Language Acquisition through Intention Reading and Pattern Finding

Jens Nevens[1] and **Jonas Doumen**[2,3] and **Paul Van Eecke**[1,2,3*] and **Katrien Beuls**[4*]
[1]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels
[2]Itec, imec research group at KU Leuven, E. Sabbelaan 51, B-8500 Kortrijk
[3]KU Leuven, Faculty of Arts, Blijde Inkomststraat 21, B-3000 Leuven
[4]Faculté d'informatique, Université de Namur, rue Grandgagnage 21, B-5000 Namur
{jens,paul}@ai.vub.ac.be
jonas.doumen@kuleuven.be
katrien.beuls@unamur.be

Abstract

One of AI's grand challenges consists in the development of autonomous agents with communication systems offering the robustness, flexibility and adaptivity found in human languages. While the processes through which children acquire language are by now relatively well understood, a faithful computational operationalisation of the underlying mechanisms is still lacking. Two main cognitive processes are involved in child language acquisition. First, children need to reconstruct the intended meaning of observed utterances, a process called intention reading. Then, they can gradually abstract away from concrete utterances in a process called pattern finding and acquire productive schemata that generalise over form and meaning. In this paper, we introduce a mechanistic model of the intention reading process and its integration with pattern finding capacities. Concretely, we present an agent-based simulation in which an agent learns a grammar that enables them to ask and answer questions about a scene. This involves the reconstruction of queries that correspond to observed questions based on the answer and scene alone, and the generalization of linguistic schemata based on these reconstructed question-query pairs. The result is a productive grammar which can be used to map between natural language questions and queries without ever having observed the queries.

1 Introduction

Language is a unique hallmark of human intelligence. Our linguistic systems do not only meticulously serve our communicative needs, they are also incredibly robust to noise, adaptive to change and they can be learnt efficiently. While the processes that drive language acquisition in children are by now relatively well understood, a faithful computational operationalisation of the underlying mechanisms is still lacking. Having such a mechanistic model would, however, constitute a crucial step towards the development of truly intelligent agents in the field of artificial intelligence (Mikolov et al., 2016; Lake et al., 2019).

The idea that children acquire language by actively participating in communicative interactions and making use of general cognitive capacities has been elaborately documented in studies on usage-based language acquisition (Bybee, 2013; Ellis et al., 2016; Ellis and Ogden, 2017). In particular, two highly complementary cognitive processes have been identified to play a key role: *intention reading* and *pattern finding* (Tomasello, 2003, 2009). First, children need to understand the communicative intentions of their interlocutors. In a process called intention reading, they reconstruct the intended meaning of the utterances they observe. Then, they can gradually abstract away from concrete utterances and meaning representations in a process called pattern finding, and acquire productive schemata that generalise over form and meaning. Theoretical as well as empirical evidence has been abundantly provided for both intention reading (Bruner, 1983; Sperber and Wilson, 1986; Meltzoff, 1995; Nelson, 1998) and pattern finding (Goldberg, 1995; Croft, 2000; Diessel, 2004; Goldberg, 2006)

In this paper, we introduce a mechanistic model of the intention reading process and its integration with pattern finding capacities. Concretely, we present an agent-based simulation in which an artificial agent learns a construction grammar that enables it to ask and answer questions about scenes it observes. The learning task thus involves the reconstruction of queries that correspond to observed questions based on the answer and scene alone, as well as the generalization of linguistic schemata based on these reconstructed question-query pairs. The learner gradually acquires a fully productive grammar, consisting of form-meaning mappings, that can be used for both language comprehension, i.e. observing a question and mapping it onto a

[*]Shared last authors.

query, and language production, i.e. expressing a query in the form of an interrogative linguistic expression.

When it comes to intention reading, the learning challenge amounts to the reconstruction of a query based on a question-answer pair and a scene, without ever observing the query itself. The learner agent is endowed with an inventory of primitive operations, which it can combine to compose new queries. The query composition process allows the agent to hypothesize about the meaning of a question given the scene and the answer to the question. The space of all possible queries that lead to the observed answer in the given scene is typically very large. At the same time, most of these queries are not adequate representations of the meaning of the question and only lead to the correct answer in this specific scene.

The second challenge is to learn abstract schemata. Pairing an observed utterance with its reconstructed meaning yields a form-meaning mapping, called a construction (Fillmore, 1988). Initially, the learner has no way of knowing which parts of the form correspond to which parts of the meaning. Hence, it stores this mapping holistically. Through the observation of different form-meaning mappings, pattern finding mechanisms allow the agent to generalise over reoccurring form-meaning patterns, thereby capturing the compositional structure of the language.

Intention reading and pattern finding are highly complementary. Specifically, intention reading facilitates pattern finding by providing meaning hypotheses. In turn, pattern finding constrains the search process involved in intention reading by providing partial analyses. This interplay between intention reading and pattern finding is key in successfully tackling the learning challenge, and constitutes the main contribution of this paper.

We validate our methodology using the CLEVR benchmark dataset (Johnson et al., 2017a), in which the communicative task of the agents consists in asking and answering questions about scenes of geometric figures. Over many interactions, the learner incrementally acquires a fully operational grammar that can be used for both language comprehension and production. We show that the acquired grammar effectively solves the visual question answering task.

The contributions of this work are both theoretical and practical. On the theoretical side, the presented work provides computational evidence for the cognitive plausibility of usage-based theories of language acquisition, in particular concerning intention reading and pattern finding. On the practical side, this paper introduces a powerful new methodology that allows autonomous agents to acquire an effective communication system with human-like properties through task-oriented interactions in their native environment.

The remainder of this paper is structured as follows. Section 2 introduces the dataset. In Section 3, we introduce the technical foundations of our methodology. Section 4 describes the experimental setup for learning construction grammars through communicative interactions. Section 5 presents the experimental results. Related work is discussed in Section 6. Finally, Section 7 reflects on the results and contributions of our work. The code of this experiment is made available through the open-source Babel toolkit[1] (Steels and Loetzsch, 2010; Nevens et al., 2019b).

2 Data

The CLEVR dataset (Johnson et al., 2017a) consists of (i) rendered scenes with geometric figures of various shapes, sizes, colours and materials, (ii) English questions about these scenes, and (iii) answers to these questions. The questions test a variety of reasoning skills, including attribute identification (*"What size is the yellow cube?"*), counting (*"How many large cylinders are there?"*), existence (*"Is there a red ball?"*), comparison (*"Are there more spheres than cylinders?"*), spatial relationships (*"What shape is the thing right of the purple cube?"*) and logical operations (*"How many things are either spheres or cylinders?"*).

This dataset was chosen because it satisfies two criteria. First, it offers visually grounded linguistic expressions. The objects about which the agents communicate are actual referents in the agents' environment. Second, it contains a large number of scenes and plenty of similar, yet non-identical questions. Such examples are necessary for any kind of generalisation process and are consistent with theoretical and empirical evidence of how children learn language (Tomasello, 2003; Tamminen et al., 2015). Other datasets that fit these two criteria could also be used.

For the experiment in this paper, we have selected a subset of the CLEVR questions. Specif-

[1] `https://emergent-languages.org`

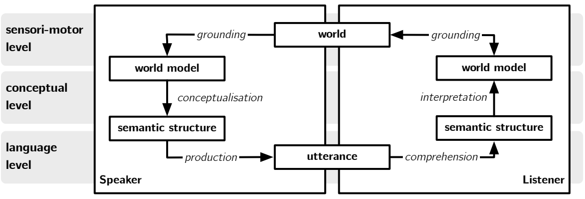


Figure 1: Processes involved in a language game.

ically, questions concerning comparison, spatial relations and logical operations have been left out. The main reason for this is that these are more complex cognitive operations that correspond to longer and more complex questions. Such questions are far removed from the type that children are faced with. Starting from CLEVR's validation split, our final dataset consists of 10,044 unique questions. Each question can be used in any of the 15,000 scenes of the validation split.

3 Technical Foundations

Our methodology builds on three main technical foundations: (i) the language game paradigm (Section 3.1), (ii) procedural cognitive semantics (Section 3.2) and (iii) computational construction grammar (Section 3.3).

3.1 The Language Game Paradigm

The *language game* paradigm (Steels, 1995, 2001) studies how linguistic conventions arise in a population of agents through local interactions and coordination. Every interaction, or language game, takes place between two agents, called the *speaker* and the *listener*, and models a particular communicative task, e.g. drawing the attention to an object in the environment. The semiotic cycle (Steels, 2012) in Figure 1 provides a schematic overview of the processes involved for the speaker and the listener. These processes take place across three different levels: the sensorimotor level, the conceptual level and the language level. In the following sections, we elaborate on the technical foundations of the processes taking place on the conceptual level (Section 3.2) and on the language level (Section 3.3). In Section 4, we concretely describe how the various processes in the semiotic cycle have been implemented in terms of these technical foundations in order to operationalise the mechanistic model of intention reading and its integration with pattern finding.

3.2 Procedural Cognitive Semantics

Incremental Recruitment Language (IRL) (Van den Broeck, 2008; Spranger et al., 2012) operationalises key insights from procedural cognitive semantics (Woods, 1968; Winograd, 1972; Johnson-Laird, 1977). Specifically, it treats the meaning of natural language utterances as programs that can be executed algorithmically in terms of the agents' representation of the environment, i.e. its world model. Such programs capture the logical structure underlying utterances in the form of *semantic networks*. An example semantic network is shown in Figure 2. The symbols preceded by question marks, as in ?OBJECT-1, are logic variables. Semantic networks are made up of predicates that are declaratively combined by sharing variables. The predicates in semantic networks represent either *semantic entities* or *primitive operations*.

Semantic entities are concepts known by the agent. They are introduced in semantic networks through BIND statements, as in (BIND SHAPE ?SHAPE-1 CUBE), binding the concept CUBE of type SHAPE to the variable ?SHAPE-1. In this experiment, a repertoire of semantic entities is given a priori to the agents. This repertoire includes the various colours, shapes, sizes and materials present in the CLEVR dataset. However, these concepts can also be learned through communicative interactions, e.g. as in Nevens et al. (2020).

Primitive operations represent the basic cognitive capabilities of the agent. In this experiment, six operations are made available. These are based on annotations provided with the CLEVR dataset. Primitive operations are implemented as multidirectional predicates with typed arguments that operate over the agents' world model and semantic repertoire. From the argument(s) that are bound, e.g. via a BIND statement or via the output of other predicates, a predicate can compute new bindings

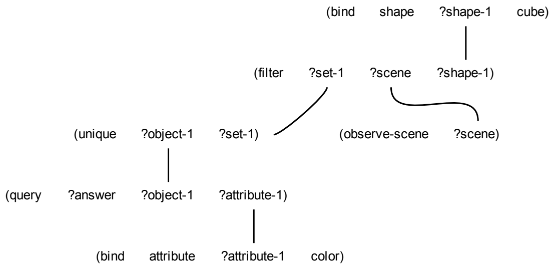

(bind shape ?shape-1 cube)
(filter ?set-1 ?scene ?shape-1)
(unique ?object-1 ?set-1)
(observe-scene ?scene)
(query ?answer ?object-1 ?attribute-1)
(bind attribute ?attribute-1 color)

Figure 2: Semantic network for the question *"What color is the cube?"*.

for the unbound argument(s).

IRL provides the computational architecture for (i) automatically composing predicates into programs and (ii) evaluating programs in terms of data structures that represent the agents' environment. The composition of predicate networks is a combinatorial search process where predicates are added incrementally and linked together by unifying their variable arguments until a communicative goal is reached. Type information of the arguments is used to determine which arguments of predicates can be linked. The evaluation of semantic networks consists in finding values, i.e. are concrete referents in the environment or the agent's semantic repertoire, for all variables in the network. A variable-value pairing is called a binding. Every semantic network has exactly one *target variable* of which the binding holds the communicative goal or intention of the corresponding utterance. To illustrate, the evaluation of the semantic network in Figure 2 goes as follows. The predicate OBSERVE-SCENE retrieves the set of all objects in the scene and binds this to the variable ?SCENE. The FILTER predicate accesses this set via the shared variable ?SCENE, together with the shape CUBE via ?SHAPE-1. The predicate then filters this set such that only cubes remain and the result is bound to ?SET-1. Next, the UNIQUE predicate checks if ?SET-1 contains a single object. If so, that object is bound to ?OBJECT-1. Finally, the QUERY predicate retrieves the color of ?OBJECT-1 and binds the result to ?ANSWER. The binding of this variable is indeed the answer to the question. Other primitive operations that are available are EXIST, which checks whether the cardinality of a set is greater than zero, and COUNT, which computes the cardinality of a set.

3.3 Computational Construction Grammar

The agents' language comprehension and production capabilities are operationalised using Fluid Construction Grammar[2] (FCG – Steels, 2017; van Trijp et al., 2022). FCG is a computational operationalisation of the basic tenets of construction grammar (Fillmore, 1988; Goldberg, 1995; Kay and Fillmore, 1999; Croft, 2001) and supports bidirectional construction-based language processing.

Corresponding to different stages of child language acquisition (Tomasello, 2003), we consider three types of constructions in this experiment:

Holophrase constructions constitute a holistic mapping between an entire form and its entire meaning representation. For example, a mapping between the question *"What color is the cube?"* and the semantic network shown in Figure 2 would be a holophrase construction.

Item-based constructions are generalisations over holophrase constructions that capture their similarities and differences, both with respect to form and meaning. For example, a construction associating the form '*What is the color of the ?X?*' with its meaning of querying the color of the referent of ?X would be an item-based generalisation over '*What is the color of the cube?*' with its meaning of querying the color of the cube and '*What is the color of the sphere?*' with its meaning of querying the color of the sphere.

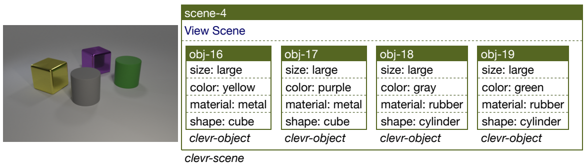
Lexical constructions provide arguments that can fill slots in item-based constructions. For example, the form *"cube"* associated with its meaning of filtering for the prototype of the concept CUBE can fill the ?X slot in the item-based construction covering '*What is the color of the ?X?*'.

Holophrase constructions only allow to comprehend the exact same utterance or produce the exact same meaning as the observation it was learnt from. Item-based constructions, on the other hand, cover a wider range of utterances and meanings through their slots, but require lexical constructions for filling those slots. When item-based and lexical constructions combine, the lexical arguments are inserted into the item-based slots, resulting in a complete utterance and a complete semantic network. The possible combinations of slots and arguments emerge through language use (Pine and Lieven, 1997; Croft, 2001). In FCG, these combinations are captured as links in a dynamically updated network, called the categorial network (Van Eecke, 2018; Steels et al., 2022). This network is consulted during constructional language processing. Hence, the links in this network determine which item-based and lexical constructions can combine.

4 The Elicitation Game

We set up a language game in a tutor-learner scenario, which we call the *elicitation game*. The agents are situated in scenes from the CLEVR dataset. The tutor is an agent that has an established linguistic inventory which allows to comprehend and produce all questions from the CLEVR

[2] https://www.fcg-net.org



scene-4

View Scene

obj-16
size: large
color: yellow
material: metal
shape: cube
clevr-object

obj-17
size: large
color: purple
material: metal
shape: cube
clevr-object

obj-18
size: large
color: gray
material: rubber
shape: cylinder
clevr-object

obj-19
size: large
color: green
material: rubber
shape: cylinder
clevr-object

clevr-scene

Figure 3: The agents are situated in scenes from the CLEVR dataset (left). These are represented symbolically (right).

dataset (cf. the grammar presented in Nevens et al. (2019a)). The learner starts with an empty construction inventory, but is endowed with the six primitive operations described above, a repertoire of semantic entities, and a number of learning mechanisms. Both the tutor and the learner can take on the discourse roles of speaker and listener. The communicative task of the elicitation game is the following. The speaker has a concept in mind and wants to elicit that concept from the listener. Therefore, it has to come up with a question about the objects in the scene to which the listener should provide the answer. The game succeeds if the listener's answer refers to the concept the speaker had in mind. Following the semiotic cycle (Figure 1), the interactions consist of the following steps:

1. Both agents perceive a randomly selected scene. Scenes are represented symbolically, as shown in Figure 3.

2. Each agent is randomly assigned a discourse role: speaker or listener.

3. The speaker selects a concept from the CLEVR dataset. This can be an object attribute (colour, size, material or shape), a number between 0 and 10, 'yes' or 'no'.

4. The speaker tries to come up with a question that has the chosen concept as its answer. This process involves two steps, namely *conceptualising* a semantic network and *producing* a question that expresses this meaning representation.

5. The listener observes the question produced by the speaker.

6. The listener tries to answer the question. This process also involves two steps, namely *comprehending* the question, i.e. mapping it onto a meaning representation, and *interpreting* this meaning representation in the current scene to come up with the answer.

7. The speaker checks whether the listener replies with the answer it had in mind. This determines the outcome of the game: success or failure.

8. If the game fails, the tutor provides feedback to the learner. Specifically, it reveals the correct answer to the question that was asked. This is a learning opportunity for the learner agent.

In the following sections, the processes of conceptualisation, production, comprehension and interpretation are discussed in detail. Afterwards, the learning mechanisms operationalising intention reading and pattern finding are introduced.

4.1 Conceptualisation

Conceptualisation is performed by the speaker to come up with the query it wants to ask. The speaker uses its own inventory of primitive operations to compose a semantic network such that the evaluation of that network, i.e. the answer to the constructed query, results in the concept the speaker wants to elicit.

4.2 Production

In production, the speaker uses its own inventory of form-meaning mappings, or constructions, to map the semantic network composed in the previous step to a natural language utterance, in this case a question. The tutor can express all valid semantic networks. When acting as the speaker, the learner will try to use its acquired holophrase, item-based and lexical constructions to express the semantic network. However, the learner's construction inventory may be inadequate for performing this mapping, causing the interaction to fail.

4.3 Comprehension

Comprehension is the inverse process of production. The listener uses its own construction inventory to try and map the observed utterance, in this case a question, to its underlying meaning representation. When acting as the listener, the learner's construction inventory may be inadequate for performing this mapping, causing the interaction to proceed with a blank answer.

4.4 Interpretation

Interpretation is performed by the listener to compute its hypothesis about the answer to the question.

This is done by evaluating the semantic network that results from comprehension. The listener's hypothesis is the value of the target variable of that semantic network.

4.5 Learning Mechanisms

Learning takes place when the interaction has failed, i.e. when the learner acting as the listener cannot retrieve the meaning underlying an observed question or the applied form-meaning mappings result in an incorrect hypothesis for the answer. The outcome of the learning mechanisms is to make new form-meaning mapping(s), through intention reading and pattern finding, in order to be more successful in future interactions.

Intention reading is performed by the learner to reconstruct a hypothesis of the meaning underlying the observed question. Similar to conceptualisation, this is done by composing a semantic network. The goal of the composition process is to construct a semantic network leading to the tutor's intention, i.e. the answer that was revealed at the end of the interaction. Crucially, the number of possible semantic networks that lead to the provided answer in the current scene is typically very large, and most of those networks will not be adequate representations of the meaning of the question. The problem faced by intention reading is thus twofold. First, the agent needs to overcome the enormous search space of possible semantic networks. Second, the agent needs to overcome incorrect meaning hypotheses.

Pattern finding allows the learner to generalise over reoccurring patterns on both the form side, which can be observed, and the meaning side, which is reconstructed through intention reading. The goal is not to learn holophrase constructions for every observation, but to learn more general item-based and lexical constructions that cover multiple observations, including novel ones. Given that both the form side and the meaning side of constructions are represented as sets of predicates, set difference operations that use unification to compare the elements are used to find the overlapping and non-overlapping parts.

The learner is endowed with five learning mechanisms that operationalise intention reading and pattern finding. These mechanisms are active in the inverse order of their presentation below.

Learning holophrases At the start of the experiment, the learner's construction inventory is empty. When it observes novel utterances, the only thing it can do is to create holophrase constructions. Specifically, the meaning is hypothesised through intention reading and paired with the observed utterance. Holophrase constructions form the basis of the learning process. Other learning operators will generalise over them.

Generalising over holophrases Whenever possible, pattern finding will compare newly created mappings between observed utterances and their reconstructed meanings against previously acquired holophrase constructions. When a minimal difference is found on both the form side and the meaning side, a generalisation can be learned. On the form side, a minimal difference refers to a single token, while on the meaning side, this is a single predicate. An item-based construction will capture the overlapping parts of the form and meaning, while a lexical construction captures the non-overlapping parts. A link is added to the categorial network indicating that the arguments of the lexical construction are suitable for filling the item-based slots on the form side and the meaning side. Three cases of this learning mechanism can be identified: (i) the new form-meaning pairing differs from the holophrase construction by substituting a minimal difference, (ii) the new form-meaning pairing extends the holophrase construction by a minimal difference, and (iii) the new form-meaning pairing reduces the holophrase construction by a minimal difference.

Learning from partial meanings This learning mechanism creates new constructions that can combine with existing constructions to analyse the observed utterance. Concretely, the acquisition of item-based and lexical constructions can lead to the partial comprehension of novel utterances. The resulting partial meaning is used by intention reading to hypothesise about the meaning underlying the observed question. Crucially, the partial meaning drastically reduces the search space faced by intention reading, as large parts of the search space that do not contain this partial meaning can be pruned. This is how the interplay between intention reading and pattern finding allows to overcome the intractability of the intention reading process. Three cases of this learning mechanism exist. First, partial meaning provided by one or more lexical constructions results in an item-based construction with an equal number of slots. This case is illus-

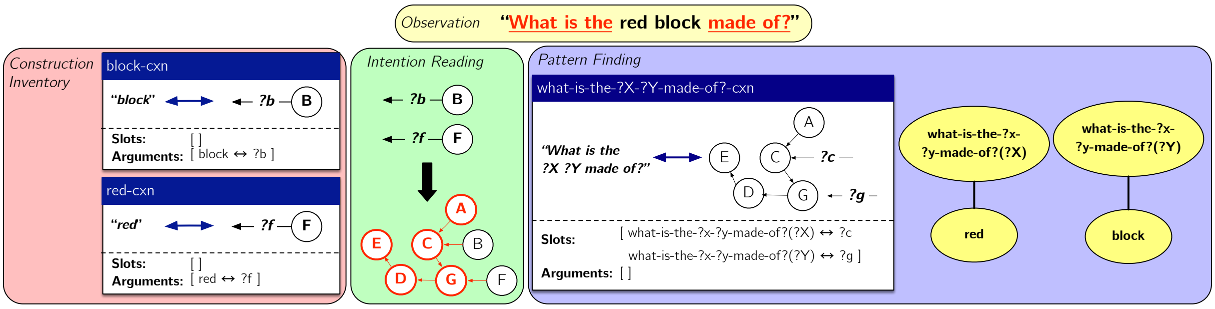

Observation
"What is the red block made of?"
Construction Inventory
block-cxn
"block"
?b
B
Slots: [ ]
Arguments: [block ↔ ?b]
red-cxn
"red"
?f
F
Slots: [ ]
Arguments: [red ↔ ?f]
Intention Reading
?b
B
?f
F
A
E
C
B
D
G
F
Pattern Finding
what-is-the-?X-?Y-made-of?-cxn
"What is the ?X ?Y made of?"
A
E
C
?c
D
G
?g
Slots: [what-is-the-?x-?y-made-of?(?X) ↔ ?c
what-is-the-?x-?y-made-of?(?Y) ↔ ?g]
Arguments: [ ]
what-is-the-?x-?y-made-of?(?X)
what-is-the-?x-?y-made-of?(?Y)
red
block

Figure 4: Schematic representation of the learning mechanism that completes a partial meaning provided by lexical constructions. The RED-CXN and BLOCK-CXN provide a partial meaning (red box) for the observed utterance (yellow box). Intention reading creates a meaning hypothesis, reusing the partial meaning (green box). Pattern finding creates an item-based construction with two slots and corresponding links in the categorial network (blue box).

trated in Figure 4. Second, partial meaning provided by an item-based construction leads to a single lexical construction. If multiple lexical items are missing, there would be referential uncertainty, which is not explored here. Third, partial meaning provided by a combination of one or more lexical constructions and an item-based construction also leads to maximally one lexical construction. In all three cases, slot-and-argument links are also added to the categorial network.

Learning slot-argument links This learning mechanism handles cases where previously acquired item-based and lexical constructions cover the observed utterance, but where the slot-argument combination has not been observed before. Due to the absence of that link in the categorial network, the corresponding constructions cannot combine, causing comprehension or production to fail. In comprehension, the learner simply adds the slot-argument combination it observed. In production, the learner creatively tries out slot-argument combinations. However, these links are only consolidated when the interaction turns out to be successful. Note that intention reading is not required for this learning mechanism.

Lateral inhibition Lateral inhibition facilitates the self-organisation of the learner's construction inventory (Steels, 1995). It is used at the end of every interaction, including successful ones. Concretely, it models the entrenchment of constructions (Schmid, 2007) by updating their scores. New constructions obtain a default score of 0.5. Scores are bound between 0 and 1.

The scores of constructions are updated based on the outcome of the game. If the game fails, the scores of the constructions used during the game are decreased by 0.4. These constructions were inadequate for the communicative task and should therefore be used less often in the future. If the game succeeds, the scores are increased by 0.1, while scores of competing constructions are decreased by 0.1. Competing constructions are constructions that also could have contributed to the comprehension or production process. When reaching a score of 0, constructions are removed from the construction inventory. The exact values used to alter construction scores do not influence the global dynamics of the learning process, as long as these values are positive and negative respectively.

The presented learning mechanisms do not posit a built-in bias towards more abstract constructions. However, given that more abstract constructions are inherently applicable in a wider range of situations, they will therefore be used more frequently. As a result of lateral inhibition, this will result in higher entrenchment scores for more abstract constructions and in lower scores for less abstract constructions. By updating scores of constructions through lateral inhibition and by preferring constructions with a higher score during comprehension and production, a positive feedback loop is created between the success and use of constructions. This feedback loop ensures that only constructions that can be used successfully in the communicative task remain, while unsuccessful constructions gradually disappear. This way, incorrect meaning hypotheses generated by intention reading can be overcome.

5 Experimental Results

This section presents the validation of our methodology on the CLEVR data. The presented results

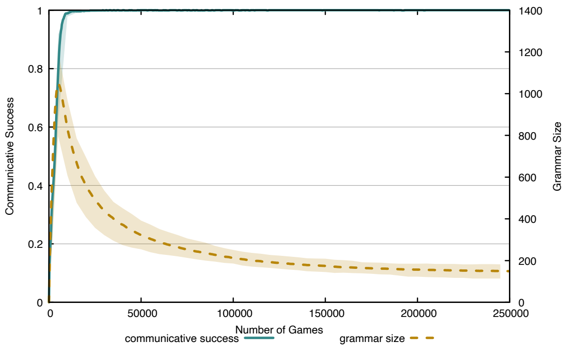


Figure 5: Evolution of communicative success (left axis) and grammar size (right axis) over time.

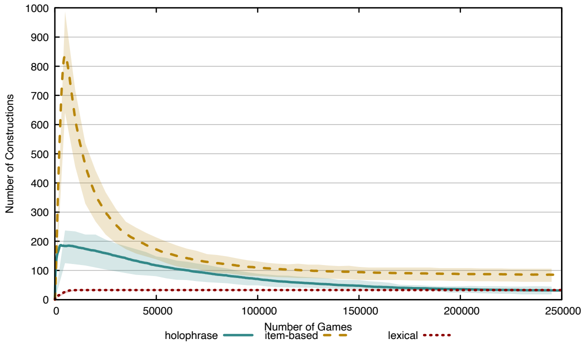


Figure 6: Evolution of the number of constructions over time, split per type.

are based on ten independent runs of 250,000 interactions each. The filled areas around the lines on the plots represent the 5th and 95th percentile.

Figure 5 presents the main results of the experiment. It shows the *communicative success* and the *grammar size* over time. Both metrics start at 0. The communicative success rises quickly, with more than 78% of the interactions being successful after 5,000 games. This is half the number of possible utterances from our subset of data. The grammar size also reaches its peak at this point, with on average 1,048 constructions being learned. Success reaches over 98% after 10,000 interactions and does not go below 99.9% from interaction 25,000 onwards. It is only after 200,000 interactions that the success reaches a stable 100%. This is because it takes a long time for all incorrect constructions to be cleared from the construction inventory. Specifically, the learner needs to observe just the right question in just the right scene to find out that a previously acquired form-meaning mapping is incorrect. The grammar size gradually decreases, reaching 492 constructions after 25,000 interactions and 149 constructions by the end of the experiment.

Figure 6 breaks down the grammar size per construction type. At the start of the experiment, only holophrase constructions are learned. Soon after, the learner can start to generalise over them. The construction inventory peaks after 5,000 interactions, reaching 184 holophrase constructions and 838 item-based constructions. Afterwards, more abstract item-based constructions gradually become dominant and overtake their less abstract competitors, including the holophrase constructions. By the end of the experiment, 31 holophrase constructions and 85 item-based constructions remain. There is less competition among the lexical constructions. After 10,000 interactions, 33 lexical constructions are learned and this remains stable until the end. The theoretical maximum of 35 lexical constructions was reached in four out of ten experimental runs. After 10,000 interactions, 33 lexical constructions are learned and this remains stable until the end. We note that it is *not* the goal of the experiment to reach one particular set of constructions. Hence, the absolute number of constructions at a given time is not important. Instead, the goal is to become successful at the communicative task and learn an efficient construction inventory for doing so.

6 Related Work

Prior agent-based models have also studied the constructivist co-acquisition of syntax and semantics through task-oriented interactions (Gerasymova and Spranger, 2010; Beuls et al., 2010; Spranger and Steels, 2015). While these models have provided important insights, this paper advances the state of the art in two ways. First, the presented experiment operates on a much larger scale. In contrast to prior models, this work does not focus on a specific linguistic phenomenon, such as the Russian aspectual system (Gerasymova and Spranger, 2010), the Hungarian agreement system (Beuls et al., 2010) or English spatial language (Spranger and Steels, 2015). The utterances being considered in this work are far more complex, both in terms of morpho-syntax and semantics. Second, the presented experiment provides fewer scaffolds. The agent does not receive a segmentation of input utterances, a predefined lexicon as in Beuls et al. (2010) or a taxonomy guiding the generalisation process of constructions as in Spranger and Steels (2015). The agent only relies on a number of basic cognitive operations and a collection of concepts. The latter can also be learned from sub-symbolic observations through the language game methodology, as shown by Nevens et al. (2020).

When it comes to the field of visual question answering, existing approaches typically tackle the task in one of two ways. Either, a large end-to-end neural network maps an image and a question to the answer, e.g. as in Malinowski et al. (2015). Alternatively, RNNs are used to map the question onto a query which is then executed on the image, e.g. as in Johnson et al. (2017b). Both of these approaches rely on huge amounts of training data. The second approach additionally requires questions annotated with queries to train the RNN. Further, both approaches rely on black-box architectures, making it unclear how and why an answer was generated. Our methodology overcomes these shortcomings. Similar to the first approach, the agent is only presented with images, questions and their answers and autonomously reconstructs the underlying queries. Our methodology is more data-efficient and the agents' representations and reasoning processes are fully transparent. Finally, the agents' communication system is open-ended and completely bidirectional using the same representations and processing mechanisms, which is not possible using current neural network architectures.

7 Concluding Discussion

The contributions of this paper span two areas.

Usage-based Language Acquisition The experiment presented in this paper provides computational evidence for the cognitive plausibility of theories from usage-based language acquisition, in particular intention reading and pattern finding (Tomasello, 2003). We have operationalised these capacities and their interplay in an agent-based simulation, which has indeed revealed learning dynamics that are similar to those observed in the literature. Starting from holophrases, the agent learns to generalise over them, gradually leading to more and more abstract schemata.

Autonomous Agents Most importantly, this paper pushes forward the state of the art in the development of autonomous agents with communication systems offering human-like properties. In particular, we have introduced a novel methodology that allows an agent to acquire an inventory of constructions that facilitates bi-directional language processing and is suitable for solving a communicative task. This grammar is acquired through situated, task-oriented interactions with indirect supervision only. Given only utterances, feedback on their underlying intentions and a collection of primitive cognitive operations, the agent engages in a highly non-trivial process of meaning creation, operationalised through intention reading, and combines this with a process of schema abstraction, operationalised through pattern finding. We show that the search space involved in intention reading, i.e. the composition of semantic programs, can effectively be constrained through its integration with pattern finding and that together, these processes allow the agent to bootstrap a successful communication system. The presented methodology is completely transparent, both in terms of the applied learning operators and the resulting inventory of constructions. The agent learns incrementally, acquiring productive linguistic structures even after a single interaction. The agent's grammar is open-ended and the lateral inhibition dynamics enable the agent to remain ever-adaptive, e.g. when the environment or the communicative task changes.

Acknowledgements

The research reported in this paper was funded by the Research Foundation Flanders (FWO) through grants 1SB6219N (JN) and 75929 (PVE), imec's Smart Education research programme, with support from the Flemish government (JD), and the European Union's Horizon 2020 research and innovation programme under grant agreement no. 951846 (KB).

References

Katrien Beuls, Kateryna Gerasymova, and Remi van Trijp. 2010. Situated learning through the use of language games. In *Proceedings of the 19th Annual Machine Learning Conference of Belgium and The Netherlands (BeNeLearn)*, pages 1–6.

Jerome Bruner. 1983. *Learning to use language*. Oxford University Press, Oxford.

Joan L. Bybee. 2013. Usage-based theory and exemplar representations of constructions. In Thomas Hoffmann and Graeme Trousdale, editors, *The Oxford Handbook of Construction Grammar*, pages 49—-69. Oxford University Press.

William Croft. 2000. *Explaining language change: An evolutionary approach*. Pearson Education, Harlow.

William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press, Oxford.

Holger Diessel. 2004. *The acquisition of complex sentences*. Cambridge University Press, Cambridge.

Nick C. Ellis and Dave C. Ogden. 2017. Thinking about multiword constructions: Usage-based approaches to acquisition and processing. *Topics in Cognitive Science*, 9(3):604–620.

Nick C. Ellis, Ute Römer, and Matthew Brook O'Donnell. 2016. *Usage-Based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar*. Wiley-Blackwell, Malden.

Charles J Fillmore. 1988. The mechanisms of "construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, volume 14, pages 35–55.

Kateryna Gerasymova and Michael Spranger. 2010. Acquisition of grammar in autonomous artificial systems. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI-2010)*, pages 923–928.

Adele Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, Chicago.

Adele Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press, Oxford.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017a. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017b. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2989–2998.

Philip N Johnson-Laird. 1977. Procedural semantics. *Cognition*, 5(3):189–214.

Paul Kay and Charles Fillmore. 1999. Grammatical constructions and linguistic generalizations: The what's x doing y? construction. *Language*, 75(1):1–33.

Brenden M Lake, Tal Linzen, and Marco Baroni. 2019. Human few-shot learning of compositional instructions. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 611–617.

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

Andrew Meltzoff. 1995. Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31(5):838–850.

Tomas Mikolov, Armand Joulin, and Marco Baroni. 2016. A roadmap towards machine intelligence. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, pages 29–61.

Katherine Nelson. 1998. *Language in cognitive development: The emergence of the mediated mind*. Cambridge University Press, Cambridge.

Jens Nevens, Paul Van Eecke, and Katrien Beuls. 2019a. Computational construction grammar for visual question answering. *Linguistics Vanguard*, 5(1):20180070.

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

Jens Nevens, Paul Van Eecke, and Katrien Beuls. 2020. From continuous observations to symbolic concepts: A discrimination-based strategy for grounded concept learning. *Frontiers in Robotics and AI*, 7:84.

Julian M Pine and Elena VM Lieven. 1997. Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, 18(2):123–138.

Hans-Jörg Schmid. 2007. Entrenchment, salience, and basic levels. In Dirk Geeraerts and Hubert Cuyckens, editors, *The Oxford handbook of cognitive linguistics*, pages 117–138. Oxford University Press, New York.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*. Harvard University Press, Cambridge, MA.

Michael Spranger, Simon Pauw, Martin Loetzsch, and Luc Steels. 2012. Open-ended procedural semantics. In Luc Steels and Manfred Hild, editors, *Language Grounding in Robots*, pages 153–172. Springer, New York.

Michael Spranger and Luc Steels. 2015. Co-acquisition of syntax and semantics: an investigation in spatial language. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1909–1915, Palo Alto, Ca. AAAI Press.

Luc Steels. 1995. A self-organizing spatial vocabulary. *Artificial Life*, 2(3):319–332.

Luc Steels. 2001. Language games for autonomous robots. *IEEE intelligent systems*, 16:16–22.

Luc Steels. 2012. Grounding language through evolutionary language games. In *Language Grounding in Robots*, pages 1–22. Springer, New York.

Luc Steels. 2017. Basics of Fluid Construction Grammar. *Constructions and Frames*, 9(2):178–225.

Luc Steels and Martin Loetzsch. 2010. Babel: A tool for running experiments on the evolution of language. In *Evolution of Communication and Language in Embodied Agents*, pages 307–313. Springer, Berlin.

Luc Steels, Paul Van Eecke, and Katrien Beuls. 2022. Usage-based learning of grammatical categories. *arXiv preprint arXiv:2204.10201*.

Jakke Tamminen, Matthew H. Davis, and Kathleen Rastle. 2015. From specific examples to general knowledge in language learning. *Cognitive Psychology*, 79:1–39.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Harvard.

Michael Tomasello. 2009. The usage-based theory of language acquisition. In *The Cambridge handbook of child language*, pages 69–87. Cambridge University Press, Cambridge.

Wouter Van den Broeck. 2008. Constraint based compositional semantics. In *The Evolution of Language: Proceedings of the 7th International Conference (EVOLANG7)*, pages 338–345. World Scientific.

Paul Van Eecke. 2018. *Generalisation and specialisation operators for computational construction grammar and their application in evolutionary linguistics Research*. Ph.D. thesis, Vrije Universiteit Brussel, Brussels: VUB Press.

Remi van Trijp, Katrien Beuls, and Paul Van Eecke. 2022. The FCG editor: An innovative environment for engineering computational construction grammars. *PLOS ONE*, 17(6):1–27.

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

W. A. Woods. 1968. Procedural semantics for a question-answering machine. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, pages 457–471, New York, NY, USA.