Modelling the co-emergence of linguistic constructions and action concepts: the case of action verbs

Maximilian Panzner, Judith Gaspers and Philipp Cimiano

***Abstract*—In this paper, we are concerned with understanding how linguistic and conceptual structures co-emerge, shaping and influencing each other. Most theories and models of language acquisition so far have adopted a 'mapping' paradigm according to which novel words or constructions are 'mapped' onto existing, priorly acquired or innate concepts. Departing from this mapping approach, we present a computational model of the co-emergence of linguistic and conceptual structures. We focus in particular on the case of action verbs and develop a model by which a system can learn the grounded meaning of a verbal construction without assuming the prior existence of a corresponding sensomotorically grounded action concept. Our model spells out how a learner can distill the essence of the meaning of a verbal construction as a process of incremental generalization of the meaning of action verbs, starting from a meaning that is specific to a certain situation in which the verb has been encountered. We understand the meaning of verbs as evoking a grounded simulation rather than a static concept and propose to capture the meaning of verbs via generative statistical models that support simulation, in our case Hidden Markov Models. Statistical models can represent the essence of a verb's meaning while modelling uncertainty and thus variation at the surface level of (observed) action performances. We show that by extending an existing framework for construction learning, our approach can account for the co-emergence of linguistic and conceptual structures. We provide proof-of-concept for our model by experimentally evaluating it on matching, choice and generation tasks, showing that our model can not only understand but also produce language.**

***Index Terms*—Incremental multi-modal learning, grounded learning, qualitative models of action, QTC, model merging**

I. Introduction

Linguistic and conceptual development are assumed to go hand in hand [1]. For one, it has been argued that language structures thought and shapes the concepts we acquire. This is indeed the main claim behind the theory of linguistic relativism, more widely known as the Sapir-Whorf hypothesis [2]. For another, conceptual development is also a prerequisite for language learning as linguistic constructions need to be 'mapped' to some concept that "represents" the meaning of the construction (see [3], [4]). This mapping paradigm underlies most of the work on associational language learning involving cross-situational analysis (see [5], [6], [7], [8]).

However, the detailed mechanisms that are involved in the co-emergence of linguistic and conceptual structures have not received prominent attention so far. Computational models can contribute to enhance our understanding of such processes by providing an implementable and thus explicit theory that can account for the co-emergence of cross-modal representations.

Many computational models and theories of language acquisition have so far assumed that concepts are available prior to learning the meaning of a certain construction. This simplification has been described by Lila Gleitmann as follows:

'This is a large simplification of the learning problem for vocabulary, to be sure. It's not likely that learning in this regard is always and only a matter of mapping the words heard onto a preset and immutable set of concepts priorly available to the prelinguistic child. Rather, there is bound to be some degree of interaction between the categories lexicalized in a language and the child's conceptual organization; moreover, that conceptual organization is changing during the period of vocabulary growth, to some degree affecting the nature of lexical entries...' [9].

In spite of being a simplification, most proposals for computational models of language acquisition have factored out the conceptual development dimension and focused on models explaining how systems learn to map novel words onto existing concepts. This is in fact the main assumption made in models and theories relying on cross-situational associational learning paradigms (e.g. [5], [6], [7], [8]). Exceptions exist nevertheless. The work of Roy et al. [10] for instance has proposed a model called CELL allowing a system to learn cross-modal patterns on the basis of sensory input. The model acquires a lexicon by finding consistent cross-modal patterns between sound sequences and shapes observed in images using a probabilistic model.

A crucial question that has not received prominent attention is how a learner can acquire the (grounded) meaning of verbal constructions, in particular capturing their dynamic meaning aspects, in such a way that a learner can both understand verbal constructions by simulating them but also generate verbalizations when observing a certain action, closing the loop between the different modalities.

In this paper, we propose a computational model and thus an implemented theory that accounts for the co-emergence of linguistic and conceptual structure for the case of action verbs and the (grounded) action concepts they denote. We propose a model by which a system can learn the grounded meaning of a verbal construction without assuming the prior existence of a corresponding concept. Our model spells out how a learner can distill the essence of the meaning of a verbal construction as a process of incremental generalization, starting from a meaning that is specific to a certain context in which the verb has been encountered. We understand the meaning of verbs as evoking a simulation rather than a static concept and propose to capture the meaning of verbs via generative probabilistic models, Hidden Markov Models in particular. The Hidden Markov Models represent the essence of the verb's meaning and can capture variation at the surface level to account for variation in action performance.

The model we propose is inspired by usage-based theories of language acquisition that assume that language learning proceeds from specific to general with specific constructions being incrementally generalized and entrenched, leading to different levels of generalization for different words [11]. This is empirically backed up by findings demonstrating different levels of generalization in development within the same part-of-speech, e.g. for verbs ([12]) or for determiners [13]. The level of generalization is thus word-specific rather than category-specific. We attempt to carry this idea over to the conceptual domain to yield comparable principles describing linguistic and conceptual development and how they interact with each other.

We thus apply similar principles to the domain of action in that our model also implements a usage-based approach to learning actions in the sense that actions are incrementally generalized. We build on Hidden Markov Models as representation of actions that are incrementally merged to yield more general models. We thus hypothesize that linguistic and conceptual development might rely on akin principles, i.e. the incremental merging of specific models to yield more general models or concepts. This generalization is driven by the desire of a learner to yield a compact description of these domains while not loosing too much descriptional accuracy. The first corresponds to Occam‘s razor principle and is implemented in our model as a prior that prefers simpler models. The second is implemented through a model merging procedure that merges specific models guided by the desire to yield generalizable models while at the same time maximizing the likelihood of generating the observed linguistic and action sequences under the generalized model in order to avoid over-generalization.

In our proposed model, language and concept acquisition go hand in hand in the sense that these generalizations are not applied only separately, but generalization at the linguistic level triggers a learner to look for potential generalizations of two actions observed in the context of the same (generalized) sentence. Equivalent linguistic constructions are thus expected to denote equivalent or unifiable grounded concepts. Generalization at the conceptual level forces a learner to induce near-synonym relations, that is to postulate relations between linguistic constructions that look different at the surface level, but clearly have commonalities in their meanings. This supports the acquisition of equivalence classes of linguistic constructions for which the evoked action concepts can be unified in one model.

In previous work, we have proposed a (computational) model of language learning that explains the usage-based incremental development of a construction grammar [14]. The model assumed that concepts, in particular action concepts, are already acquired. In this paper we extend the model towards explaining how linguistic constructions and action concepts are learned in interaction with each other.

To our knowledge, our model is the first model that explains how linguistic verbal constructions and the action concepts they represent co-emerge, following similar principles relying on incremental generalization driven by the desire to yield more compact models that maintain predictive accuracy. Our model spells out these mechanisms in detail and thus provides a detailed implemented theory explaining how linguistic and conceptual development go hand in hand. Further, we provide a model that can both be used to ‘understand’ but also to ‘generate’ language. Our model allows a learner both to “talk” about observed actions, being able to categorize actions and verbalize them, but also to “simulate” an action given an (input) sentence that describes the action. In this sense our model is one of the few (cognitive) models of language acquisition bringing also comprehension and generation together in the sense of Pickering and Garrod [15]. Further, it is the first model that explains how synonyms emerge as a byproduct of grouping similar action models.

Our experiments are carried out on a dataset consisting of action performances for four types of actions (jump on, jump over, circle around and push) carried out by subjects when prompted with sentences verbalizing the action in question.

To evaluate the performance of our model, we evaluate the model under three conditions: i) a matching task consisting of deciding whether a given sentence describes a given action instance, ii) a selection task consisting in selecting one out of three action instances that is described by a given sentence, and iii) a generation task consisting in generating a sentence describing a given action instance.

The paper is structured as follows: in the next Section II we present our model that accounts for the co-emergence of linguistic constructions and corresponding (action) concepts. The model builds on a previous model for the acquisition of constructions that was published before (see Gaspers et al. [14]). We describe this model for the sake of completeness and to make this paper self-contained. We then present the approach we follow for modelling action concepts using Hidden Markov Models (HMM) and the qualitative trajectory calculus (QTC) [16]. We assume that a learner is able to extract qualitative relations from the perceptual input and rely on QTC to capture such qualitative relations. We then explain how incremental generalization is performed via a model merging approach that is guided by the desire to maximize likelihood while minimizing model complexity. In Section IV we present results of our model on the three tasks mentioned above. Before concluding, we discuss implications of our work for the larger field of language acquisition in Section V.

II. Model

In this section, we describe in detail our model for accounting for the co-emergence of linguistic constructions and the action concepts they denote. In essence, the model consists of two components. One component is based on a model that was published before and that models the acquisition of syntactic constructions using symbolic meaning representations (see Gaspers et al. [14]) The second component is responsible for inducing general action concepts from specific examples of action performances following an incremental model merging

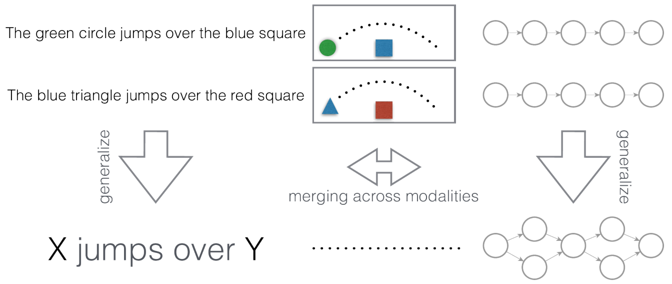

The green circle jumps over the blue square
The blue triangle jumps over the red square
generalize
merging across modalities
generalize
X jumps over Y

Fig. 1. Overview over the joint model. The model consists of a component to induce syntactic constructions and a component responsible for inducing generalized action concepts. The components take sequences of words or sequences of discrete qualitative relations between two objects as input. They create a category most specific for that particular input (top) and generalize by gradually merging specific categories into more general categories (bottom) using similarity cues from both modalities. The resulting generalized models consist of slot and frame constructions for language (bottom left) and action representations in terms of Hidden Markov Models (bottom right).

approach based on generative probabilistic models, in our case Hidden Markov Models, which were specifically chosen because the underlying network is structurally similar to the linguistic construction networks and can be learned incrementally. The observation alphabet corresponds to a set of qualitative relations that a learner is assumed to be able to recognize in a visual context. In essence, both models take as input sequences of words and qualitative relations describing the relations between a trajector and some reference object, respectively. Upon first occurrence of a certain sentence together with an action sequence, both models create a category most specific for that given sequence of words and action. Later, these most specific categories are generalized as more and more similar examples are observed, leading to entrenchment and generalization. The overall model is depicted in Figure 1. It shows two input sentences: 'The green circle jumps over the blue square.' and 'The blue triangle jumps over the red square' as well as two corresponding action sequences. Our component for learning generalized constructions from sequences of words would generate the hypothesis that both sentences can be merged into a more general construction *'X jumps over Y'*, abstracting from the specific slot fillers of the corresponding verbal construction. This mergeability of both sentences into a more general sentence would trigger our second action concept learning component to try to unify both action sequences into a more generalized action sequence in terms of a probabilistic model that still generates both sequences with high likelihood while not being overly complex.

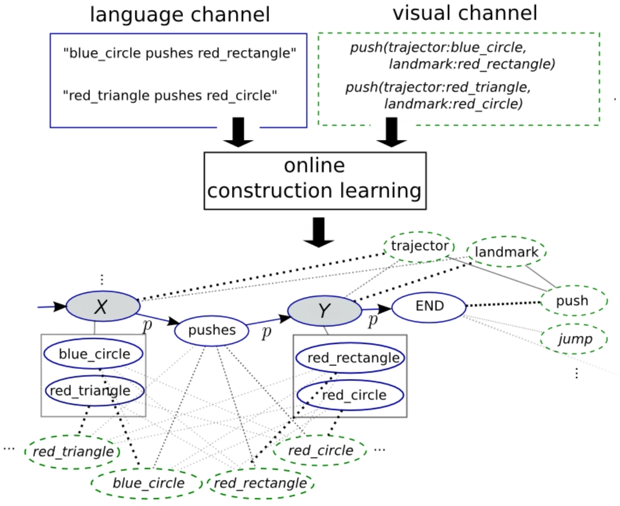
At the same time, the interaction between both components can be reversed: when two actions are regarded as being sufficiently compatible or similar to be merged into a single action concept model, the system could infer that their corresponding sentences or linguistic instructions might also be regarded as equivalent. In this way, our model can also detect synonym relations. In what follows we present the first part of our model, the part responsible for inducing generalized linguistic constructions in Section II-A. This section summarizes in a nutshell the model presented in earlier work. We refer the interested reader for details to the original model [14]. Extending this model to deal with non-symbolic representations, we present our approach for representing action concepts using Hidden Markov Models with an observation alphabet based on the Qualitative Trajectory Calculus (QTC) in Section II-B. We describe how our model induces such representations in interaction with the component for learning generalized linguistic constructions in Section II-C.

A. *Learning syntactic constructions*

The existing computational model for inducing generalized linguistic constructions acquires a lexicon and syntactic constructions from examples comprised of input across two different input channels: a language channel and a visual channel. The language channel presents sentences as sequences of characters to the system while input from the visual channel is represented as a symbolic description of the visual context. The visual meaning representation (*MR*) is comprised of a set of actions performed according to the description given in the sentence. Each action $mr_i \in MR$ is represented by means of predicate logic formulas, comprising a predicate ξ along with a set of arguments. The learning process and an example of a verb-specific construction stored in the network are shown in Figure 2.

The learned network consists of two interrelated subnetworks, the **lexical subnetwork** and the **syntactic subnetwork**, which is comprised of two sublayers, the **slot and frame layer** (**S&F**) and the **mapping layer**. The **lexical subnetwork** encodes simple phrases, i.e. (short sequences of) words along with their associated semantic referents as nodes in the network, i.e. the sequence "red triangle" and the corresponding

Fig. 2. Schematic overview showing an example construction learned from the two examples of the "pushes" action given as paired input across the language and visual channel. The figure shows the learned construction stored in the network.


language channel
visual channel
"blue_circle pushes red_rectangle"
"red_triangle pushes red_circle"
push(trajector:blue_circle, landmark:red_rectangle)
push(trajector:red_triangle, landmark:red_circle)
online construction learning
trajector
landmark
X
p
pushes
p
Y
p
END
push
jump
blue_circle
red_triangle
red_rectangle
red_circle
red_triangle
red_circle
blue_circle
red_rectangle

semantic referent *'red_triangle'* in Fig 2. The **S&F pattern layer** represents syntactic constructions as sequences of nodes that together constitute a path. Paths can contain variable nodes that represent slots in the syntactic pattern. These slots can be filled with elements contained in specific groupings. This layer also encodes the associated semantic frames. For instance, in Fig. 2, a syntactic construction is represented as a path p which expresses a pattern "X pushes Y", where X and Y represent syntactic slots in the pattern, which can be filled with groupings of elements such as "blue circle" and "red triangle" in the case of X or "red circle" and "red rectangle" in the case of Y. The semantic frame associated with the pattern is *push(trajector,landmark)*. The **mapping layer** contains networks representing construction-specific argument mappings between syntactic patterns and semantic frames together with mappings of the syntactic arguments to semantic arguments. For example, in Fig. 2 an individual mapping network captures the correspondences between X and the *trajector* role as well as Y and the *landmark* role. Form-meaning mappings as correspondences between linguistic and semantic entities, are established by capturing their co-occurrence frequencies across observed examples/situations in associative networks [17] (see Gaspers et al. [14] for details).

Learning is organized in an online fashion where each input example causes immediate changes in the network structure. The learning process is roughly divided into two steps: i) update of the lexical layer, where connections between lexical units and semantic referents are established and reinforced, and ii) update of the construction layer, where the model mainly attempts to merge paths, and thus generalizes over specific linguistic and action examples observed. For generalization, the model exploits type variations at the linguistic level in relation to semantic observations. More specifically, there are two different generalization steps, both of which are applied to each observed input example, i.e. i) a slot-driven generalization step and ii) a syntactic generalization step. In the slot-driven generalization step, the model searches for sentences and (partially generalized) patterns for which linguistic variation in a position yields corresponding semantic variation in a slot in an associated semantic frame. In the syntactic generalization step, the model searches for patterns which show linguistic variation in a position but are associated with the same semantic frame. Thus, syntactic generalization may yield groupings of lexical units which are synonyms.

To illustrate the intuition behind the learning steps, consider the following example: A learner observes "the blue circle jumps" and "the red triangle jumps" in the visual context *jump(trajector:blue_circle)* and *jump(trajector:red_triangle)*, respectively. To learn across situations, during updates of the lexical layer, the model would use its knowledge that the linguistic phrase "red triangle" refers to the semantic entity *red_triangle* and that the phrase "blue circle" refers to the semantic entity *blue_circle*. Such knowledge would, in turn, be applied during updates of the construction layer in the slot-driven generalization step to learn that the type variation in the sentences' first position ("blue circle" vs. "red triangle") reflects the meaning difference in the *trajector* role of *jump*. The model would use its knowledge to acquire the more general pattern shown in (1), where X = [blue circle $\rightarrow blue\_circle$, red triangle $\rightarrow red\_triangle$].

(1)

Syntactic pattern	X jumps
Semantic frame	*jump(trajector)*
Mapping	$X \rightarrow$ *trajector*

Now let's assume that after observation of some more input examples the model has also acquired the constructions shown in (2), where again X = [red triangle $\rightarrow red\_triangle$, blue circle $\rightarrow blue\_circle$].

(2)

Syntactic pattern	X hops
Semantic frame	*jump(trajector)*
Mapping	$X \rightarrow$ *trajector*

Since the two syntactic patterns show linguistic variation in one position ("jumps" vs. "hops"), but are associated with the same semantic frame, the model would group these two words and assume that both can be used interchangeably (without yielding semantic change). The model would thus use its knowledge to acquire the more general pattern shown in (3), where X = [red triangle $\rightarrow red\_triangle$, blue circle $\rightarrow blue\_circle$] and *SYN*$_1$ = [jumps, hops].

(3)

Syntactic pattern	X *SYN*$_1$
Semantic frame	*jump(trajector)*
Mapping	$X \rightarrow$ *trajector*

B. Action models

This section describes how action performances are represented as Hidden Markov Models (HMM) over sequences of qualitative relations between a trajector and a landmark expressed in the Qualitative Trajectory Calculus (QTC). We focus on actions in which some trajector moves or is moved relative to some landmark or ground. We assume that a system is able to observe qualitative relations that describe the relation between a moving trajector relative to a given landmark. Our Hidden Markov Models in essence thus model the action specific probability of a given sequence of qualitative relations describing the relation between a trajector and a landmark over time.

In order to describe the relative position and movement between landmark and trajector, we build on the qualitative trajectory calculus - double cross (QTC_{C1}) [16] as a formal foundation. In general, the QTC family of representations describes the interaction between two moving point objects k and l with respect to the reference line RL that connects them at a specific point t in time. As we only have one actively moved object in our experiments, we decided on QTC_{C1} among the QTC family of representations to give the best trade off between generalization and specificity of the qualitative relations. The QTC_{C1} framework defines a 4-element state descriptor (C_1, C_2, C_3, C_4) where each $C_i \in \{-, 0, +\}$ represents a so called constraint with the following interpretation:

C_1 Distance constraint: Movement of k with respect to l at time t_1:

- - k is moving towards l
- 0 k is not moving relative to l
- + k is moving away from l

C_2 Distance constraint: Movement of l with respect to k at time t_1: analogously to C_1

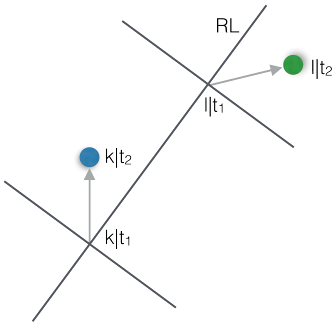


Fig. 3. This figure shows two moving objects k and l at two different time points t_1 and t_2. In this example k is moving towards l at time t_1 on the left hand side of the reference line RL, l is moving away from k on the left hand side of the reference line from l to k. The corresponding QTC_{C1} relation is $(-+--)$. Reproduced from [16].

C_3 Side constraint: Movement of k with respect to RL at time t_1:
- - k is moving to the left-hand side of RL
- 0 k is moving along RL or not moving at all
- + k is moving to the right-hand side of RL

C_4 Side constraint: Movement of l with respect to RL at time t_1: analogously to C_3

According to the above definition, QTC_{C1} defines a total of $3^4 = 81$ different basic relations. The framework provides a rather coarse discretization of the relations between two objects, leading to situations where the qualitative relation between the two objects holds for a longer portion of the trajectory. As these parts of the trajectory do not carry much discriminative information, we apply logarithmic compression of repetitive subsequences as described by Panzner et al. [18], which allows to preserve information about the acceleration along the trajectory, increasing the overall performance especially for very similar actions like "jumps over" and "jumps upon", while still allowing to generalize over high variations in relative pace of the action performances. As an illustration of our action representation consider Figure 4, which depicts a rectangle circling once around a circle on an elliptic trajectory. At the first marked position, P_1, the square is moving on the top left of the circle, corresponding to the QTC descriptor (-,0,-,0). At P_2, the square is on top of the circle and instead of approaching the circle the rectangle veers away from the circle now, resulting in the first constraint of the QTC relation to change from $-$ to $+$ yielding (+,0,-,0) as the new relation. In this very smooth trajectory the QTC relations would only change at the positions P_2, P_4, P_6 and remain unchanged in between, leading to subsequences with many repeated QTC relations in between which are subject to the logarithmic compression. Trajectories from actions performed by humans however are much more cluttered, so that the sequences of QTC relation contain many additional transitions.

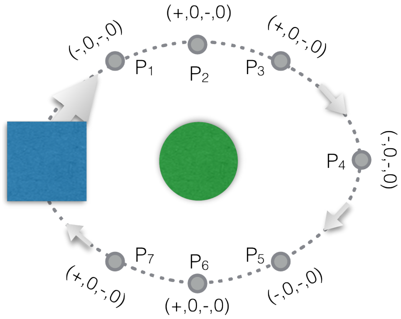


Fig. 4. Example sequence of a "circles around" action. The blue rectangle circles around the green circle on a smooth elliptic trajectory. The QTC relations only change at the marked positions P_2, P_4, P_6 and remain unchanged in between.

C. Induction of action models

In our approach, induction of generalized action models is performed by incrementally merging specific HMMs into more general HMMs that have a higher entropy compared to the very specific HMMs. At the same time, the generalized HMMs should still assign substantial probability mass to the observed example while minimizing model complexity. Our incremental model merging approach follows the approach described by Omohundro et al. [19] and is inspired by the observation that when faced with new situations, humans and animals alike drive their learning process by first storing individual examples (memory based learning) where few data points are available and gradually switching to a parametric learning scheme to allow for better generalization as more and more data becomes available [20]. Our approach mimics this behavior by starting with simple models generating exactly one sequence which evolve into more complex models as more data becomes available. Eventually, our goal is to have one HMM for each action type.

The process to evolve simple models into complex ones relies on three basic operations. **Data incorporation** integrates a new observation sequence into an existing (possibly empty) model. **State merging** consolidates the resulting model in a way which allows it to generalize to yet unseen trajectories by merging paths corresponding to similar action performances. **Model evaluation** approximates how well a given model fits its constituting dataset.

This scheme allows our models to achieve good generalization performance when faced with new examples while also being capable of one-shot learning after just one seen example. Learning, as generalization over the concrete observed examples, is driven by structure merging in the model in a way that we trade model likelihood against a preference or bias for models of lower complexity. This is well known as the Occam's Razor principle. This principle suggests that among equally well predicting hypothesis one should choose the simplest hypotheses requiring the fewest assumptions.

As graphical models, HMMs are particularly well suited for a model merging approach because data incorporation, state merging and model evaluation are straightforward to apply in this framework as basic graph manipulation operations:

Data incorporation: To integrate a new sequence into a given model we first construct a unique path between the initial and the final state of the model where each symbol in the sequence corresponds to a state in the new path. Each of these states emits its respective symbol in the underlying sequence and simply transitions to the next state, yielding a maximally specific sub path in the model which exactly reproduces the corresponding sequence. After integrating the new path, the probability distribution governing the outgoing transitions from the start state is rebalanced according to the relative frequencies of the pre-existing paths.

State merging: The conversion of the memory based learning scheme with unique maximally specific sub-paths for each sequence in the underlying dataset into a model which is able to generalize to a variety of similar trajectories is achieved by merging states which are similar according to their emission and transition densities. Merging two states q_1 and q_2 means replacing these states with a new state $\hat{q}$ whose transition and emission densities are a weighted mixture of the densities of the two underlying states. Transitions to q_1 and q_2 are redirected to $\hat{q}$ and their probabilities are recalculated according to their empirical estimates in the generating data. As we do not store the underlying samples explicitly, the recalculation of probabilities is tackled by tracking transition and emission statistics corresponding to nodes in the network. Transitions emanating from one of the two old states are simply accumulated and re-routed so that they start from $\hat{q}$. Consolidating the model through state merging abstracts from the concrete examples in the underlying dataset and allows the model to generalize to novel action performances.

Model evaluation: We evaluate the models resulting from the merging process using a score composed of a structural model prior $P(M)$ and the data dependent model likelihood $P(X|M)$:

$$\lambda P(M) + (1-\lambda)P(X|M) \tag{4}$$

The parameter $\lambda \in [0,1]$ mediates between prior and likelihood (see [21] for a detailed analysis). The model prior $P(M)$ acts as a data independent bias. In our system we employ an Occam's Razor like prior favoring simpler models. Giving precedence to simpler models with fewer states makes this measure the primary driving force in the generalization process:

$$P(M) = e^{|M|}. \tag{5}$$

The model size $|M|$ corresponds to the number of states. The complexity of the transition and emission distributions in each state could also be involved in this calculation. However, in this setting we found that the number of states alone produces the best performing models. While the structural prior favors simpler models, its antagonist, the model likelihood, has its maximum at the initial model with the maximum likelihood sub-paths. The exact model likelihood given the dataset X is computed as:

$$P(X|M) = \prod_{x \in X} P(x|M) \tag{6}$$

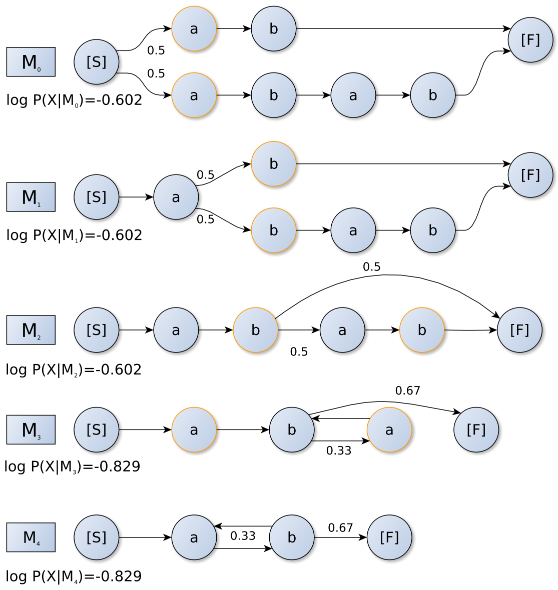


Fig. 5. Sequence of models obtained by merging samples from an exemplary language $(ab)^+$, reproduced from [19]. Transitions without special annotations and all emissions have the probability 1.0.

with

$$P(x|M) = \sum_{q_1 \ldots q_l \in Q^l} p(q_I \to q_1) p(q_1 \uparrow x_1) \ldots\; p(q_l \uparrow x_l) p(q_l \to q_F). \tag{7}$$

where l is the length of the sample and q_I, q_F denote the initial and final states of the model. As we do not want to store the underlying samples explicitly, we use an approximation which considers only the terms with the highest contribution, the Viterbi path:

$$P(X|M) \approx \prod_{q \in Q} \left(\prod_{q\prime \in Q} p(q \to q\prime)^{c(q \to q\prime)} \prod_{\sigma \in \Sigma} p(q \uparrow \sigma)^{c(q \uparrow \sigma)} \right) \tag{8}$$

where $c(q \to q\prime)$ and $c(q \uparrow \sigma)$ are the total counts of transitions and emissions occurring along the Viterbi path associated with the samples in the underlying dataset (see [19] for details). All experiments were conducted using $\lambda = 0.1$, giving the likelihood precedence over the bias for models of lower complexity[1].

The simplest model in our approach is a model which generates a single sequence and assigns the complete probability mass to this unique sequence, equally distributed to each state. We call such models 'maximum likelihood models' because they produce their respective sequences with the highest possible probability. Starting from maximum likelihood models over individual sequences we build more general HMMs by

[1] See [21] for details concerning model parameters and properties in the model for learning action concepts.

merging simpler ones and iteratively joining similar states to intertwine sub-paths constructed from different sequences, allowing them to generalize to different action performances. These initial models can be seen as being obtained by 'unrolling' the paths used in generating the samples in the target model. By iteratively merging states we attempt to undo the unrolling, searching the space of possible models back to the generating model. Merging of two models M_1, M_2 is done by first joining the start states of the models and re-balancing the outgoing transitions afterwards. In the second step, the final states of both models are merged and all transitions to the former final states are consolidated and re-routed to the final state of the joint model. When merging two maximum likelihood models, the resulting model would simply represent the underlying sequences. When generating from such a model, the actual sequence which will be generated is determined early by taking one of the possible paths emanating from the start state. A model which only consists of sub-paths which are themselves maximum likelihood models of an underlying sequence is thus unable to generalize to any other sequence not in its constituting dataset. An example illustrating the mechanism for model merging is shown in Figure 5. The first model (M_0) in this example is constructed over two sequences $\{ab, abab\}$ and thus has two sub-paths originating from the start state, each having a probability of 0.5. After taking the first transition from the start state, the model completely converges to generate either ab or $abab$. Only the transitions from the start state display stochastic behavior, while the individual sub-paths are completely deterministic.

In order to enable the model to also generate and understand yet unknown sequences, we have to intertwine these paths and their underlying characteristics. This is done trough state merging (Alg. 1), where we first build a list of possible merge candidates using a measure of similarity between state emission and transition probability densities. In this approach we use the symmetrized Kullback–Leibler divergence. Then we greedily merge the best pair of states and re-evaluate the model likelihood. We continue the merging process until we reach a point where the likelihood of the resulting model decreases to a level which cannot be compensated by the prior rewarding simpler models.

Example: An example of this merging process can be seen in Figure 5. The merging process starts with a model M_0 constructed from two of the previously mentioned maximum likelihood sequences, which were sampled from the regular language $(ab)^+$. The likelihood of this initial model is calculated according to equation 8 as $log(0.5)+log(0.5) = -0.602$. As most transitions have a probability of 1.0, only the two transitions emanating from the initial state contribute to the result. To start the state merging process we select the first two states to be merged according to their emission and transition similarities. This similarity measure can be seen as an approximation to the expected drop in likelihood of the resulting model. In this example, we selected the highlighted states. After merging, we get M_1 as the resulting model. The overall model likelihood does not change as we have again only two paths contributing to this measure, both having a probability of 0.5. After merging the next two candidates, we yield model M_2 again without a drop in model likelihood. The next merge yields M_3, leading to a first drop in likelihood, but as we have a prior favoring less states, M_3 is still more preferable compared to M_0. This model is now cyclic and able to generate more sequences than the original sequences $ab, abab$ it was created from. In fact, being cyclic enables M_3 now to generalize to the language $(ab)^+$ which was used to generate the constituting samples in the first place. The last merge simplifies the model further to the most compact form to generate the example language.

```
Data: current model $M$
best model $M_{best} = M$
best model score $S_{best} = P(X|M)$
candidates: $C \subset (NxN) \setminus \{(s_1, s_2)|a_1 = s_2\}$
for $c \in C$ do
    $\hat{M} \leftarrow$ new model with $(s_1, s_2)$ merged
    $S_{\hat{M}} = P(X|\hat{M})$
    if $S_{\hat{M}} \leq S_{max}$ then
        $\hat{M} =$ lookahead($\hat{M}$); recurse for one merge
        $S_{\hat{M}} = P(X|\hat{M})$
    end
    if $S_{\hat{M}} > S_{max}$ then
        $M_{best} := \hat{M}$; current best model
        $S_{best} := P(X|\hat{M})$; current best score
    end
end
return $M_{best}$
```

Algorithm 1: State merging algorithm. Initialize the resulting model (M_{best}) with the current model and set the score for M_{best} to the score of the current model. Generate a list of candidate state tuples (s_1, s_2) to merge according to the similarity of their emission and transition densities. Construct a new model $\hat{M}$ from M with states (s_1, s_2) merged and check if the resulting model scores higher than the current best model according to Equation 4. If the model scores higher it is remembered as the currently best model, if not the algorithm tries one more merge as *lookahead*.

D. Grounding syntactic constructions in qualitative action models

This section briefly explains how the component for inducing generalized linguistic constructions is extended to incorporate the action models captured by the HMMs. In essence, the symbolic predicates 'observed' in context are replaced by observations of which object is moving relative to which other object at which time stamp as follows:

(9)

NL sentence	blue_circle jumps over green_rectangle.
Objects	*trajector*: *blue_circle* *landmark: green_rectangle*
Moves/positions	*move(1234,blue_circle,[11:12]);* *move(1277,blue_circle,[12:13]);...*

When observing such an example, both a specific construction 'The blue circle jumps over the green rectangle' and a specific HMM capturing that movement sequence are generated. We assume here that a learning system can already recognize particular objects, e.g. blue_circle as well as green_rectangle as well as the role they play (e.g. trajector or landmark). Note that we do not assume that the system knows already the names for such objects.

The system has thus three tasks:

1) Learn the names for objects (through associational learning)
2) Induce generalized linguistic constructions abstracting over specific slots (through postulation of slots and associational learning)
3) Develop general action concepts abstracting from specific action instances (through model merging)

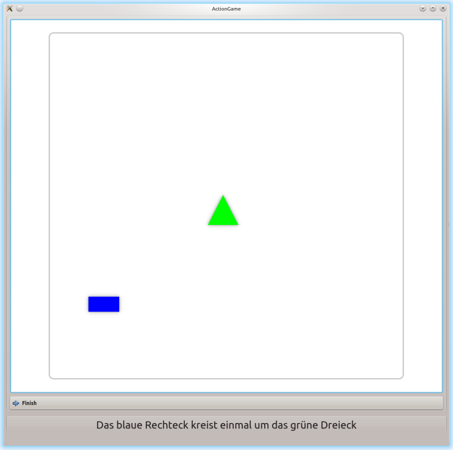
These three tasks are solved by the model inducing generalized syntactic constructions as described in Section II-A together with the representation and induction approaches described in Section II-B.

The two components for inducing generalized syntactic constructions as well as action concept induction interact bidirectionally as follows:

- When the component for inducing generalized constructions encounters two specific constructions that can be merged into a generalized slot-and-frame pattern by abstracting from parts of the sentence by introducing slots, it performs this generalization only if the HMMs associated to the specific constructions are mergeable. Mergeability means here that the similarity $0 \leq sim(M_1, M_2) \leq 1$ between the action models is above a given threshold (0.86) which was optimized independently using randomized grid search. The similarity was derived from a distance metric between HMMs that is similar to the Kullback-Leibler distance between distributions. We let both HMMs generate sequences and for each of these sequences accumulate the difference between the likelihood of that sequence given the generating model and the likelihood given the other model. This procedure is similar to the one proposed by Juang et al. [22]. If the models are mergeable, then both HMMs are merged into a new HMM that represents a more general action concept in the sense of accounting for more variability in action performance.
- In case the component for inducing generalized actions detects that two HMMs associated with different syntactic constructions are extremely similar in the sense that they very likely represent the same action, then it is inferred that both constructions are synonyms of each other.
- In case two sentences are exactly the same, the two HMMs are directly merged.

III. Learning scenario and input data

We consider a learning scenario in which the system learns from written sentences, describing different actions coupled with example 2D trajectories corresponding to these actions. We considered four actions, i.e. *jump onto*, *jump over*, *revolve around (once)*, and *pushes*. These actions were chosen because they can be performed easily in a 2D-scenario regardless of the types of objects involved and because they also provide some challenges regarding discriminability, e.g. instances of *jump onto* and *jump over* may have rather similar trajectories. In previous work [23] we already collected suitable data which is also used for the experiments presented in this paper.



Das blaue Rechteck kreist einmal um das grüne Dreieck

Fig. 6. Simple game with two geometric objects which can be freely moved on the gamefield. In this screen test subjects are asked to revolve the blue rectangle around the green triangle (instruction in the lower part of the screen).

To collect the data, we implemented a simple game in which users could slide geometric objects on a screen (see Fig. 6 for an example screen-shot). Participants were asked to play 100 game rounds, each corresponding to a unique combination of action and objects to perform the action with.

In each trial, a sentence expressing an action, e.g. “the blue rectangle circles once around the green triangle”, was displayed on the screen along with two objects named in the sentence. Subjects were asked to perform the action described by the displayed sentence accordingly by sliding the corresponding object(s). Each displayed sentence described one out of the four different actions. For each action a single syntactic pattern was used to generate sentences describing the action, with different combinations of the objects appearing in the syntactic slots of the pattern. In previous work [23], we used the following four patterns that differ in their verbs or prepositions:

- *trajector* pushes *landmark* from left to right
- *trajector* jumps onto *landmark*
- *trajector* jumps over *landmark*
- *trajector* revolves once around *landmark*

We considered 9 objects for *trajector* and *landmark*, i.e. 3 geometric forms (rectangle, triangle, circle) × 3 colors (red, blue, green), and 25 different sentences (i.e. instantiations of the pattern) were generated for each action, for example “the red circle pushes the green triangle from left to right”. Trajectories were determined by sampling the positions of both objects at a fixed rate. We collected data from 12 subjects (9 male, 3 female, mean age = 29,4 years), yielding 1200 input examples altogether.

Extending our previous experiments, in this article we also attempt to merge HMMs and corresponding sentences/patterns based on similarity between HMMs. Thus, it is also our goal to identify synonyms, i.e. words expressing the same action. In order to measure system performance regarding this issue, we modified the collected data set in that we used two different verbs for each of the four actions. More specifically, for each action performance in an input example we randomly choose one out of two possible sentences as the descriptions; we used the following patterns to generate descriptions:

- *trajector* [pushes|shoves] *landmark* from left to right
- *trajector* [jumps|hops] onto *landmark*
- *trajector* [jumps|hops] over *landmark*
- *trajector* [revolves|circles] once around *landmark*

Notice that the patterns were chosen such that similarity between action models, i.e. HMMs, is indeed an important criterion in order to determine synonyms correctly. For instance, taking solely linguistic variation into account could also yield an incorrect merging of patterns "*trajector* jumps onto *landmark*" and "*trajector* jumps over *landmark*" into "*trajector* jumps [onto|over] *landmark*".

IV. Experimental evaluation

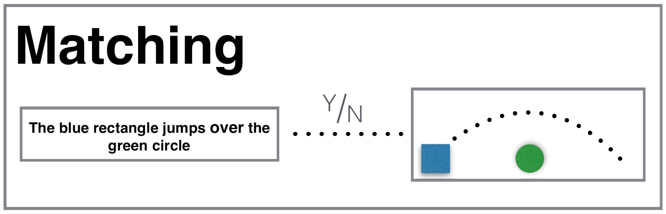
Since we explore grounded language learning, we are interested in the system's generalization abilities both at the linguistic and conceptual level. That is, the main goals of the system are to i) understand and generate novel sentences, and ii) abstract over concrete trajectories of actions, in particular to also recognize actions performed by novel subjects. Thus, we consider two evaluation scenarios:

1) *novel-performer:* 12-fold cross-validation over all subjects, i.e. training on data collected for 11 subjects and testing on the data of the 12-th subject.
2) *novel-sentence:* 25-fold cross-validation in which all sentences observed during testing are novel, i.e. none of them has been observed during training and thus cannot be understood or generated by performing rote-learning. Folds are generated by first collapsing data from all 12 subjects and then partitioning into 25 folds so that in each fold we have the same number of examples for each of the 4 action categories and 4 sentences which are not contained in any other fold.

The developed system is evaluated in two different experimental settings: one concerning the understanding and one concerning the generation abilities. To measure the system's performance we compute precision, recall and f-measure (the harmonic mean of precision and recall). Recall is computed as the percentage of testing examples for which the system generates the correct result and precision as the percentage of correctly generated results of the number of testing examples for which the system actually generates a result (i.e. the system may choose that it cannot determine the result, for instance, because it has not been able to determine a suitable syntactic pattern and/or action model). In order to estimate to what extent the system is able to detect synonyms for actions, we present a (mainly qualitative) analysis of the learned grammars.

In the following, we will first focus on language understanding abilities using a matching and a choosing test, and subsequently explore a language generation experiment. Afterwards, we put our results into context.

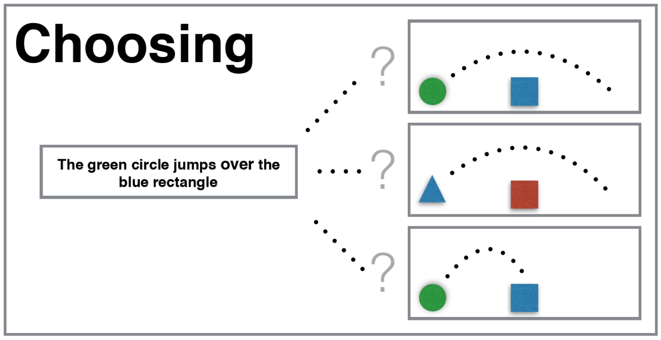
A. Matching test



In the first experiment, we evaluate the system's understanding abilities in a matching task. The test is depicted in the figure above. The system receives as input pairs of sentences and action performances, presented as QTC sequences. The system has then to decide whether the sentence describes the action. These testing data are generated such that the action corresponds to the sentences in about 50% of the examples. More specifically, we keep the appropriate action for half of the testing examples and shuffle the action sequences for the other half such that the action does not correspond to the sentence. Any system has thus a 50% chance level of providing the correct response.

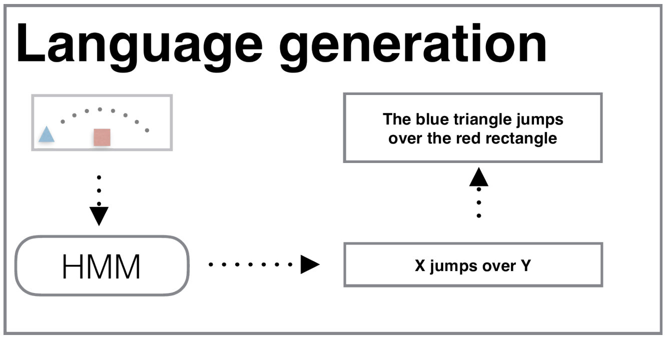
The matching test is implemented using our model as follows: given an input sentence, the systems retrieves a generalized syntactic construction from the construction network that matches the input sentence. It then retrieves the associated HMM. If this HMM is the model that has the highest likelihood of generating the specific QTC sequence, then the system determines that the sentence matches the action.

B. Choosing test



This task is schematically depicted in the figure above. When presented with a sentence paired with three action instances represented as QTC relations, the system has to decide which of the three actions the sentence refers to. Hereby, it is guaranteed that the sentence refers to exactly one action. The other two actions are confounder actions that depict an unrelated action type as well as an action of the same type as described by the sentence but with other objects. Given the sentence ‘The blue circle jumps over the red rectangle’, one example would indeed encode a blue circle jumping over a red rectangle, while the others would encode, e.g., a blue triangle jumping over a red rectangle as well as a blue circle jumping *on* a red rectangle.

C. Language generation test



To evaluate the system’s language generation abilities we first generate sentences for given testing actions. The sentences are generated by extracting all learned knowledge from the construction network – i.e. syntactic patterns, lexical units and groupings of elements – along with their associated HMMs and subsequently reversing the associations. For example, we might extract a pattern “X pushes Y” associated with a meaning comprising an HMM and the information what lexical units can occur in positions X and Y along with their corresponding role in the meaning (such as *trajector*). Given a testing action, the system first determines the HMM that has the highest likelihood of generating the sequence. Based on the grammar, it can then retrieve the corresponding syntactic pattern along with the information about lexical units and their roles. The generated sentence is considered correct only if it is identical with the example’s actual sentence or the alternative sentence describing the same action.

We compare our results against a baseline that was established by choosing a sentence from the training data that has been observed with a similar meaning representation. Similarity is rated on both the involved objects (referents) and the action sequences. For the action sequences we implemented a simple matching score based on the Levenshtein distance between the compressed QTC_c sequences. For all pairs of trajectories t_1, t_2 we calculate a matching score as the Levenshtein distance normalized with respect to its theoretical upper bound $lev(t_1, t_2)/max(|t_1|, |t_2|)$. Because the distances are calculated over the compressed sequences it can also be considered an instance of Dynamic Time Warping. This baseline can however only yield matches in the *novel-performer* condition; in the *novel-sentence* condition none of the testing sentences has been observed during training and thus cannot be found by simply taking a sentence observed with a similar meaning.

D. Results

Results for all three tests along with their corresponding baseline values are presented in Table I. The results reveal that the system achieves a large increase in performance over the random baseline, i.e. performing well above chance level, in both language understanding tests when synonym detection is active.

TABLE I
RESULTS FOR THE MATCHING (LANGUAGE UNDERSTANDING), THE CHOOSING (LANGUAGE UNDERSTANDING) AND THE LANGUAGE GENERATION TEST IN THE NOVEL-PERFORMER AND NOVEL-SENTENCE CONDITIONS WITH AND WITHOUT SYNONYM DETECTION.

Matching test				
Setting	Synonym Detection	F_1	Precision	Recall
Baseline			50% chance	
novel-performer	No	75,42	75,42	75,42
novel-performer	Yes	99,08	99,08	99,08
novel-sentence	No	47,32	90,11	32,08
novel-sentence	Yes	86,50	88,69	84,42
Choosing test				
Setting	Synonym Detection	F_1	Precision	Recall
Baseline			$\sim$33% chance	
novel-performer	No	67,70	100,00	51,17
novel-performer	Yes	99,33	100,00	98,76
novel-sentence	No	55,86	88,00	40,92
novel-sentence	Yes	80,62	92,00	71,75
Language generation test				
Setting	Synonym Detection	F_1	Precision	Recall
Baseline		89% Levenshtein Distance		
novel-performer	No	68,67	68,67	68,67
novel-performer	Yes	74,00	74,00	74,00
novel-sentence	No	54,82	79,49	41,83
novel-sentence	Yes	64,09	67,63	61,08

For the **matching test** in the novel-performer condition, F_1, precision and recall are alike, since the system answers yes if the HMM retrieved for the sentence has the highest likelihood of generating the observed action sequence, otherwise the system answers with ‘no match’, thus yielding an answer for each testing example. Since most sentences were parsed correctly (as indicated by high values for precision and recall), the system appears to have induced suitable grammar and action models in most cases, i.e. for most folds. The learned action models appear to generalize well to a novel performer for most human subjects. For the *novel-sentences* condition values are slightly lower, especially when the synonym detection is not used, which likely results from an insufficient determination of syntactic patterns, i.e. syntactic patterns may not have been learned before testing. When faced with sentences which are instances of unknown syntactic patterns the system would not generate an answer, resulting in the low recall of 32% in the condition without synonym detection. But even in this condition, 90% of the actually generated answers were correct. With synonym detection the system is mostly able to respond even to unknown sentences, resulting in a F_1 score of 84%, which is clearly above the baseline of 50%. Taken together, the results are promising, showing a large increase in performance over the baseline. It is remarkable that in both the novel-performer as well as the novel-sentence case, performance increases substantially when the synonym detection condition is active, by 25% as well as by almost 40% F-Measure. The reason for the impact of the synonym detection is clearly due to the way the dataset has been constructed, replacing 50% of the sentences by synonym sentences. Nevertheless, the results on the novel-sentence condition show that the component for inducing synonyms is indeed working very well.

On the **choosing test**, the system outperforms the baseline condition by large (67,70% and 55,86% vs. 33% on the novel-performer and novel-sentence conditions without synonym detection). Especially in the novel-perfomer condition the precision is 100%, indicating that if the system is given several potential meanings for a sentence and cannot determine the correct match it does not confuse the sentence with distractor meanings, even if these are also somewhat similar to the observed sentence, i.e. corresponding to the same action or involving the same objects. The impact of the synonym detection is also very large with increases in terms of F-Measure by over 30% (from 67,70% to 99,33%) for the novel performer setting and close to 25% (from 55,86% to 80,62%) for the novel-sentence setting. Thus, taking the results for both tests together, the learned models appear to be suitable to yield generalized linguistic constructions and action models that generalize to unseen sequences as well as a reasonable discrimination ability between different actions.

In the **language generation test**, the system performs only slightly below the baseline in the *novel-performer* condition, showing that by merging observed action trajectories for several subjects into generalized action models the discriminative power is mostly retained. However, the learned grammar and models yield the additional benefit that the system is able to also generate sentences not observed during training. In particular, in the *novel-sentence* condition the system is still able to generate several sentences correctly, even though it has never observed them or their corresponding meanings before, which corroborates the generalization abilities of our model.

V. Discussion

We have presented a model that accounts for the co-emerge of linguistic constructions as well as corresponding (grounded) action concepts, mutually influencing each other to fit the observed reality. At both levels, representation learning is performed in a bottom-up incremental fashion, unifying and merging specific instances into generalized representations that capture the essential characteristics of linguistic structures and action concepts, being ‘generative’ in the sense of being able to produce different surface forms. The mutual influence of emerging representations across both modalities is bidirectional in our model. On the one hand, recognition of equivalent or mergeable linguistic structures drives the model towards merging/unifying action representations into equivalence classes, that is, into one generative model. This allows learning of the essence of action concepts denoted by action verbs. On the other hand, similarity in action models leads our system to the inference that two verbal constructions might indeed be synonymous, e.g. as in the case of *‘X jumps over Y’* and *‘X hops over Y’*.Neues In what follows we discuss the implications of our work with respect to work on i) grounded cognition and in particular computational models of cognitive grammar, as well as, ii) linguistic relativity and the role of language in cognitive development, thinking and category/concept formation. We also discuss the relation of our work to current theories and models of language acquisition.

Grounded Cognition and Cognitive Grammar: Our work is related to work that postulates that conceptual knowledge is grounded in modality-specific systems ([24], [25]). As a special case of conceptual knowledge, language is also regarded as being grounded in perception and modality-specific systems [26]. In fact, our generative models are able to generate modality-specific simulations of perception and are thus inherently modal.

As learning proceeds, our system develops a grounded representation of the action denoted by the verb in form of a generative model, a Hidden Markov Model in our case. These Markov Models can be seen as intensional - vs. extensional - meaning representations that are grounded in perception and allow to perceptually ‘simulate’ the action denoted by the verb in question. Our HMMs can to some extent be seen as a specific implementation of the perceptual symbol systems proposed by Barsalou [27].

Our work is also related to attempts to provide a simulation-based and embodied semantics for natural language. Feldman et al. have proposed X-Schemas as a way to capture the (embodied) meaning of a certain linguistic construction. The work by Feldman et al. on Embodied Construction Grammar (ECG) ([28], [29]) is very related to our approach. However, the X-Schemas by which the meaning of linguistic constructions are represented are very symbolic compared to our qualitative models. Our qualitative models are still far away from a full grounding in the sensoric and actuator systems of an embodied system, but clearly go one step further than the X-Schemas used in Embodied Construction Grammar (ECG). The closest related work is the one of van Trijp et al. [30], who have developed approaches by which robots can learn linguistic knowledge in the framework of Fluid Construction Grammar (FCG). However, as far as we know, they have not developed any approach that can actually induce these X-Schemas from observation. Further, the work of Feldman and colleagues has neither considered how synonym relations could be inferred by a system on the basis of detecting similarity between X-Schemas. This would indeed presuppose a notion of similarity between X-Schemas. Such a notion of similarity is inherent in our model, operationalized as ‘unifiability’ of two models. In general, there are to our knowledge no models that make detailed predictions about how synonyms or near-synonyms are acquired.

Our work is related both to approaches to grounded acquisition of language in robots and cognitive systems, but also to approaches to the representation and acquisition of actions. With respect to approaches to grounded acquisition of language, there has been a lot of work on developing models which can acquire single words and their meanings (e.g. [31], [5]). In some approaches, this meaning is grounded in perception, but is typically limited to objects ([10], [32]).

Other approaches (e.g. [33], [34]) deal with the acquisition of syntactic constructions as we do, but typically do not ground these constructions in qualitative action models. With respect to the representation and acquisition of actions, different approaches based on prototypes [35], markov models with referential representations [36] or representations based on QTC [37] and neural networks [38] have been proposed.

Many authors have emphasized the cognitive interaction between action and language ([39], [40], [41], [42]). We have attempted to provide a detailed model that explains how action and language structures emerge in interaction, influencing each other.

Linguistic Relativity and Language as Enhancer of Cognitive Abilities: Our proposal is further in line with the paradigm of linguistic relativity, corresponding to the claim that the language one speaks or hears influences one's own conceptualization and the categories one forms. While the strong claim that language determines thought has been largely abandoned [43], there is increasing empirical evidence showing that language influences thought. Wolff and Holmes [43] have described four ways in which language could influence thought, or more accurately, an emerging conceptualization, distinguishing the functions of i) Language as Meddler, ii) Language as Augmenter, iii) Language as Spotlight, and iv) Language as Inducer.

The understanding of language as a tool that enhances the computational and cognitive abilities of humans has lucidly been spelled out by Clark [44]. In our account, language acts both as a spotlight as well as as an inducer of a schematic representation of experience. In fact, as suggested by Waxman and Markow [45], language might serve as an invitation to form a new category. In our case, observing a sentence for the first time leads to induce a category specific for that sentence. Generalization of several sentences leads to recognizing a pattern and to induce a cognitive schema that represents the essence of the action category denoted by the generalized verbal construction. Language is thus playing the role of triggering the search for a schematic category, in our case an HMM, that supports conceptualizing experience. There is indeed a lot of empirical evidence showing that language can facilitate category formation in the above sense. Xu [46] found that the presence of distinct labels facilitated object individuation. Xu concludes that language may play an important role in the acquisition of sortal/object kind concepts in infancy and that words may play as *'essence placeholders'*. This is exactly what is happening in our model. On encountering a sentence the first time, our system creates a placeholder for the essence of this sentence. This early 'essence' is very specific for the given situation in which the sentence was heard and lacks any generalization. Later, when observing similar sentences, the corresponding essences are generalized by merging them into more general essences. Other researchers have shown that language can help to acquire the distinction between approachable and non-approachable creatures [47]. Gentner and Boroditsky have suggested two processes that are active in learning the concepts that are denoted by words. They refer to *cognitive dominance* when concepts emerge from cognitive-perceptual processes and later the name for these categories is acquired. They refer to *linguistic dominance* when *'the world presents perceptual bits whose clumping is not predefined and language has a say how the bits get conflated into concepts'*. Clearly, as argued by Gentner and Boroditsky [48], this is not a dichotomy, but a continuum that spans a space in which the acquisition of a certain concept for a name can be located. According to Gentner and Borodtiksy, the acquisition of categories denoted by proper names or concrete nouns rather lies on the cognitive dominance side of the continuum. While kinship terms and verbs lie somewhere in the middle of the continuum, prepositions, conjunctions and determiners are rather positioned at the right side of the continuum. Our model explains the acquisition of categories towards the right end of continuum. The formation of such categories is triggered by the fact that learners are confronted with a new construction or name. It is certainly an open question where on the dominance continuum our specific actions such as pushing, jumping on, circling around are positioned at. This could be certainly determined experimentally. For the sake of providing a proof-of-concept for our model, we have assumed that the categories are not available previous to encountering the corresponding verbal constructions. While this is a mere assumption, our model is certainly not depending on that.

Emergence of meaning as a mapping: Some researchers have criticized the 'mapping metaphor', that is the idea that language needs to be mapped to some priorly existing 'concept'. As mentioned in the introduction, Lila Gleitmann has emphasized that the assumption that language acquisition consists in the acquisition of (new) names for existing concepts is clearly an over-simplification. Overcoming this simplifying assumption has been one of the goals of our approach. Tomasello [3] has criticized the 'mapping metaphor' on the grounds that it neglects that learning the meaning of words is actually a process of contextual inference in which the intentional structure of an action is considered to infer what the speaker is actually referring to. Rohlfing and colleagues [4] have recently criticized the mapping metaphor on other grounds arguing that children would *not necessarily remember the connection between the word and the referent unless it is framed pragmatically*, that is, it is introduced in the context of a recurring interactional pattern with the purpose of achieving a joint goal between tutor and learner. While presented as an alternative to the mapping paradigm, they rather hypothesize that a communicative pragmatic frame facilitates to learn which concept a certain word evokes. The work of Rohlfing et al. and the proposal of pragmatic frames can be regarded as an elaboration of the general theory of Tomasello claiming that recognition of intention, shared attention and goals as well as the ability to simulate others as intentional agents are crucial ingredients by which children infer the meaning of a certain word or expression in context.

Part of the above mentioned criticisms on the mapping approach stem from the fact that the term 'mapping' is not clearly defined. As lucidly highlighted by McMurray et al. [6], there are two notions of meaning: the referential meaning of a sentence or expression in a given situation and the intensional meaning of an expression. The referential meaning is inferred in a particular situation on the basis of an understanding of the situation. The intensional meaning corresponds to the situation-independent meaning of a linguistic expression, that is to its 'essence'.

Neither the theory of Tomasello nor the work of Rohlfing et al. make any predictions about how the intensional aspects of the meaning are learned over time and across situations as a byproduct of experiencing the word in different contexts. In a standard formal semantics paradigm, the intensional or situation-independent meaning aspects of a sentence can be captured using a truth-conditional approach, capturing the logical constraints that need to be fulfilled for the sentence to be true in a given world or situation. In this sense, our model tries to capture the essential meaning of a linguistic expression as it appears in our data, albeit not using a standard truth-conditional semantics approach. In contrast, in our model intensional meaning is captured via probabilistic models that, instead of modelling logical conditions on the worlds in which the sentence is true, models the distribution or likelihood of observing a certain action sequence in a world or situation described by a verbal construction that is associated with the given HMM.

Most works dealing with the question how systems can learn the intensional meaning of a word have investigated how systems distill the meaning of a word by cross-situational learning, that is contrasting the different situations in which a word has been heard (see [7], [6]). First, as argued above, referential and intensional meaning is often confounded. Second, according to our understanding of the notion of mapping, it does not imply that the corresponding concept pre-exists independently. In particular, our understanding of the term "mapping" does not necessarily imply that the category to which a word is mapped to exists already. In our approach, upon first encounter of a construction, it is 'mapped' to a maximally specific novel category that is created in the very moment in which the sentence and action are observed.

Nevertheless, the term mapping is not only problematic for the above reasons, but for the fact that it suggests that the meaning of words can be 'mapped' to a static symbol or handle. In contrast, we prefer to talk about *'evocation'* of a situation-independent meaning following the theory of Frame Semantics [49] that postulates that words evoke more than just a static referent or handle, but more complex semantic frames, cognitive schemas or, in our case, grounded representations of the essence of the word's meaning that can be used to simulate. This is compliant with the view of Taylor and Zwaan [50], who argue that: *'when a person hears or reads text involving action, there is activation of the motor systems in his or her brain', which corresponds to the referential semantic content of the description'*.

Regarding how a learner infers the meaning of a novel verb in context, Gleitman (see above) has proposed the syntactic bootstrapping theory according to which a syntactic construction (e.g. a transitive construction) can give cues about the potential meaning of a transitively used verb in context. However, for bootstrapping mechanisms to work, a learner needs to have learned a generic transitive construction and the general concept of *an agent that does something to another agent*. In earlier work we have provided an account for how such abstract constructions and categories might emerge [51]. They presuppose that thematic roles such as *agent*, *patient*, etc. are already acquired. In previous work, we have also proposed a learning architecture that combines top-down and bottom-up processing to learn constructions starting from sequences of phonemes [52].

Surely, syntactic knowledge as well as inference about the intentions of others plays an important role in inferring the meaning of unknown words in context. The recognition of intentions and of the teleology of actions emerges very early in childhood [53]. So far, our model is limited in that it neither models how pre-existing linguistic knowledge nor reasoning about the intentions of others or of the purposes of actions are factored in into the task of figuring out the meaning of a new verbal constructions. In fact, our model so far only considers the spatio-temporal structure of action concepts. The incorporation of these factors into our model is an obvious avenue for future work, albeit a very challenging one.

Analogy making: Hofstädter and Sanders [54] have recently argued that analogy making is a fundamental process in cognitive processing. According to their theory, category formation and language learning is driven by analogy-making, that is by fitting previously acquired categories to a given observation and then extending the category to the new observation, leading to generalization, which can be possibly perceived as an over-generalization for mature systems the language of which is heavily influenced by conventions. Take the example of the sentence *'I undressed the banana'*, which for mature speakers is an overt over-generalization, while from a cognitive perspective it makes a lot of sense. In some sense, this is what our model is making: it is constantly making analogies. It induces very specific action categories for a very specific sentence and then, when observing a similar sentence, it attempts to extend both the linguistic construction and the category to cover the new observation as well. In doing this, it is guided by the desire to minimize the complexity of the model, that is the number of bits needed to store the action model, while not loosing predictive power, that is not over-generalizing. This mechanism could be regarded as a direct implementation of the theory of Hofstädter and Sanders. We have provided the proof-of-concept in simulated experiments that this principle works for the case of simple verbal constructions denoting actions in which some object (an agent) moves with respect to some reference object. Whether our principles could be extended to other constructions and more complex cognitive models and simulations is an open question. However, there seems to be no principled reason why our approach would not be extensible to other categories. Arguably, HMMs would not be able to model all kind of representations. But we stress that our proposal is not specific to HMMs. Our proposal requires that there is some (probabilistic) generative model that represents some (induced) category. For dynamic categories such as (action) verbs that involve a spatio-temporal signature (what linguistically is called often 'path') surely a sequential model will be needed, while for other static categories (e.g. objects denoted by nouns) a static prototype as simulation might be sufficient. Most likely, however, even object representations have a non-static meaning component representing the potential to act on the object, corresponding to so called 'affordances' [55]. In terms of the theory of Hofstädter and Sanders, our generative models capture the *'essences'* of concepts of which we see a specific realization or *'surface'* in the real world.

VI. Conclusion

We have presented a computational model of language acquisition that models the joint emergence of linguistic constructions and the conceptual categories they evoke. In line with 'slow mapping' approaches that claim that the acquisition of the meaning of a word is a long-term process that starts with the first encounter ([56], [6]), we have provided an account of language learning in which, upon the first encounter of a linguistic construction or expression, a learner induces an ad-hoc category that is specific for the given sentence and context. From many encounters of similar sentences, learners distill the essence of the category evoked by the word through incremental generalization, guided by the desire to produce compact models while not reducing the accuracy of the predictions made by the model on sequences observed so far. Concepts in our approach are 'essences' evoked by the linguistic constructions and are generative models that can generate various 'surfaces'. The goal of a learner is to distill these essences from the many encounters with a given word or linguistic construction. In doing this, they incrementally stretch and extend the meaning of early categories to subsume other early categories by a process that could be understood as one of constant analogy making and extension of categories. Our proposal is thus very much in line with the understanding of cognition as analogy making.

We have looked in particular at the case of verbal constructions and corresponding actions in which a trajector moves with respect to some reference object along a characteristic spatio-temporal path. As actions are temporal sequences, the models evoked by verbs denoting actions have to be, at the very least, sequential models capturing the likelihood of sequences generated by the model. We have decided to capture the embodied meaning of verbs via Hidden Markov Models that can be used also to anticipate the completion of actions (see Panzner et al. [21]). Capturing concepts of both modalities in a model that is graphical in principle allows not only to interrelate concepts across modalities but also to unify both models in a single joint representation in future work. The choice for graphical models, in contrast to e.g. neural models, is also substantiated by the hypothesis that cognitive processes could be partially explained and conceptualized by a cognitive architecture that is based on graphical models (Danks 2014 [57]).

Our approach models the first encounter of a word as the start of the acquisition of a category and thus is in line with the theory of linguistic relativism in the sense that language triggers the induction of a category. It follows from our approach that if cultures use different constructions, their speakers will as a corollary induce different categories. Future work should look at the predictions of our model and design experiments that can test whether categories belong to the linguistic or more to the cognitive domain.

Our model so far has concentrated only on modelling the spatio-temporal essence of actions in which some trajector is moved relatively to some reference object. This is a very restricted subset of actions. To account for other verbs, the teleological structure of actions would need to be modeled. A richer modelling of actions will also require representing thematic roles such as patient and agent in a developmentally appropriate and grounded fashion. Our model has assumed that a system can recognize such relations from the input and model them symbolically. In terms of the abilities of a developing language learner, our system assumes that the following abilities are already acquired or inborn: the ability to segment and track objects, the ability to segment and identify words, the ability to recognize and conceptualize basic scenes as well as the ability to jointly attend to a tutor that demonstrates actions. All these are in itself complex abilities the emergence of which needs to be explained. However, this is out of the scope of the current paper. Finally, while our proposed computational theory for modelling the co-emergence of linguistic and action representations has been shown to be effective in our experiments, it is still a very 'mechanistic' theory, assuming that a system can represent construction networks and HMMs explicitly, perform statistical inference and merging operations on these representations, etc. It is indeed an open question if learning systems can encode and manipulate such representations on a neural substrate. In future work, we will seek to find architectures that show a similar behaviour but do not postulate the existence of data structures such as proposed in this paper but instead rely on deep learning and representational learning to explain the co-emergence of representations across modalities.

Our work also paves the path for developing the foundations for a physically grounded simulation-based semantics of natural language.

VII. Acknowledgements

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

[1] M. Bowerman and S. C. Levinson, Eds., *Language Acquisition and Conceptual Development*. Cambridge University Press, 2001.

[2] P. Wolff and K. J. Holmes, "Linguistic relativity," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 2, no. 3, pp. 253–265, 2011.

[3] M. Tomasello, "Could we please lose the mapping metaphor, please?" *Behavioral and Brain Sciences*, vol. 24, no. 6, pp. 1119–1120, 2001.

[4] K. J. Rohlfing, B. Wrede, A.-L. Vollmer, and P.-Y. Oudeyer, "An alternative to mapping a word onto a concept in language acquisition: Pragmatic frames," *Frontiers in Psychology*, vol. 7, no. 470, 2016.

[5] J. M. Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings," *Cognition*, vol. 61, no. 1-2, pp. 39–91, 1996.

[6] B. McMurray, L. K. Samuelson, and J. S. Horst, "Word learning emerges from the interaction of online referent selection and slow associative learning," *Psychological Review*, vol. 119, no. 4, pp. 831–877, 2012.

[7] L. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," *Cognition*, vol. 106, no. 3, pp. 1558–1568, 2008.

[8] L. B. Smith, "How to learn words: An associative crane," *Breaking the word learning barrier*, pp. 51–80, 2000.

[9] L. Gleitman, "The structural sources of verb meanings," *Language Acquisition*, vol. 1, no. 1, pp. 3–55, 1990.

[10] D. Roy and A. Pentland, "Learning words from sights and sounds: a computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.

[11] H. Behrens, "Usage-based and emergentist approaches to language acquisition," *Linguistics*, vol. 47, no. 2, pp. 383–411, 2017.

[12] M. Tomasello, "The item-based nature of children's early syntactic development," *Trends in Cognitive Sciences*, vol. 4, no. 4, 2000.

[13] J. M. Pine and E. V. Lieven, "Slot and frame patterns and the development of the determiner category," *Applied psycholinguistics*, vol. 18, no. 2, pp. 123–138, 1997.

[14] J. Gaspers and P. Cimiano, "A computational model for the item-based induction of construction networks," *Cognitive Science*, vol. 38, no. 3, pp. 439–488, 2014.

[15] M. Pickering and S. Garrod, "An integrated theory of language production and comprehension," *Behavioral and Brain Sciences*, vol. 36, no. 4, pp. 329–247, 2013.

[16] N. Weghe, B. Kuijpers, P. Bogaert, and P. Maeyer, "A Qualitative Trajectory Calculus and the Composition of Its Relations," *GeoSpatial Semantics SE - 5*, vol. 3799, no. Dc, pp. 60–76, 2005.

[17] R. Rojas, *Theorie der neuronalen Netze*. Springer-Verlag, 1993.

[18] M. Panzner and P. Cimiano, "Comparing hidden markov models and long short term memory neural networks for learning action representations," in *Machine Learning, Optimization, and Big Data*. Springer International Publishing, 2016, pp. 94–105.

[19] S. Omohundro, "Best-first model merging for dynamic learning and recognition," in *Advances in Neural Information Processing Systems 4*. Morgan Kaufmann, 1992, pp. 958–965.

[20] R. N. Shepard, "Toward a universal law of generalization for psychological science," *Science*, vol. 237, no. 4820, pp. 1317–1323, 1987.

[21] M. Panzner and P. Cimiano, "Incremental learning of action models as HMMs over qualitative trajectory representations." Workshop on New Challenges in Neural Computation (NC2), 2015.

[22] B.-H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden markov models," *AT&T technical journal*, vol. 64, no. 2, pp. 391–408, 1985.

[23] M. Panzner, J. Gaspers, and P. Cimiano, "Learning linguistic constructions grounded in qualitative action models," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2015.

[24] L. W. Barsalou, W. K. Simmons, A. K. Barbey, and C. D. Wilson, "Grounding conceptual knowledge in modality-specific systems," *Trends in Cognitive Sciences*, vol. 7, no. 2, pp. 84–91, 2003.

[25] L. W. Barsalou, "Grounded cognition," *Annu. Rev. Psychol.*, vol. 59, pp. 617–645, 2008.

[26] F. Pulvermüller, "How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics," *Trends in cognitive sciences*, vol. 17, no. 9, pp. 458–470, 2013.

[27] L. W. Barsalou, "Abstraction in perceptual symbol systems," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 358, no. 1435, pp. 1177–1187, 2003.

[28] J. A. Feldman, *From Molecule to Metaphor: A Neural Theory of Language*. MIT Press, 2006.

[29] N. Chang, J. Feldman, and S. Narayanan, "Structured connectionist models of language, cognition and action," in *Ninth Neural Computation and Psychology Workshop*, 2004.

[30] R. Van Trijp, L. Steels, K. Beuls, and P. Wellens, "Fluid construction grammar: The new kid on the block," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 63–68.

[31] A. Fazly, A. Alishahi, and S. Stevenson, "A probabilistic computational model of cross-situational word learning," *Cognitive Science*, vol. 34, no. 6, pp. 1017–1063, 2010.

[32] K. Hsiao, S. Tellex, S. Vosoughi, R. Kubat, and D. Roy, "Object schemas for grounding language in a responsive robot," *Connect. Sci.*, vol. 20, no. 4, pp. 253–276, 2008.

[33] X. Hinaut, M. Petit, G. Pointeau, and P. F. Dominey, "Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks," *Frontiers in Neurorobotics*, vol. 8, no. 16, pp. 1–17, 2014.

[34] P. F. Dominey and J.-D. Boucher, "Learning to talk about events from narrated video in a construction grammar framework," *Artificial Intelligence*, vol. 167, no. 1-2, pp. 31–61, 2005.

[35] E. Kruse, R. Gutsche, and F. M. Wahl, "Acquisition of statistical motion patterns in dynamic environments and their application to mobile robot motion planning," in *IROS'97.*, vol. 2. IEEE, 1997, pp. 712–717.

[36] K. Sugiura, N. Iwahashi, H. Kashioka, and S. Nakamura, "Learning, generation and recognition of motions by reference-point-dependent probabilistic models." *Advanced Robotics*, vol. 25, no. 6-7, pp. 825–848, 2011.

[37] M. Hanheide, A. Peters, and N. Bellotto, "Analysis of human-robot spatial behaviour applying a qualitative trajectory calculus," in *RO-MAN, 2012 IEEE*. IEEE, 2012, pp. 689–694.

[38] A. Droniou, S. Ivaldi, and O. Sigaud, "Learning a repertoire of actions with deep neural networks," in *Joint International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob)*, 2014, pp. 6–p.

[39] A. C. Anna M. Borghi, "Action and language integration: From humans to cognitive robots," *Trends in Cognitive Science*, vol. 6, pp. 344–358, 2014.

[40] A. Cangelosi, V. Tikhanojff, J. F. Fontanari, and E. Hourdakis, "Integrating language and cognition: A cognitive robotics approach," *IEEE Computational Intelligence Magazine*, vol. 2, no. 3, pp. 65–70, 2007.

[41] R. M. Willems and P. Hagoort, "Neural evidence for the interplay between language, gesture and action: A review," *Brain and Language*, vol. 101, p. 278–289, 2007.

[42] A. M. Glenberg and M. Kaschak, "Grounding language in action," *Psychonomic Bulletin and Review*, vol. 9, pp. 558–565, 2007.

[43] P. Wolff and K. J. Holmes, "Linguistic relativity," *WIREs Cogn. Sci.*, 2010.

[44] A. Clark, "Magic words: how language augments human computation," in *Language and thought: Interdisciplinary themes*, P. Carruthers and J. Boucher, Eds. Cambridge University Press, 2001.

[45] S. Waxman and D. Markow, "Words as invitations to form categories: evidence from 12-to 13-month-old infants," *Cognitive Psychology*, vol. 7, no. 470, 1995.

[46] F. Xu, "The role of language in acquiring object kind concepts in infancy," *Cognition*, vol. 85, pp. 223–250, 2002.

[47] G. Lupyan, D. H. Rakison, and J. L. McClelland, "Language is not just for talking: Redundant labels facilitate learning of novel categories," *Psychological science*, vol. 18, no. 12, pp. 1077–1083, 2007.

[48] D. Gentner and L. Boroditsky, "Individuation, relativity, and early word learning," in *Language Acquisition and Conceptual Development*, M. Bowerman and S. C. Levinson, Eds. New York: ambridge University Press, 2001.

[49] C. J. Fillmore, "Frame semantics and the nature of language," *Annals of the New York Academy of Sciences*, vol. 280, no. 1, pp. 20–32, 1976.

[50] L. Taylor and R. Zwaan, "Action in cognition: the case of language," *Language and Cognition*, vol. 1, no. 1, pp. 45–58, 2009.

[51] J. Gaspers, A. Foltz, and P. Cimiano, "Towards the emergence of verb-general constructions and early representations for verb entries: Insights from a computational model," in *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2014.

[52] J. Gaspers, P. Cimiano, K. J. Rohlfing, and B. Wrede, "Constructing a language from scratch: Combining bottom-up and top-down learning processes in a computational model of language acquisition," *IEEE Trans. Cognitive and Developmental Systems*, vol. 9, no. 2, pp. 183–196, 2017.

[53] G. Gergely and G. Csibra, "Teleological reasoning in infancy: the naive theory of rational action," *Trends in Cognitive Sciences*, vol. 7, no. 7, pp. 287–292, 2003.

[54] D. Hofstadter and E. Sander, *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic Books, 2013.

[55] J. J. Gibbson, "The theory of affordances," in *Perceiving, Acting, and Knowing*, R. Shaw and J. Bransford, Eds., 1977.

[56] S. Carey, "The child as word learner," in *Linguistic Theory and Psychological Reality*, M. Halle, J. Bresnan, and G. Miller, Eds., 1978.

[57] D. Danks, *Unifying the mind: Cognitive representations as graphical models*. Mit Press, 2014.