Introduction

Humans are the only species with a communication system which is fundamentally variable at all levels (Levinson and Evans 2010). Decades of work by typologists have shown that syntax and semantics of language are largely (if not completely) culture specific. This is even true for perceptually close language domains such as the semantics and syntax of spatial language (Brown 2008; Levinson 1996). If this hypothesis is true, then learners of a language face a tremendous problem. There are roughly 6000 languages in the world. All normally developed children are capable of learning any one of these languages at rapid pace. How is that possible?

The answer to this question must be that children have very general learning mechanisms that can account for the acquisition of any language. The most promising approach to understanding language learning comes from proponents of usage-based language learning and Construction Grammar (Tomasello 2003; Lieven et al. 2003). This approach stresses that (1) the basis of child language acquisition is the utterance as an expression of communicative intentions; (2) children’s earliest utterances are concrete in the sense that they are instantiations of item-based schemas or constructions; (3) abstractions result from children generalizing across variation they observe at particular “slots” in otherwise same utterances; and (4) children create novel utterances for themselves based on observed examples and similarity to such examples.

The usage-based approach to language learning has given rise to a number of computational and robotic models that examine specific leaning of aspects of language such as ditransitive construction (Chang and Maia 2001), Aktionsarten (Gerasymova and Spranger 2010; 2012) and event structure constructions (Steels et al. 2012). All these models are interesting in that they show that some constrained set of constructions can be fruitfully learnt based on prerequisites such as a known lexicon.

In this paper we discuss a recently developed model of construction learning (Spranger 2015; Spranger and Steels 2015) that does not require any prior knowledge of the lexicon or other constructions but learns from scratch embodied constructions for referring to objects, colors, spatial relations and co-reference. The model is validated in grounded interactions between a tutor and a learner robot that interact in an office environment. The model integrates category learning for the acquisition of concepts (Spranger 2013), acquisition of compositional semantics (Spranger, Pauw, and Loetzsch 2010) and learning of constructions for mapping semantics to strings and back.

In this paper we focus on the construction learning part of the model. The paper starts by explaining the interaction patterns we have designed to study language learning. We then dive into the function of syntax for expressing semantics and how we represent semantics. This is followed by a description of the model. The paper finishes with a description of the experimental setup, results and discussion.

Robot Interaction Setup

We study the acquisition of language using robots that interact in an office space. There are two robots. One is the tutor. The other robot is the learner. Both robots try to draw each others attention to objects in the environment using natural language utterances and various extra-linguistic signals (pointing, head shaking and head nodding).

The set-up is shown in Figure 1. The environment consists of a number of objects (represented in the graphical representation of the situation model as circles), boxes (rectangles) and interlocutors (arrows). The vision system of each robot tracks objects in the vicinity and establishes a model of the environment with real-valued distances and orientations of objects with respect to the body of the robot. The environment is open-ended. Objects and boxes are added or removed and their spatial configuration changed. Moreover, robots are free to move around. For the purpose of evaluation we recorded more than 1000 spatial scenes with different numbers of objects (up to 15 objects) and different configurations of objects.

Tutor and learner engage in scripted interactions. The flow of an interaction is sketched in the following bullet points.

1. Each agent perceives the scene. This basically means that one scene from the 1000 available scenes is chosen and one view is loaded by the learner and the other by the tutor.
2. Each robot is assigned a role: either speaker or hearer. If the tutor is the speaker, then the learner is the hearer and vice-versa. The agents then engage in the following interaction.

3. The speaker selects an object (further called the topic \( T \)) from the situation model. The speaker tries to find a meaning (this can involve spatial relations, landmarks and perspective) which discriminates \( T \) from other objects in the scene. Subsequently, the speaker produces an utterance for this meaning.

4. The listener parses the utterance and tries to interpret the (potentially partial) meaning of the observed utterance to identify possible topics in the scene. The listener then points to the topic he thinks is the most likely interpretation of the phrase.

5. The speaker checks whether the listener points to \( T \). If the listener pointed correctly, the game is a success and the speaker signals this outcome to the listener.

6. If the game is a failure, then, depending on the tutoring strategy, additional things may happen. If the tutor is the speaker then he can point to the object he had in mind. Otherwise, if the tutor is the listener, then he points to the object he understood to be the topic. If the learner is the speaker and the tutor the listener, the tutor might nevertheless say how he would refer to the topic. For the experiments described here, the tutor only points when he is the speaker and the game is a failure.

Figure 2 details the systems involved in these interactions. In particular there are two systems that are important for this paper. Incremental Recruitment Language (IRL) and Fluid Construction Grammar (FCG). IRL is a framework for representing and learning concepts, categories and semantic and conceptual structure. It bridges between real-valued, continuous sensorimotor representations and symbolic meaning. FCG is a construction grammar formalism allowing to express mappings between symbolic meanings and Natural language strings and back.

**Computational Cognitive Semantics with IRL**

The Incremental Recruitment Language - or short IRL - (Spranger et al. 2012) enables the robots to process the symbolic meaning of Natural language phrases. IRL is different from logic-based approaches because it emphasizes the procedural (Haddock 1989) aspect of the meaning of utterances. Rather than describing truth-value, IRL represents the meaning of utterances as general programs consisting of different representations and algorithms.

Figure 3 shows the IRL-program (meaning) of the phrase “left of the block”. The structure contains pointers to categories and spatial relations in the form of bind statements (in bold), as well as a number of cognitive operations. For example, construct-region-lateral constructs a region representation. Cognitive operations and bind statements are linked using variables (which are symbols starting with ?). For instance, the variable ?landmark links a subpart of the IRL-program identifying “the block” to the landmark input slot of the operation construct-region-lateral thereby capturing the fact that “the block” should act as the landmark to the region.

For the experiments here are four types of categories and rela-
tions are important: selectors, concepts, color categories and spatial relations.

Object categories Each objects in the world can be broadly categorized into various concepts: robots, boxes, blocks and balls. The tutor knows these categories and they allow him to identify objects. Object categories are basically expressed as nouns in noun phrases.

Color Categories Objects are colored and robots can choose to express color categories. We use here 4 color categories: “red”, “blue”, “yellow” and “green”. We represent color categories using a similarity function based on a prototypical color (Bleys et al. 2009).

\[
sim_c(o, c) := e^{-\frac{1}{2\sigma^2} |o_c - c_c|}
\]

where \(o\) is some object, \(c\) the category, \(o_c\) the color of a particular object \(o\) and \(c_c\) the prototypical color of category \(c\).

Spatial relations English locative spatial relations can be broadly categorized into three different classes. Proximal categories such as “near” and “far” rely on proximity to some particular landmark object. Projective categories are categories such as “front”, “back”, “left” and “right”. These categories are primarily angular categories signifying a direction. The direction can come from the reference object itself (intrinsic) or can be induced by the observer or some perspective (relative frame of reference (Retz-Schmidt 1988)). Absolute categories such as “north”, “south”, “east” and “west” which rely on a compass direction, with the pivot direction to the magnetic north pole. Other absolute systems rely on features of the environment to determine the layout of the angles (Brown 2008). In the experiments discussed here, the wall marker is used as a global direction on the scene.

We represent spatial categories using a similarity function (Herkovits 1986) based either on a prototypical angle (for absolute, projective) or distance (for proximal) enveloped by an exponential decay:

\[
sim_a(o, c) := e^{-\frac{1}{2\sigma^2} |a_o - a_c|}
\]

where \(o\) is some object, \(c\) the category, \(a_o\) the angle to a particular object \(o\) and \(a_c\) is prototypical angle of category \(c\). Importantly, angular and proximal distances are always defined relative to a coordinate system origin. By default this is the robot observing the world.

Selectors The determiners “the” and “all” are modeled as selectors that select objects from sets of objects. Given a particular input set these selectors choose objects from the set. For instance, “the” requires that the object is unique in the set with respect to similarity. In practice that means that there needs to be an object with a much higher similarity than other objects. “All” just chooses all objects from a set.

Agents can use spatial relations (and other concepts) in IRL-programs combined with different cognitive operations.

Set operations such as picking the highest scored member of a set etc., which are important for dealing with determiners such as “the”.

Categorization operations take a set as input and score objects according to some similarity functions defined by categories. Examples are apply-class and apply-category.

Mental rotations are implemented as linear algebra operations that transform a feature space such as angle and direction to another point of origin, e.g. geometric-transform. These operations also handle different frames of reference (intrinsic, relative and absolute).

An important insight from cognitive linguistics is that there is a deep connection between semantics and syntax. The following gives two examples from English to highlight this.

1. The left block.
2. left of the block.

Both phrases consist of almost the same lexical material (the, block and left) but their grammatical structure and their meaning structure is quite different. In Example 1 (Figure 4) the spatial relation is used as modifier on the set of objects denoted by the noun, whereas in Example 2 (Figure 3) the spatial category is applied to a landmark denoted by the determined noun phrase. In Example 1 (Figure 4) the spatial relation refers to a group-based reference system (Tenbrink and Moratz 2003). Importantly, these differences are signaled by word classes. When “left” is used as adjective (Example 1, Figure 4) then the group-based reference operation is needed. When “left” is used prepositionally - as in “left of” (Example 2, Figure 3) - then the meaning is a region construction operation.

Expressing Cognitive Semantics using Construction Grammar

In order to compute utterances for meaning (production) and meaning of utterances (parsing), we use a recent variant of a computational construction grammar system called Fluid Construction Grammar or FCG (Steels 2011). FCG allows to specify bidirectional mappings between meanings and utterances in the form of a single grammar. The tutor robot operate a spatial grammar comprised of roughly 46 constructions (bidirectional rules) - primarily lexical constructions, as well as a number of functional and phrasal constructions.

The important constructions are lexical, functional and phrasal constructions.

Lexical constructions are bidirectional mappings between semantic entities (concepts, relations) and words. For instance, there is a lexical construction for “the” that maps (bind selector ?the unique) to the stem “the”.

Functional constructions map smaller semantic context to word classes. One difference between semantics Examples 1 and 2 is precisely in which cognitive operation the same category is used.
learning. The learner starts out with no syntactic knowledge (no words or any knowledge about phrase structure). This is obviously a strong assumption that is not necessary for the complete model (Spranger 2015) but allows us to focus on construction learning. The learner starts out with no syntactic knowledge (no words or any knowledge about phrase structure).

There are various learning mechanisms that together enable the learner to master the language: adoption, generalization and consolidation. Adoption happens when there is unknown syntactic material: words, relations between words etc. Generalization is a mechanism that given enough accumulated constructions starts to extract larger, more abstract patterns. Generalization leads to additional constructions that cover the same utterances and meanings as more specific constructions. Consolidation removes this unused or unnecessary material over time from the memory of the learner.

Adoption Suppose the learner encounters unknown phrase \( s \). So the listener does not know the phrase or some part of the phrase (step 4 in the interaction script fails). In that case the listener points (if he can still interpret the phrase given the context) or signals failure and the speaker points to the topic \( t \). In any case, the listener will have constructed an IRL-program for the topic object based on the pointing received from the speaker. The listener invents a mapping from the complete IRL-program to the complete phrase (or parts thereof).

Generalization Once the learner has acquired enough exemplars, he tries to extract more abstract constructions. Suppose the learner first hears the determined noun phrase “the box”. Initially this will allow him to successfully produce and interpret that exact phrase. Upon hearing another example of a determiner and a noun “the ball”, the learner can now deduce that likely he can build phrases of the form “the \( X \)” where \( X \) is something else namely a particular concept. The learner then invents an item-based construction and other constructions by breaking up the holophrases (see Figure 6 for a graphical explanation).

The splitting of concrete constructions into more abstract constructions adds slots for e.g. lexical items. These items can be constrained in multiple ways. In the experiments here we incorporate the hierarchy of semantic entities (see Figure 7 for a part of that taxonomy). In particular, we use the least abstract common class of two items that can fill the slot. For instance, knowing that a slot can be filled by “front” and “back” allows frontal categories to participate in the slot. Knowing that a slot can be filled with “front” and “near” enables spatial relations for the slot.

Figure 5: Schematic of an item-based construction (and two more lexical constructions) invented by the learner through reasoning over holophrases that have been heard before. The input constructions differ in semantics and in the form in a single item, which allows the learner to make a structural inference and split up the existing constructions.

Figure 6: Schematic of an abstract construction invented by the learner through reasoning over item-based constructions that have been learned before. The input constructions differ in semantics and in the form in a single item, which allows the learner to make a structural inference and split up the existing constructions.
Consolidation Once an item-based construction and its associated more lexical constructions have emerged, they are in competition with the holophrase constructions since they cover the same communicative situations (same meanings and phrases). The learner in production and interpretation knows this and so a competition takes place in the learner between the new constructions and the old holophrases. Setting up the right alignment dynamics for the learner can eliminate the holophrase constructions. Initially the learner will choose the new constructions over the older ones. However, decreasing the score of competing constructions after every interaction leads to a forgetting of the holophrase constructions over time.

These simple learning operators have a number of nice consequences for the model. For example, the same generalization and consolidation operators lead from specific phrases to holophrases and then later to abstract constructions (see Figure 6). Once multiple item-based constructions are learned, they can be further broken up using the exact same learning operator. More and more abstract constructions will emerge with more possible arguments until aspects of English phrase structure constructions emerge.

Experimental Setup and Results
We can test the learning by running multiple tutor-learner simulations. The tutor can express a number of meanings and phrases. In total we see that 546 phrases are used by the tutor (on this particular data set) including determined noun phrases, e.g. “the block” or “the box”, more complex adjective noun phrases, e.g. “the left block”, “the red block”, and very complex phrases such as “the block left of the box from your perspective”.

Since we are only interested in construction learning for this paper, we focus on syntax-semantics mappings and we measure only communicative success and the number of constructions. Communicative success measures whether the learner is able to produce phrases given a particular situation and meaning. If that is the case this is measured with 1. If not the interaction counts as 0. It also measure whether the tutor is able to correctly parse a sentence. If he can do that in an interaction, the interaction counts as 1, otherwise 0.

Figure 8 shows the overall dynamics of learning grammar (for 100 experiments). In all 100 experiments, the learner learns to be successful in communication after roughly 350 interactions (average 80% success). At this point the learner has been exposed to an average of 120 utterances (for about 550 total possible utterances used by the tutor), which is before the learner has seen all possible utterances, but enough to be successful for that stage. Initially the tutor chooses to expose the learner to simple scenes. This makes the learner immediately successful. Over time the tutor increases the complexity of the environment and the language needed to cope with the environment. That’s why the learner does not reach 100% success but keeps learning.

At the same time, we can see an overshoot of the number of constructions in the inventory of the learner. This is due to an initial rise of holophrases, which later slowly die out, being overtaken by more item-based constructions and the emergence of lexical constructions. The lexical constructions give already a lot of information and often allow the learner to successfully interpret the phrase given simple environments. Later functional constructions (handling a single cognitive operation or a set of them) and more phrase structure like constructions (only mapping variable linkings to word order) emerge. Slowly memory is consolidated. Typically, the learner does not end up with 46 constructions (which is what the tutor is using). A few more survive because the learner has not seen some enough examples of complex constructions. Importantly, the learner starts talking using phrases he observed more or less from the start. When there are initial holophrases and lexical constructions, the learner can also parse and make utterances he has never seen before.

One final point: when we inspect the learning trajectories of agents, we can see various pathways are possible through construction learning in different experimental runs depending on the order the learner hears phrases. So for instance, in some experimental runs, color adjectives are picked up faster than spatial adjectives.

Discussion
The experiments done here rely on a number of representational choices that make learning easier but have unintended consequences. For instance, we do not split plural markers from words but leave them as is. Therefore, the learner ends up with 2 lexical constructions one for “blocks” and one for “block” with the same meaning. In comprehension that is fine because both constructions give the same result. However, when the learner produces this can lead to phrases such as “the blocks” and “all block”. These in turn
are no problem for the tutor but they are wrong. The only problem where this currently happens is with singular/plural distinctions. To solve this problem requires slightly advanced methods not reported here.

The present model relies on simple learning operators. It is actually quite surprising that this works. Related work has shown that simple operators can work in conjunction with the right choice of the tutor strategy (Spranger 2015). The tutor reduces the search space of generalizable constructions which in principle is as large as the power set of all possible sentences. He does so by presenting similar yet slightly different examples and gradually increasing the complexity of the phrases used.

The model sketched here is not the first to attempt to model construction learning. However, this model is unique in its attempt to learn a construction grammar of English end-to-end, without pre-given lexicon or other linguistic knowledge. Currently ongoing work is validating this end-to-end learning by testing the model on other linguistic domains such as event structure and other languages such as German. Lastly, we want to integrate the model sketched here with models of the emergence of gesture.

References


