoni, A., Bernardi, R., & Gatt, A. (2018). Grounded textual entailment. In *Proceedings of the 27th international conference on computational linguistics* (pp. 2354–2368). Association for Computational Linguistics.

Walker, J., Gupta, A., & Hebert, M. (2014). Patch to the future: Unsupervised visual prediction. In *2014 IEEE conference on computer vision and pattern recognition* (pp. 3302–3309).

Wang, J., Madhyastha, P. S., & Specia, L. (2018). Object counts! Bringing explicit detections back into image captioning *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 2180–2193). Association for Computational Linguistics. (Long Papers)

Wang, P., Wu, Q., Shen, C., Dick, A., & Van Den Henge, A. (2017a). Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the 26th international joint conference on artificial intelligence* (pp. 1290–1296).

Wang, P., Wu, Q., Shen, C., Dick, A., & Van Den Hengel, A. (2017b). FVQA: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(10), 2413–2427.

Webb, N., & Webber, B. (2009). Special issue on interactive question answering: Introduction. *Natural Language Engineering*, *15*(1), 1–8.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*, 1–191.

Wisniewski, A (2015). *Handbook of contemporary semantic theory, chapter Semantics of questions*. Blackwell.

Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & van den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, *163*.

Xie, N., Lai, F., Doran, D., & Kadav, A. (2019). Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 21–29).

Yatskar, M., Zettlemoyer, L., & Farhadi, A. (2016). Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5534–5542).

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. B. (2018). Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *32nd conference on neural information processing systems (NeurIPS 2018)*.

Yu, L., Park, E., Berg, A. C., & Berg, T. L. (2015). Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2461–2469).

Zarrieß, S., & Schlangen, D. (2019). Know what you don't know: Modeling a pragmatic speaker that refers to objects of unknown categories. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 654–659).

Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., & Parikh, D. (2016). Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5014–5022).

Zheng, C., Guo, Q., & Kordjamshidi, P. (2020). Cross-modality relevance for reasoning on language and vision. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7642–7651).

Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., & Fergus, R. (2015). Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*.

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & Gao, J. (2020). Unified vision-language pre-training for image captioning and VQA *AAAI* (Vol. 34, pp. 13041–13049).

Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4995–5004).

AUTHOR BIOGRAPHIES

Raffaella Bernardi is Associate Professor at DISI (Department of Information Engineering and Computer Science) and CIMeC (Center for Mind/Brain Science), University of Trento. From 2002 till 2010, she has been assistant professor with a temporary contract at the Faculty of Computer Science, Free University of Bozen-Bolzano where she taught Computational Linguistics and acted as local coordinator of the Erasmus Mundus European Masters Programme in LCT. She studied at the Universities of Utrecht and Amsterdam specializing in Logic and Language, in 1999 she joined the international PhD Programme at the University of Utrecht and wrote a dissertation on categorial type logic (defended in June 2002). Since then she has continued to contribute extensively to this field by organizing several international workshops and disseminating the topic by means of teaching activities. After her PhD defence, she has worked within the Network of Excellence in Computational Logic (CoLogNET) for the area Logic and Natural Language Processing and she has been part of the Management Board of FoLLI (European Association for Logic, Language and Information) for several years. Furthermore, she has done joint work with the University of Bologna, the University of Pisa, the University of Amsterdam, INRIA Loraine, the University of Utrecht, the Australia University and New York University, carrying out work on Corpora, Grammar Induction within the Categorial Grammar framework, Question Answering and Categorial Type Logic. Her research interests took a computational turn in October 2002 when she moved to the Free University of Bozen-Bolzano and started working on Natural Language Interfaces to Structured Data. In 2011, she has started working on Distributional Semantics

investigating its compositional properties and its integration with Computer Vision models. She has supervised PhD projects on controlled natural language to access ontology and databases and on Interactive Question Answering Systems in the Library Domain. She is now mostly working on Multimodal Models. She is the local coordinator of the Erasmus Mundus European Masters Programme in LCT and of the Language and Multimodal Interaction track of the MSc in Cognitive Science offered by the University of Trento. She is the author of more than 100 publications in proceedings of international workshops, conferences and journals. She has been the PI within the EU Project "CACAO" (CP 2006 DILI 510035 CACAO Program: eContentplus), she has been member of the unitn team for the EU projects "Galateas" (CIP-ICT-PSP-2009-3 250430), LiMoSINe (FP7-PEOPLE 214905), CogNET (ICT-14-2014 RIA). She has been part of the team which won the ERC 2011 Starting Independent Research Grant COMPOSES (project nr. 283554). She has been member of the Management Board of the Cost Action The European Network on Integrating Vision and Language and she is part of the Executive Board for the Special Interest Group on Computational Semantics (SIGSem) and for Formal Grammar. She is the EU representative within the ACL Sponsorship Board.

Sandro Pezzelle is a postdoctoral researcher at the Institute for Logic, Language and Computation (ILLC) at the University of Amsterdam. He is member of the Dialogue Modelling Group led by Prof. Raquel Fernández and, since 2020, part of the DREAM (Distributed dynamic REpresentations for diAlogue Management) ERC project funded by the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455). In 2018, he obtained a PhD cum laude in Cognitive and Brain Sciences, track Language, Interaction and Computation at the Center for Mind/Brain Sciences (CIMeC), University of Trento, under the supervision of Prof. Raffaella Bernardi. Before, he graduated cum laude in Linguistics (MA) and Modern Literature (BA) at the University of Padova. He authored more than 20 peer-reviewed publications appeared in the proceedings of top-tier NLP conferences (ACL, EACL, NAACL, EMNLP) and workshops, journals (Cognition, Journal of Natural Language Engineering), and books. He (co-)supervised a dozen of student theses and projects at the Bachelor, Master, and PhD level. He gave lectures at various MSc courses on NLP and cognitive science topics; he acted as TA and TA coordinator for NLP courses. In 2021, he will teach and coordinate his first Master's course on cognitive science topics. He gave presentations at several conferences and workshops of various communities including NLP, linguistics, cognitive science, psycholinguistics, and computer vision. He is recognised by the scientific community, within which he plays an active role: he served as a reviewer for many events, including top-tier conferences in NLP (ACL, EACL, NAACL, EMNLP), cognitive science (CogSci), and machine learning (NeurIPS); he acted as the Publication chair for a well-established workshop in computational semantics (*SEM 2017), as well as co-Editor of a Frontiers Special Issue on Language and Vision. He co-organised three editions of a workshop at the crossroads of NLP, computer vision, and structured knowledge (LANTERN 2019-2021); he is Program co-Chair of a new symposium on NLP (EurNLP 2021). In 2020, together with Margot van Der Goot and Raquel Fernández, he was awarded a RPA Human(e) AI seed grant by the University of Amsterdam (project: "Exploring Adaptation of Conversational Systems to Different Age Groups"). His research is at the intersection of Computational Linguistics, Cognitive Science, and Computer Vision. In particular, he

focuses on how the semantics of words and sentences is affected by the (multimodal) context in which they are uttered or interpreted; how speakers can successfully communicate with each other by converging to a same or similar-enough semantic interpretation through a dialogical interaction; whether and how state-of-the-art AI models can mimic these processes.

How to cite this article: Bernardi R, Pezzelle S. Linguistic issues behind visual question answering. *Lang Linguist Compass.* 2021;elnc3.12417. https://doi.org/10.1111/lnc3.12417