Construction grammar and procedural semantics for human-interpretable grounded language processing

De Vos, Liesbet; Nevens, Jens; Van Eecke, Paul; Beuls, Katrien

Published in:
Linguistics Vanguard

DOI:
10.1515/lingvan-2022-0054

Publication date:
2024

License:
Unspecified

Document Version:
Accepted author manuscript

Construction grammar and procedural semantics for human-interpretable grounded language processing

Liesbet De Vos[1], Jens Nevens[2], Paul Van Eecke[*2], and Katrien Beuls[*1]

[1]Faculté d'informatique, Université de Namur
[2]Artificial Intelligence Laboratory, Vrije Universiteit Brussel

Abstract

Grounded language processing is a crucial component in many artificial intelligence systems, as it allows agents to communicate about their physical surroundings. State-of-the-art approaches typically employ deep learning techniques that perform end-to-end mappings between natural language expressions and representations grounded in the environment. Although these techniques achieve high levels of accuracy, they are often criticized for their lack of interpretability and their reliance on large amounts of training data. As an alternative, we propose a fully interpretable, data-efficient architecture for grounded language processing. The architecture is based on two main components. The first component comprises an inventory of human-interpretable concepts learned through task-based communicative interactions. These concepts connect the sensorimotor experiences of an agent to meaningful symbols that can be used for reasoning operations. The second component is a computational construction grammar that maps between natural language expressions and procedural semantic representations. These representations are grounded through their integration with the learned concepts. We validate the architecture using a variation on the CLEVR benchmark, achieving an accuracy of 96%. Our experiments demonstrate that the integration of a computational construction grammar with an inventory of interpretable grounded concepts can effectively achieve human-interpretable grounded language processing in the CLEVR environment.

Keywords— grounded language understanding; procedural semantics; computational construction grammar; Fluid Construction Grammar

1 Introduction

Intelligent systems that need to communicate about the world in which they are situated must be equipped with capacities for grounded language processing. Grounded language processing is the subarea of artificial intelligence that studies the connection between natural language on the one hand and perception and action in the world on the other (cf. Winograd 1972; Mooney 2008; Steels 2012; Kazemzadeh et al. 2014; Cirik et al. 2018; Nevens et al. 2019a; Persson et al. 2019; Alomari et al. 2022). For instance, when asked to answer the question *What color is the small sphere?* about the scene that is depicted in Figure 1,

[*] Joint last authors.

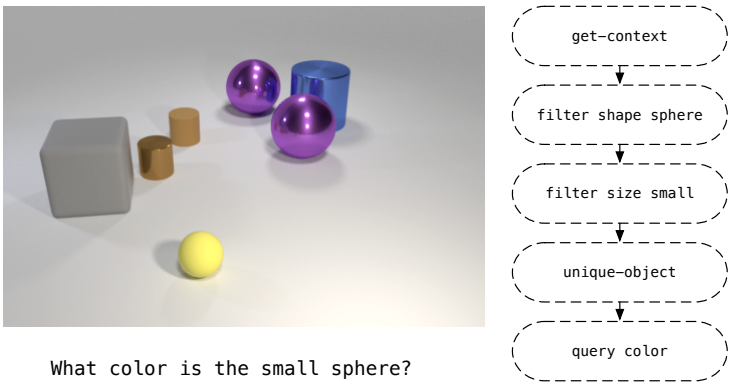


Figure 1: Grounded language processing in a visual question answering scenario.

an intelligent agent must be able to map the question onto its semantic structure, as well as to use this semantic structure to find an answer to the question. This semantic structure can for example take the form of a query that filters the scene for spheres, filters the set of spheres for small objects, checks whether the set of small spheres contains a single object, and retrieves the colour of this object. Upon execution of this query with respect to the scene depicted in Figure 1, the answer YELLOW would then be retrieved.

Methodologically, the state of the art in grounded language processing is dominated by deep neural network techniques (see e.g. Chen et al. 2015; Lu et al. 2016; Das et al. 2017; Jang et al. 2017; Yu et al. 2019). While these techniques are lauded for their high performance, they are often criticized for their reliance on large amounts of training data and their lack of interpretability (Massiceti et al. 2018; Marcus 2018; Mitchell 2020, 2021). As an alternative to these techniques, we propose a data-efficient and human-interpretable architecture for grounded language processing. The architecture is based on two main components. The first component comprises an inventory of human-interpretable concepts that were learned through task-based communicative interactions (Nevens et al. 2020). These concepts associate symbolic labels (e.g LEFT, BLUE, or SPHERE) with combinations of prototypical values on human-interpretable feature channels (e.g. POSITION-ON-X-AXIS, POSITION-ON-Y-AXIS, RED-VALUE, GREEN-VALUE, BLUE-VALUE, NUMBER-OF-CORNERS). The second component is a computational construction grammar that can be used to map between English questions and procedural semantic representations (Nevens et al. 2019b). These procedural semantic representations are composed of *primitive cognitive operations* that can be executed by an agent (e.g. FILTER, COUNT, or QUERY), and which make use of the concept representations provided by the first component. For example, the FILTER SHAPE SPHERE and FILTER SIZE SMALL operations used in Figure 1 make use of the concept representations for SPHERE and SMALL respectively. By integrating the concept inventory into the primitive operations, we provide the agent with effective building blocks to perform interpretable grounded language processing.

The remainder of this paper is structured as follows. Sections 2 and 3 introduce the two foundational components on which our system builds, namely Nevens et al.'s (2020) discrimination-based learning of grounded concepts and Nevens et al.'s (2019b) computational construction grammar-based approach to visual question answering. Section 4 describes our integrated architecture for human-interpretable grounded language processing, which constitutes the main contribution of this paper, and evaluates it with respect to a variation on the CLEVR visual question answering benchmark (Johnson, Hariharan, van der Maaten, Fei-Fei, et al. 2017). Finally, Section 5 reflects on the evaluation results and concludes the paper.

2 Discrimination-based learning of grounded concepts

As humans, we observe the world through our sensory experiences. We see colour and shape, feel texture and temperature, hear pitch and loudness, taste sweetness and bitterness, and smell a wide range of scents. To share these sensory experiences with others, we rely on a variety of concepts that can be expressed using language. While linguistic expressions can be shared between individuals, the concepts and their grounding in underlying experiences are personal and differ across individuals. Concepts and their linguistic expressions form thus an abstraction layer over personal experiences that is crucial for communication (Steels and Belpaeme 2005). Correspondingly, intelligent agents should have at their disposal an inventory of concepts that allows them to abstract away from individual sensor values. Ideally, this inventory should be established incrementally and in a dynamic manner, so that new concepts can easily be added, and existing ones can adapt over time (Wellens 2012; Loetzsch 2015; Bleys 2016; Steels et al. 2016).

To learn such adaptive concepts, Nevens et al. (2020) introduce a discrimination-based architecture for grounded concept learning. Adopting the language game methodology (Steels 2001, 2012), in which a population of agents engages in task-oriented communicative interactions in order to converge on a shared communication system, the authors set up a scenario with two agents: a tutor and a learner. The tutor is provided with an inventory of ontological categories that can be used to discriminate objects in the CLEVR dataset (Johnson, Hariharan, van der Maaten, Fei-Fei, et al. 2017). The learner, on the other hand, starts the experiment without any concepts in place, any notion of how many concepts should be learned, nor any knowledge of the ontological categories that the tutor makes use of (e.g. colours, shapes, materials, etc.). Each interaction is situated in a randomly generated scene comparable to the one depicted in Figure 1 and proceeds as follows:

1. The tutor agent selects a topic from the scene, that is, the object to which they will try to draw the learner's attention.
2. The tutor agent searches for a concept that can uniquely discriminate this topic from the other objects in the scene.
3. The tutor agent produces the linguistic expression associated to the chosen concept.
4. The learner agent then retrieves the concept associated to the observed linguistic expression in their inventory.
5. If a concept is found, the learner agent points to the object in the scene most similar to this concept. Otherwise, the learner signals that they do not know the word.
6. The tutor provides feedback to the learner by pointing to the topic object.
7. The learner agent updates its concept inventory according to the outcome of the interaction.

A game succeeds if the learner agent correctly pointed to the topic object, and fails otherwise. As a consequence of the updating phase that takes place after each interaction, the concept inventory of the learner gradually converges towards the inventory of the tutor.

Figure 2 shows an example of the concept CUBE, learned through a series of tutor-learner language games. We see that the label "cube" has been associated with two attributes: number of sides (with a prototypical value of six sides) and number of corners (with a prototypical value of eight corners). A score is assigned to each attribute, representing how strongly that attribute is associated to the concept.

In order to use the learned concepts in communicative interactions, an agent must be able to calculate

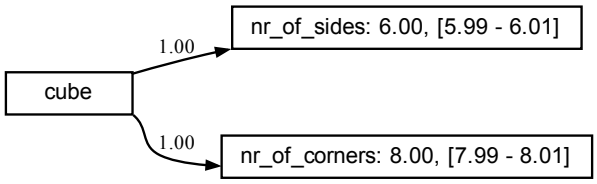


Figure 2: Example of the concept CUBE, learned over the course of 1000 interactions. The label "cube" is associated to two attributes, namely "nr_of_sides = 6" and "nr_of_corners = 8".

the fit between concepts and objects in the scene. This fit is calculated by a weighted similarity measure that takes into account all attributes of the concept and weighs them using the attribute's certainty score. Through a series of such communicative interactions with the tutor agent, the learner agent acquires a dynamic repertoire of concepts that enables them to communicate about the objects in their environment.

3 A computational construction grammar for visual question answering

Grounded language understanding requires an autonomous agent to relate the meaning of linguistic expressions to the concrete situations in which these are uttered. For example, in order to answer the question in Figure 1, a mapping needs to be established between its semantic structure and the image about which the question is asked. Operationalizing a grounded language understanding system involves thus mapping between natural language expressions and semantic structures that ground entities or events in the world.

The field of semantic parsing is concerned with mapping natural language utterances onto a formal representation of their meaning. Coarsely, techniques for semantic parsing can be grouped into approaches that model the grammar of a language formally and those that do not. The second group consists of neural and statistical approaches that rely on regularities extracted from large amounts of annotated corpus data (e.g. Hu et al. 2017; Johnson, Hariharan, van der Maaten, Hoffman, et al. 2017; Yi et al. 2018; Mao et al. 2019; Andreas et al. 2016). While these technologies generally achieve high levels of accuracy on broad-coverage benchmark datasets, they are often criticized for their lack of transparency (Marcus 2018; Mitchell 2020). The first group consists of grammar-based techniques, including approaches based on formalisms such as Head-Driven Phrase Structure Grammar (McFetridge et al. 1996; Frank et al. 2007), Lexical-Functional Grammar (Yarmohammadi et al. 2008), Combinatory Categorial Grammar (CCG) (Zettlemoyer and Collins 2005), and Fluid Construction Grammar (FCG) (Marques and Beuls 2016; Beuls et al. 2021; Van Eecke et al. 2022). These approaches have the advantages of transparency, linguistic motivation, and human interpretability, but they are notoriously difficult to operationalize on a large scale as the grammars typically need to be designed by a grammar engineer (van Trijp et al. 2022). Yet, advances in the automatic learning of CCG (Liang 2016) and, more recently, FCG grammars (Van Eecke 2018; Nevens et al. 2022; Doumen et al. 2023; Beuls and Van Eecke 2023) show that this limitation has been lifted.

For the purposes of this paper, we will employ the FCG formalism (van Trijp et al. 2022; Beuls and Van Eecke 2024), including the grammar proposed by Nevens et al. (2019b). This grammar adequately covers the CLEVR dataset and makes use of a procedural semantic representation formalized in Incremental Recruitment Language (IRL) (Van den Broeck 2008; Spranger et al. 2010). The meaning representation used by the grammar is executable on symbolic scene representations but not grounded in continuous data. In other terms, it does not deal with the problem of relating the semantic structures to perceptual sensor

values. For example, when asked to filter a set of objects for cubes, the IRL engine considers each element of the set and returns all those that have the value CUBE for their SHAPE attribute. Likewise, when asked to query the colour of an object, it literally retrieves the value of the object's COLOR property. Figure 3 shows an example of such a symbolic scene representation, a meaning representation for the question *What material is the blue cube?*, and its execution with respect to the symbolic scene. An overview of the 15 primitive operations is provided in Table 1.

4 An integrated architecture for human-interpretable grounded language processing

This section describes our architecture for grounded language understanding, which integrates the discrimination-based learning of concepts introduced in Section 2 and the grammar-based approach to parsing natural language expressions into procedural semantic representations introduced in Section 3. It thereby presents the main contribution of this paper.

The cornerstone of the integration of both strands of research is an implementation of the primitive operations presented in Table 1 that is grounded in continuous perceptual data through concepts that bridge between symbolic and numerical representations as exemplified in Figure 2. In order to achieve this, the first step was to replace the ontology underlying the semantic representation used by the computational construction grammar (cf. Section 3) with a new ontology that consists of the concepts learned through discrimination-based language games (cf. Section 2). A schematic overview of this ontology is shown in Figure 4.

The second step consisted in operationalizing the actual primitives using these concepts, that is, to use the concepts for supporting the operations of (i) filtering objects according to their perceptual categories (e.g. retrieving the set of blue objects in a scene), (ii) querying the perceptual categories applicable to a given object (e.g. querying the colour of an object), (iii) spatially relating objects to each other (e.g. retrieving the objects to the left of a given object), and (iv) comparing objects with respect to a given category (e.g. retrieving all objects with the same colour as a given object). When operationalizing these primitives, it is important to consider that concepts are not only meaningful on their own, that is, they take the form of a prototype, but also in relation to each other. For example, an object can be situated at a certain distance from prototypical blue, yet this is not sufficient for this object to be called blue. Indeed, calling the object blue would also need to entail that no other colour prototype is situated closer to this object than the blue prototype. This behaviour is achieved in the primitives by computing the distance between an object and all concepts on an ontological axis (e.g. large and small for the size axis), and selecting the concept that is closest to the object. In our example, an object would be considered blue if no other colour category prototype is closer to the object than the blue one. The same reasoning is applied to all other ontological axes, such as size, shape, and material.

An additional challenge consists in the integration of the spatial concepts LEFT, RIGHT, FRONT, and BEHIND. As these concepts were learned in isolation, they represent absolute prototypes. For example, *left* refers to the left portion of the scene. However, the CLEVR dataset uses these concepts in a relative fashion. Indeed, *the objects left of the red cube* establishes the red cube as a landmark and refers to all objects that appear left relative to this landmark. This is exemplified in Figure 5, where the absolute concept is shown on the left-hand side and the relative concept is shown on the right-hand side. In the primitives, this behaviour is achieved by dynamically shifting the prototypical values of a concept with respect to the landmark object. As such, the neutral value for left and right is shifted from the middle of the image to the position of the landmark.

An example of the integrated architecture in action is shown in Figure 6. In this figure, the question *What*

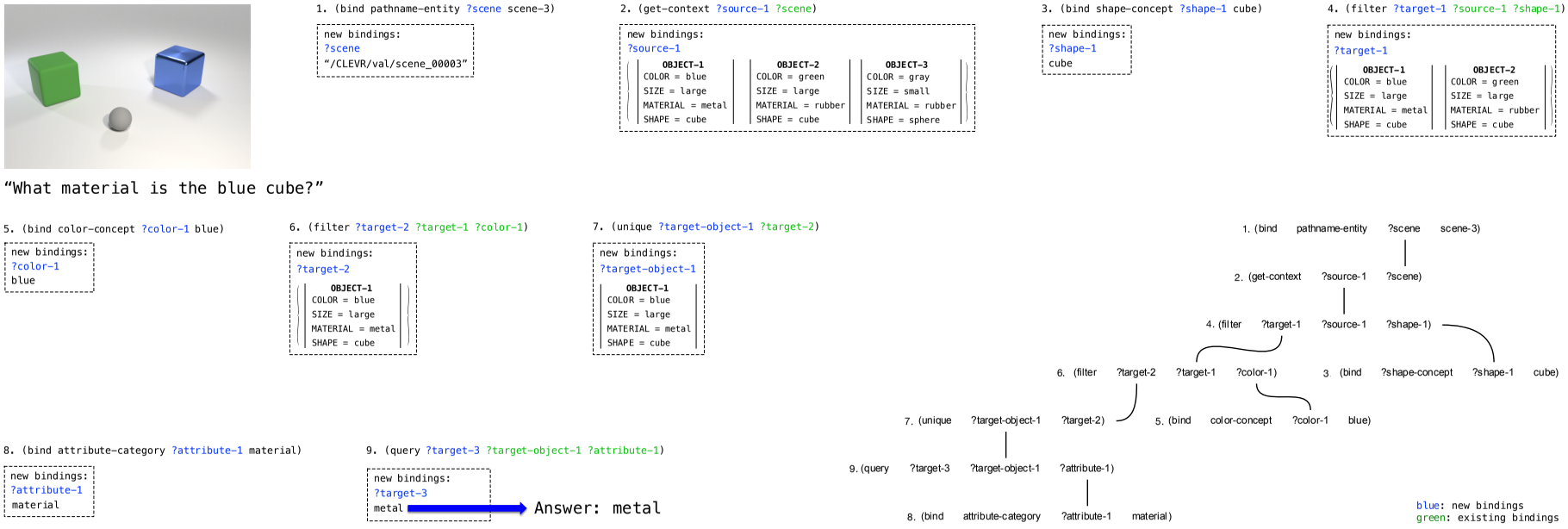


Figure 3: Example of a symbolic scene representation (upper left corner), a meaning representation for the question *What material is the blue cube?* (lower right corner) and its execution with respect to the symbolic scene (in between).

Table 1: CLEVR IRL primitives as implemented by Nevens et al. (2019b).

Primitive	**Arguments**	**Implemented behaviour**
1. BIND	`type, ?target, category`	Binds `category`, of `type` (concept or pathname) to `?target`.
2. GET-CONTEXT	`?target, ?scene`	Loads `?scene` and binds it to `?target`.
3. COUNT	`?target-num, ?source-set`	Counts the number of objects in `?source-set` and binds that number to `?target-num`.
4. EQUAL-INTEGER	`?target-bool, ?num-1, ?num-2`	Compares `?num-1` to `?num-2` and binds the resulting Boolean value to `?target-bool`.
5. LESS-THAN	`?target-bool, ?num-1, ?num-2`	Checks whether `?num-1` is smaller than `?num-2` and binds the resulting Boolean value to `?target-bool`.
6. GREATER-THAN	`?target-bool, ?num-1, ?num-2`	Checks whether `?num-1` is greater than `?num-2` and binds the resulting Boolean value to `?target-bool`.
7. EQUAL	`?target-bool, ?concept-1, ?concept-2`	Checks whether `?concept-1` and `concept-2` are the same and binds the resulting Boolean value to `?target-bool`.
8. EXIST	`?target-bool, ?source-set`	Checks whether `?source-set` is non-empty and binds the resulting Boolean value to `?target-bool`.
9. INTERSECT	`?target-set, ?source-set-1, ?source-set-2`	Finds the intersection of `?source-set-1` and `?source-set-2` and binds it to `?target-set`.
10. UNION	`?target-set, ?source-set-1, ?source-set-2`	Finds the union of `?source-set-1` and `?source-set-2` and binds it to `?target-set`.
11. UNIQUE	`?target-object, ?source-set`	Checks whether there is exactly one object in `?source-set` and binds that object to `?target-object` if it is the case.
12. FILTER	`?target-set, ?source-set, ?concept`	Filters `?source-set` using `?concept` and binds the resulting set of objects to `?target-set`.
13. QUERY	`?target-category, ?source-object, ?attribute`	Finds out which `?target-category` applies to `?source-object` when querying for `?attribute`.
14. SAME	`?target-set, ?source-set, ?source-object, ?attribute`	Finds all objects in the `?source-set` that belong to the same category for `?attribute` as the `?source-object`, and binds them to `?target-set`.
15. RELATE	`?target-set, ?source-set, ?source-object, ?spatial-concept`	Finds all objects in `?source-set` that are in the given spatial relation (indicated by `?spatial-concept`) to the `?source-object`, and binds them to `?target-set`.

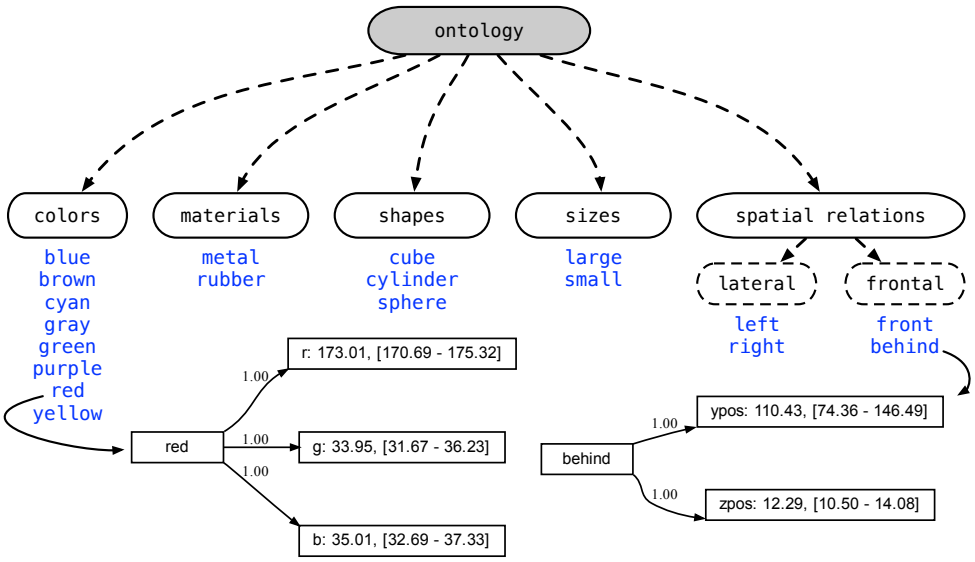
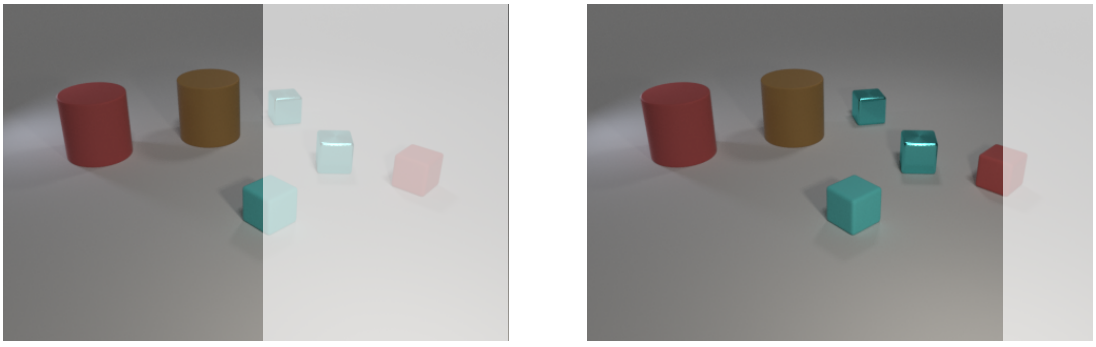


Figure 4: The ontology underlying the grounded language understanding experiment.



left

How many objects are left of the red cube?

Figure 5: The use of the spatial concept LEFT in an absolute and a relative fashion.

material is the blue cube? is asked about the image shown in the upper left corner. First, the utterance is analysed by the grammar into the meaning representation shown in the lower right corner of the figure. Then, this meaning representation is executed on the scene. Steps 1 (BIND) and 2 (GET-CONTEXT) involve retrieving the image and mapping the objects in the image to visual features, according to the procedure described in Nevens et al. (2020). Steps 3 (BIND) and 4 (FILTER) involve filtering the set of retrieved objects for cubes. Steps 5 (BIND) and 6 (FILTER) involve filtering the resulting set of cubes for blue objects. Step 7 (UNIQUE) verifies whether the resulting set contains a single object and steps 8 (BIND) and 9 (QUERY) query the material of this object.

In order to evaluate our integrated architecture, its accuracy was computed on the 149,991 questions of the CLEVR validation split, by comparing the predicted answers to the ground truth annotation. This process was repeated 10 times with a different series of concepts, that is, concepts resulting from a different experimental run during the learning phase. All experimental runs led to an accuracy between 95% and 98%, with an average of 96%. These results are in line with most state-of-the-art approaches to the CLEVR dataset, although they are not directly comparable due to the different procedure for generating continuous scene representations (see Nevens et al. 2020). It is important to keep in mind that the CLEVR dataset is conceived as a diagnostic dataset that aims to boost the development of innovative approaches to visual question answering. This means that the primary objective is not to compete on a percentage point accuracy level, but to diagnose possible problems with new techniques. In fact, a wide variety of approaches have been shown to obtain an accuracy of 95–99.9% on the CLEVR dataset, including approaches that make use of scene graphs (Yi et al. 2018; Mao et al. 2019), MAC networks (Hudson and Manning 2018), and neural module networks (Johnson, Hariharan, van der Maaten, Hoffman, et al. 2017; Hu et al. 2018).

An important asset of our methodology concerns the human interpretability of all steps involved in the language understanding process. A human-interpretable grammar maps between natural language questions and a meaning representation that lays out the reasoning steps involved in answering the questions. These reasoning steps correspond to human-interpretable operations which in turn make use of human-interpretable concepts. Consequently, if the model provides an unexpected answer to a question, a human expert can trace back the chain of operations that led to this answer. Figure 7 shows an example where the system could not answer the question *What color is the thing in front of the large cube?*. We can see that the grammar has mapped the question to a suitable semantic representation, but that the UNIQUE operation fails as a result of a wrong value returned by the RELATE operation. Indeed, the RELATE operation falsely considered two objects to be in front of the large cube. As a result of this failure, the system detects that it cannot adequately perform all reasoning steps and signals that the question could not be answered. In fact, this behaviour sets our system apart from the deep learning-based approaches that dominate the field today, as these have great difficulties at self-diagnosing reasoning uncertainties and thereby often make overconfident false predictions (see e.g. Guo et al. 2017; Thulasidasan et al. 2019).

5 Discussion and conclusion

This research has brought together two distinct strands of research and has integrated them into an architecture that aims to achieve human-interpretable grounded language processing on both the conceptual and the grammatical level. The first strand of research concerned the learning of concepts that map between continuous sensor values and categories that are meaningful in the environment. These concepts were learned through discrimination-based language games and it was previously shown that they can be learned in a data-efficient manner, that they generalize well to unseen instances, that they are transparent and human-interpretable, that they are adaptive to changes in the environment and that they can be

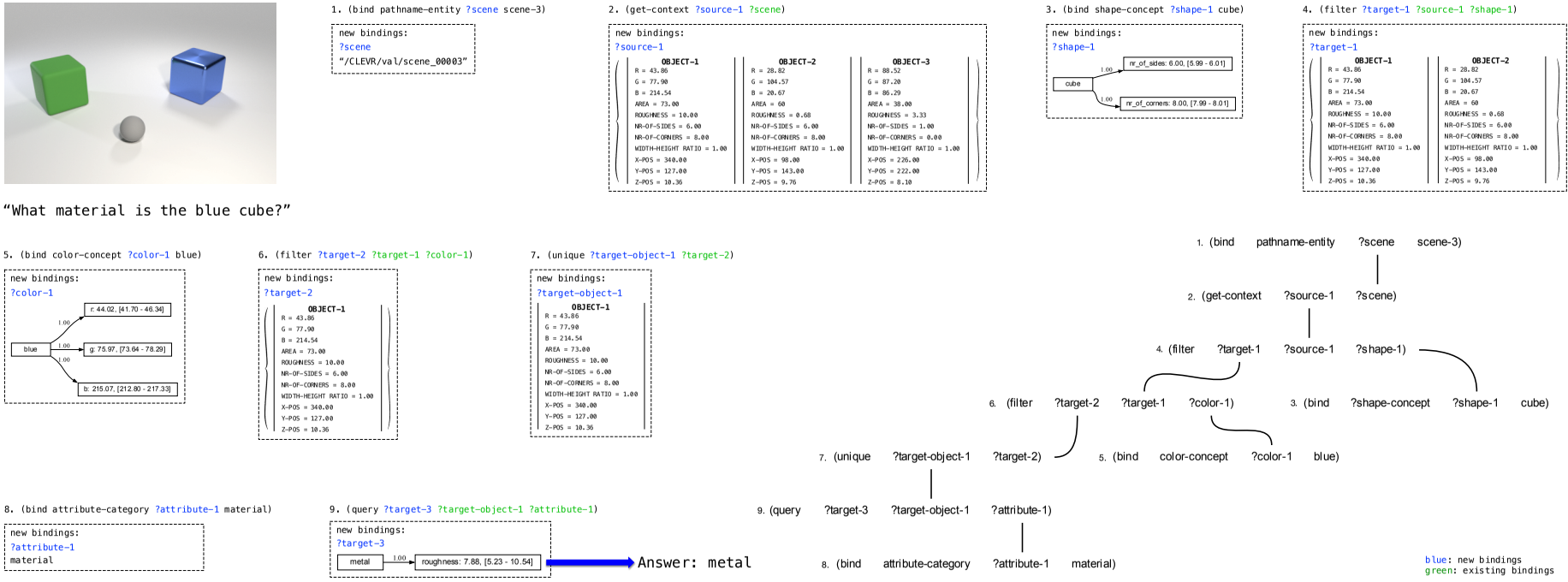


Figure 6: Example of the execution of the semantic network underlying the utterance *What material is the blue cube?* (lower right corner) with respect to the image shown in the upper left corner.

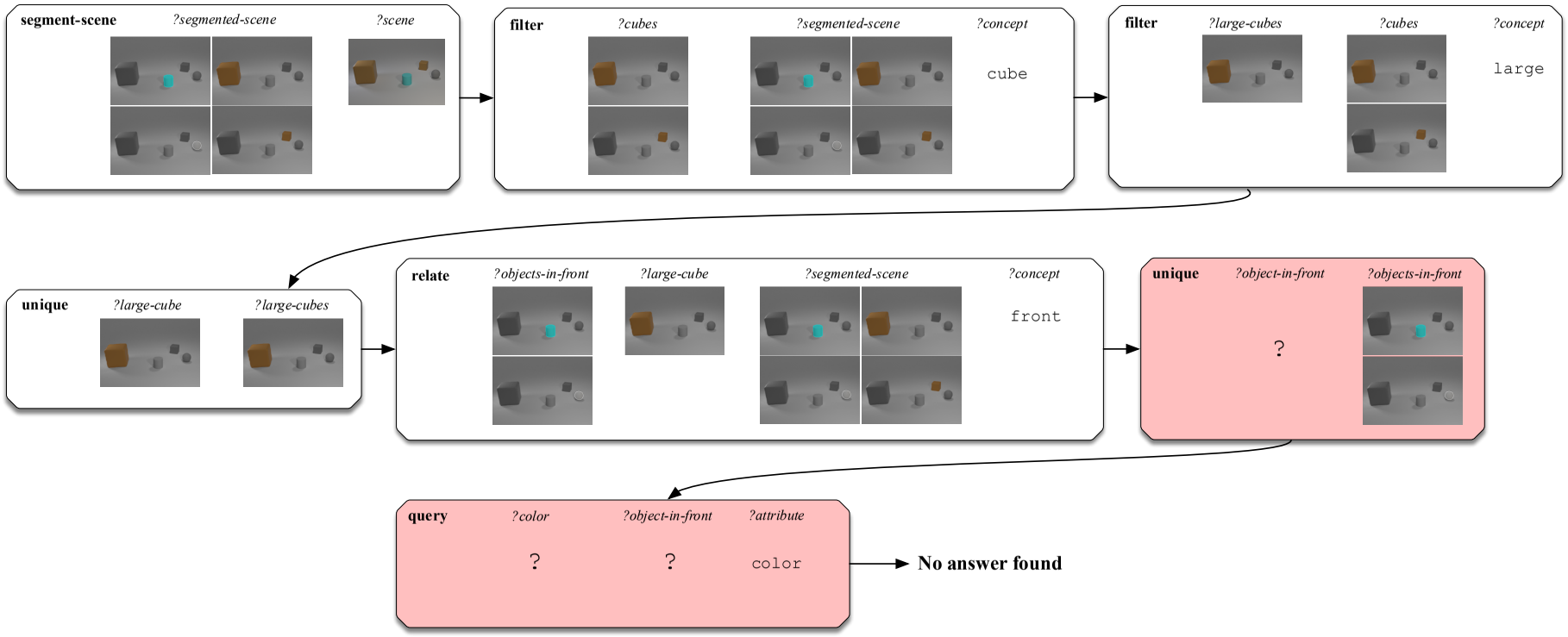


Figure 7: Example of an execution process where the model cannot answer the question *What color is the thing in front of the large cube?*. The unexpected answer is due to the execution of the RELATE primitive, where the model identifies two objects in front of the large cube instead of a single one.

combined compositionally (Nevens et al. 2020). However, this strand of research focused exclusively on concept learning and was not concerned with linguistic structures exceeding the level of individual words. The second strand of research concerned the use of construction grammar and procedural semantics to model the relation between linguistic expressions and their underlying semantic structure. We have focused on a grammar that covers the CLEVR visual question answering benchmark dataset and a procedural semantic representation that was executable on symbolic scene representations, but did not deal with the problem of grounding concepts in continuous environments (Nevens et al. 2019b). In our architecture, we have tied together the grammar-based approach to visual question answering and the discrimination-based approach to concept learning. We have evaluated the results on a variation of the CLEVR dataset and achieved an accuracy of 96% on average. We have thereby shown that both strands of research are complementary to and compatible with each other, and that their combination can be leveraged to achieve human-interpretable grounded language processing down to the level of continuous features. These results are in line with the growing body of research on neuro-symbolic artificial intelligence that combines sub-symbolic and symbolic techniques to tackle cognitive tasks that involve both perception and reasoning (Garcez et al. 2015; Manhaeve et al. 2021). The main challenge remains the scaling of the approach that we have introduced. Recent breakthroughs in the abductive learning of computational construction grammars show that the automatic learning of interpretable construction grammars that can map between utterances and procedural semantic representations is feasible (Van Eecke and Beuls 2017; Nevens 2022; Doumen et al. 2023; Beuls and Van Eecke 2023), although more research is needed in this area to make the learning methods applicable to broad-coverage corpora. When it comes to the concept representations, it is unlikely that the representations introduced by Nevens et al. (2020) would as such scale far beyond scenes of geometrical figures. More research would here be required to include more flexible ways to represent the distribution of relevant values on the individual feature channels. While this would possibly render the learning process less data-efficient, a wider range of more fine-grained concepts could then be captured.

References

Alomari, Muhannad, Fangjun Li, David C. Hogg & Anthony G. Cohn. 2022. Online perceptual learning and natural language acquisition for autonomous robots. *Artificial Intelligence* 303. 103637. https://doi.org/10.1016/j.artint.2021.103637.

Andreas, Jacob, Marcus Rohrbach, Trevor Darrell & Dan Klein. 2016. Learning to compose neural networks for question answering. In Kevin Knight, Ani Nenkova & Owen Rambow (eds.), *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, 1545–1554. San Diego, CA: Association for Computational Linguistics.

Beuls, Katrien & Paul Van Eecke. 2023. Fluid construction grammar: State of the art and future outlook. In Claire Bonial & Harish Tayyar Madabushi (eds.), *Proceedings of the first International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, 41–50. Washington, D.C.: Association for Computational Linguistics.

Beuls, Katrien & Paul Van Eecke. 2024. Construction grammar and artificial intelligence. In Mirjam Fried & Kiki Nikiforidou (eds.), *The Cambridge handbook of construction grammar*. Forthcoming. Cambridge, United Kingdom: Cambridge University Press.

Beuls, Katrien, Paul Van Eecke & Vanja Sophie Cangalovic. 2021. A computational construction grammar approach to semantic frame extraction. *Linguistics Vanguard* 7(1). 20180015. https://doi.org/10.1515/lingvan-2018-0015.

Bleys, Joris. 2016. *Language strategies for the domain of colour*. Berlin: Language Science Press.

Chen, Kan, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu & Ram Nevatia. 2015. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*. https://doi.org/10.48550/arXiv.1511.05960.

Cirik, Volkan, Taylor Berg-Kirkpatrick & Louis-Philippe Morency. 2018. Using syntax to ground referring expressions in natural images. In Sheila McIlraith & Kilian Q. Weinberger (eds.), *Proceedings of the thirty-second AAAI Conference on Artificial Intelligence*, 6756–6764. Washington, D.C.: AAAI Press. https://doi.org/10.1609/aaai.v32i1.12343.

Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh & Dhruv Batra. 2017. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1080–1089. Washington, D.C.: IEEE Computer Society.

Doumen, Jonas, Katrien Beuls & Paul Van Eecke. 2023. Modelling language acquisition through syntactico-semantic pattern finding. In Andreas Vlachos & Isabelle Augenstein (eds.), *Findings of the association for computational linguistics: EACL 2023*, 1317–1327. Dubrovnik: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-eacl.99.

Frank, Anette, Hans-Ulrich Krieger, Feiyu Xu, Hans Uszkoreit, Berthold Crysmann, Brigitte Jörg & Ulrich Schäfer. 2007. Question answering from structured knowledge sources. *Journal of Applied Logic* 5(1). 20–48. https://doi.org/10.1016/j.jal.2005.12.006.

Garcez, Artur d'Avila, Tarek R. Besold, Luc De Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C. Lamb, Risto Miikkulainen & Daniel L. Silver. 2015. Neural-symbolic learning and reasoning: Contributions and challenges. In *2015 AAAI Spring symposium series*, 18–21. Washington, D.C.: AAAI Press.

Guo, Chuan, Geoff Pleiss, Yu Sun & Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In Doina Precup & Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 1321–1330. Sydney: JMLR.org.

Hu, Ronghang, Jacob Andreas, Trevor Darrell & Kate Saenko. 2018. Explainable neural computation via stack neural module networks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu & Yair Weiss (eds.), *European conference on computer vision (eccv 2018)*, 53–69. Cham: Springer.

Hu, Ronghang, Jacob Andreas, Marcus Rohrbach, Trevor Darrell & Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 804–813. Washington, D.C.: IEEE Computer Society.

Hudson, Drew A. & Christopher D. Manning. 2018. Compositional attention networks for machine reasoning. In *6th International Conference on Learning Representations (ICLR 2018)*, 1–20. Vancouver.

Jang, Yunseok, Yale Song, Youngjae Yu, Youngjin Kim & Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *2017 IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), 2758–2766. Washington, D.C.: IEEE Computer Society. https://doi.org/10.1109/CVPR.2017.149.

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick & 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)*, 2901–2910. Washington, D.C.: IEEE Computer Society. https://doi.org/10.1109/CVPR.2017.215.

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick & Ross Girshick. 2017. Inferring and executing programs for visual reasoning. In Rita Cucchiara, Yasuyuki Matsushita, Nicu Sebe & Stefano Soatto (eds.), *2017 IEEE International Conference on Computer Vision (ICCV)*, 2989–2998. Washington, D.C.: IEEE Computer Society. https://doi.org/10.1109/ICCV.2017.325.

Kazemzadeh, Sahar, Vicente Ordonez, Mark Matten & Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang & Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 787–798. Doha: Association for Computational Linguistics.

Liang, Percy. 2016. Learning executable semantic parsers for natural language understanding. *Communications of the ACM* 59(9). 68–76. https://doi.org/10.1145/2866568.

Loetzsch, Martin. 2015. *Lexicon formation in autonomous robots*. Berlin: Humboldt-Universität zu Berlin dissertation.

Lu, Jiasen, Jianwei Yang, Dhruv Batra & Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In Daniel Lee, Masashi Sugiyama, Ulrike Von Luxburg, Isabelle Guyon & Roman Garnett (eds.), *Advances in neural information processing systems 29 (NIPS 2016)*, 289–297. Red Hook, NY: Curran Associates.

Manhaeve, Robin, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester & Luc De Raedt. 2021. Neural probabilistic logic programming in DeepProbLog. *Artificial Intelligence* 298. 103504. https://doi.org/10.1016/j.artint.2021.103504.

Mao, Jiayuan, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum & Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *7th International Conference on Learning Representations (ICLR 2019)*. New Orleans, LA.

Marcus, Gary. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*. https://doi.org/10.48550/arXiv.1801.00631.

Marques, Tânia & Katrien Beuls. 2016. Evaluation strategies for computational construction grammars. In Yuji Matsumoto & Rashmi Prasad (eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical papers*, 1137–1146. Osaka, Japan: International Committee on Computational Linguistics.

Massiceti, Daniela, Puneet K. Dokania, N. Siddharth & Philip Torr. 2018. Visual dialogue without vision or dialogue. In *Critiquing and correcting trends in machine learning workshop: NeurIPS 2018*. Montreal, Canada.

McFetridge, Paul, Fred Popowich & Dan Fass. 1996. An analysis of compounds in HPSG (Head-driven Phrase Structure Grammar) for database queries. *Data & Knowledge Engineering* 20(2). 195–209.

Mitchell, Melanie. 2020. On crashing the barrier of meaning in artificial intelligence. *AI Magazine* 41(2). 86–92. https://doi.org/10.1609/aimag.v41i2.5259.

Mitchell, Melanie. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences* 1505(1). 79–101.

Mooney, Raymond J. 2008. Learning to connect language and perception. In Dieter Fox & Carla Gomes (eds.), *Proceedings of the twenty-third AAAI conference on artificial intelligence*, 1598–1601. Washington, D.C.: AAAI Press.

Nevens, Jens. 2022. *Representing and learning linguistic structures on the conceptual, morphosyntactic, and semantic level*. Brussels: Vrije Universiteit Brussel dissertation.

Nevens, Jens, Jonas Doumen, Paul Van Eecke & Katrien Beuls. 2022. Language acquisition through intention reading and pattern finding. In Nicoletta Calzolari & Chu-Ren Huang (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, 15–25. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Nevens, Jens, Paul Van Eecke & Katrien Beuls. 2019a. A practical guide to studying emergent communication through grounded language games. In *AISB 2019 Symposium on Language Learning for Artificial Agents*, 1–8. Falmouth: AISB.

Nevens, Jens, Paul Van Eecke & Katrien Beuls. 2019b. Computational construction grammar for visual question answering. *Linguistics Vanguard* 5(1). 20180070. https://doi.org/10.1515/lingvan-2018-0070.

Nevens, Jens, Paul Van Eecke & Katrien Beuls. 2020. From continuous observations to symbolic concepts: A discrimination-based strategy for grounded concept learning. *Frontiers in Robotics and AI* 7(84). https://doi.org/10.3389/frobt.2020.00084.

Persson, Andreas, Pedro Miguel Zuidberg Dos Martires, Luc De Raedt & Amy Loutfi. 2019. Semantic relational object tracking. *IEEE Transactions on Cognitive and Developmental Systems* 12(1). 84–97.

Spranger, Michael, Simon Pauw & Martin Loetzsch. 2010. Open-ended semantics co-evolving with spatial language. In Erica A. Cartmill, Sean Roberts, Heidi Lyn & Hannah Cornish (eds.), *Proceedings of the 10th international conference (EVOLANGX)*, 297–304. Singapore: World Scientific. https://doi.org/10.1142/9789814295222_0038.

Steels, Luc. 2001. Language games for autonomous robots. *IEEE Intelligent Systems* 16. 16–22.

Steels, Luc. 2012. Grounding language through evolutionary language games. In Luc Steels & Manfred Hild (eds.), *Language grounding in robots*, 1–22. New York, NY: Springer. https://doi.org/10.1007/978-1-4614-3064-3_1.

Steels, Luc & Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences* 28(4). 469–489. https://doi.org/10.1017/S0140525X05000087.

Steels, Luc, Martin Loetzsch & Michael Spranger. 2016. A boy named Sue: The semiotic dynamics of naming and identity. *Belgian Journal of Linguistics* 30(1). 147–169. https://doi.org/10.1075/bjl.30.07ste.

Thulasidasan, Sunil, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya & Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox & Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 13843–13854. Red Hook, NY, USA: Curran Associates. https://doi.org/10.2172/1525811.

Van den Broeck, Wouter. 2008. Constraint based compositional semantics. In Andrew D. M. Smith, Kenny Smith & Ramon Ferrer i Cancho (eds.), *Proceedings of the 7th International Conference on the Evolution of Language (EVOLANG7)*, 338–345. World Scientific. https://doi.org/10.1142/9789812776129_0043.

Van Eecke, Paul. 2018. *Generalisation and specialisation operators for computational construction grammar and their application in evolutionary linguistics research.* Brussels: Vrije Universiteit Brussel dissertation.

Van Eecke, Paul & Katrien Beuls. 2017. Meta-layer problem solving for computational construction grammar. In *The 2017 AAAI Spring symposium series*, 258–265. Washington, D.C.: AAAI Press.

Van Eecke, Paul, Jens Nevens & Katrien Beuls. 2022. Neural heuristics for scaling constructional language processing. *Journal of Language Modelling* 10(2). 287–314.

van Trijp, Remi, Katrien Beuls & Paul Van Eecke. 2022. The FCG Editor: An innovative environment for engineering computational construction grammars. *PLOS ONE* 17(6). e0269708. https://doi.org/10.1371/journal.pone.0269708.

Wellens, Pieter. 2012. *Adaptive strategies in the emergence of lexical systems.* Brussels: Vrije Universiteit Brussel dissertation.

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

Yarmohammadi, Mahsa A., Mehrnoush Shamsfard, Mahshid A. Yarmohammadi & Masoud Rouhizadeh. 2008. SBUQA question answering system. In *Advances in computer science and engineering: Csicc 2008*, 316–323. Berlin: Springer. https://doi.org/10.1007/978-3-540-89985-3_39.

Yi, Kexin, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli & Josh Tenenbaum. 2018. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In Samy Bengio, Hanna Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi & Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, 1031–1042. Red Hook, NY, USA: Curran Associates.

Yu, Zhou, Jun Yu, Yuhao Cui, Dacheng Tao & Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6274–6283. Washington, D.C.: IEEE Computer Society. https://doi.org/10.1109/CVPR.2019.00644.

Zettlemoyer, Luke & Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In Fahiem Bacchus & Tommi Jaakkola (eds.), *Proceedings of the twenty-first Conference on Uncertainty in Artificial Intelligence*, 658–666. Edinburgh: AUAI Press.