OPEN-ENDED SEMANTICS CO-EVOLVING WITH SPATIAL LANGUAGE

MICHAEL SPRANGER AND SIMON PAUW
Sony CSL Paris, 6 rue Amyot, 75005 Paris, France
spranger@csl.sony.fr

MARTIN LOETZSCH
VUB AI Lab, Vrije Universiteit Brussels, Pleinlaan 2, 1050 Brussels, Belgium

How can we explain the enormous amount of creativity and flexibility in spatial language use? In this paper we detail computational experiments that try to capture the essence of this puzzle. We hypothesize that flexible semantics which allow agents to conceptualize reality in many different ways are key to this issue. We will introduce our particular semantic modeling approach as well as the coupling of conceptual structures to the language system. We will justify the approach and show how these systems play together in the evolution of spatial language using humanoid robots.

1. Introduction

Linguists have long studied spatial language as an important foundation of communication in many languages of the world. Spatial utterances such as "the building left of the train station" are typically used to discriminate an object (sometimes referred to as *figure*) in a specific spatial scene. Such utterances can contain one or more *spatial terms* (e.g. *left*) (?) as well as a marker or lexical item for a *reference object*, also called *trajector*, *landmark* or *ground* (in this case "train station")[a]. There are also other types of spatial utterances not involving an explicit reference object, for example "This ball here!". However, as already demonstrated by ? (?) who made one of the first attempts on modeling spatial language with robots, talking about objects using spatial language inevitably involves some notion of perspective, hence the choice of a particular point of reference. As ? (?) suggests, the study of spatial language is best framed using three distinctions: 1) *absolute* frames of reference that relate to fixed features of the environment, for instance, seaside/mountainside or gravity, 2) *intrinsic* frames of reference that are based on objects that have an inherent orientation, e.g. humans have an inherent front, as

[a] There is quite some debate in linguistics as to how to define spatial language (?). Here we only consider utterances that involve spatial relational terms.

in "the glass in front of you", and finally 3) *relative* frames of reference that allow to express spatial relations with respect to some object from the viewpoint of the observer, as in "left of the tree"[b]. Languages differ in which frames of reference they instantiate, e.g. the Mayan language Tzeltal features only an absolute frame of reference.

How language users choose a particular reference system to distinguish some object in the environment has been studied extensively in experimental psychology. It seems now clear that this choice depends not only on the saliency of the landmark (?), but also on the prototypicality of the spatial position of the object, with respect to the category system and the particular frames of reference available to the speaker. Moreover, speakers align their choice of reference points during interactions (?).

The phenomena encountered in natural language make it clear that in order to study the evolution of spatial language, in particular the alignment of conceptualization and language strategies, we first need to answer the question how agents flexibly compose complex semantic structure, encode the semantic structure into syntactical structure and back, thus making themselves understood, and second we need learning operators that shape the particular way of how agents choose to express themselves in a given environment. Consequently, this paper first deals with our approach to representing semantic programs, called *Incremental Recruitment Language* (IRL), followed by an introduction to the system for producing and parsing utterances, called *Fluid Construction Grammar* (FCG). We then move on to present an experimental setup that tests the expressive power of the mechanisms for studying the evolution of spatial language. Lastly, we add learning operators to the system and explore how ambiguity arising in conceptualization can be resolved collaboratively by a community of agents, thereby highlighting how and why spatial language might have evolved.

This paper builds on extensive previous work. IRL has first been described in ? (?) and extended in ? (?, ?). FCG also has a long tradition of development in our lab (?). Moreover this is not the first attempt to model spatial language. ? (?) explained alignment of choice of perspective and ? (?) highlighted exaptive mechanisms behind spatial language stemming from bodily meaning. However, the approach in this paper is unprecedented in terms of complexity of semantics, as well as integration of FCG, IRL and embodiment.

2. Flexible Representation of Conceptual Structure with IRL

What is the meaning of the utterance "the block to my left"? Presumably this utterance is an attempt to draw the attention of the hearer to some object in the environment. In procedural semantics terms: the speaker wants the hearer to ex-

[b] Trees do not feature an inherent orientation, i.e. have no left side. It is thus the position of the observer in relation to the tree, that determines what is left of the tree.

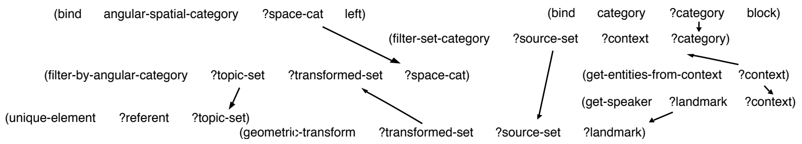
(bind angular-spatial-category ?space-cat left)
(filter-set-category ?source-set ?context ?category)
(bind category ?category block)
(filter-by-angular-category ?topic-set ?transformed-set ?space-cat)
(get-entities-from-context ?context)
(get-speaker ?landmark ?context)
(unique-element ?referent ?topic-set)
(geometric-transform ?transformed-set ?source-set ?landmark)

Figure 1. IRL network representing the meaning of the utterance "the block to my left".

ecute a program – a set of instructions that will lead him to identify the object in the environment that is the referent of the phrase. The steps that the hearer should follow are 1) search for objects in the environment that are blocks, i.e. filter the context of the interaction for things that belong to the class of blocks and than 2) filter for the blocks that are left of the speaker. This last operation can be split up even further into first put yourself in the position of the speaker and than filter for objects that are at the left side. In summary, the meaning of the utterance is in the particular combination of *cognitive operations* that are connected in a *network*.

Figure 1 shows a possible semantic structure for our example phrase. The network consists of nodes that represent cognitive operations, which are instantiations of cognitive mechanisms such as categorization, discrimination and so forth. This particular graph contains 1) the operation `filter-set-category` that filters a set using the category `block`, 2) the mechanism `geometric-transform` for changing perspective to the perspective of the `speaker` 3) the operation `filter-by-angular-category` that filters objects using the projective, angular category `left` and 4) `unique-element` as a check for uniqueness.

Each cognitive operation is characterized by a call pattern. For instance, `filter-set-category` has three slots `(filter-set-category ?target-set ?source-set ?category)` (marked by variables starting with `?`) whic means that it can take three input/output arguments. Given a source set and a category, a target set will be computed, that contains items pertaining to the particular target category. Moreover, when a target set and a source set is passed, then `filter-set-category` tries to compute the category that can filter the source set and lead to the target set. Other combinations are also possible, if we only pass a source set, then all imaginable categories together with the target sets they produce from the source set are computed. How many such combinations are possible depends on the number of categories known to the agent. The categories, but also prototypes, as well as other semantic material that can be used in cognitive operations are called *semantic entities*.

We have already hinted at the linking of cognitive operations via variables. That is, the value computed by one cognitive operation can be input of another. Moreover, semantic entities can be explicitly represented in a network with *bind statements* that assign a value of a particular type to a variable. For example, `(bind angular-spatial-category ?space-cat left)` introduces the category left, so that it can be used by the `filter-by-angular-category`-operation. When variables are explicitly bound, control flow in the network for that particular variable is clear. But, it is extremely important to understand that in all other cases it is not, mirroring the fact that natural language utterances in many cases provide semantic material but leave the actual control flow unspecified or at least under-specified. Whether the utterance “the block to my left’ means 1) filter by blocks, than 2) take the perspective of the speaker, and 3) search for left items, or whether one should first search for items left of the speaker and for items in that set, which are blocks, is ambiguous or at best irrelevant.

Next to the unspecified control flow, the system offers another dimension of flexibility. Cognitive operations can be creatively combined into networks. When speakers try to conceptualize a specific object in a specific scene they are trying to construct networks that best discriminate the target object. Just as human speakers try to conceptualize the world for spatial utterances by identifying good combinations of reference points, frames of reference and spatial categories, agents can freely compose cognitive operations and search the space of possible networks for good solutions, encodable by the particular conventionalized grammar. Trying to interpret an utterance, agents are facing a search problem as well. An utterance might only partially encode a network, which turns the process of understanding an utterance into a restoration of possibly intended meaning. For instance, suppose you hear the phrase “left of the train station”. In English this phrase is ambiguous because it under-specifies how to precisely conceptualize the landmark “train station”, both intrinsic and relative frame of reference are possible. Which interpretation of the two makes sense might be inferable from the context, nevertheless whichever choice one makes, one has filled in an operation that was not explicitly coded in the utterance.

3. Encoding Conceptual Structure into Syntactic Structure using FCG

We now turn to FCG, the computational engine for verbalizing IRL networks. FCG is specifically designed to support language evolution studies in a construction grammar approach. At the heart of the formalism are constructions, which are bi-directional form-meaning mappings. Most importantly, FCG features mechanisms for tracking the usage patterns and the success of particular rules in the inventory of agents. Constructions, as well as all other items known to agents, be it its semantic entities, cognitive operations, or even IRL networks, are scored. The score can be updated according to the success of particular items in communication and reflects how certain an agent is that the items lead to success. Moreover, similarly to conceptualization, we understand the process of verbalizing conceptual structure as a a search process. Trying to produce an utterance for an IRL network comprises searching for conventional lexical and grammatical constructions that maximize communicative success.

FCG can produce phrases starting from IRL programs by adding words and

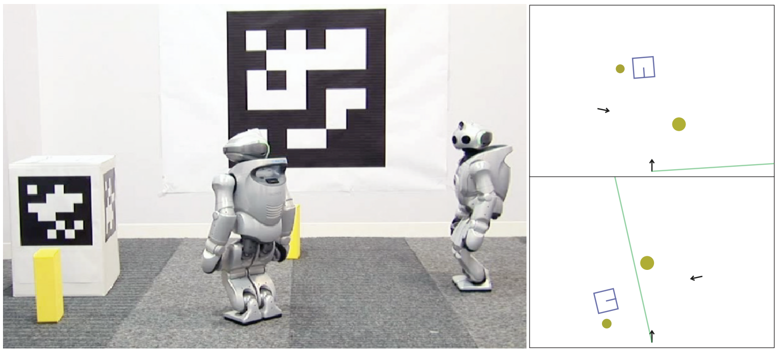

Figure 2. Left: example scene consisting of yellow colored blocks, the target objects, as well as four possible frames of reference (two robots, one box and a marker on the wall). The box to the left is an allocentric reference point and can be used in relative and intrinsic frames of reference. The big barcode attached to the wall signifies a geocentric, absolute direction, similar to north on a compass. Right: world models extracted by each robot for this particular spatial setup (arrows – robots, blue rectangle – box, yellow circles – yellow objects, green line – major direction geocentric frame of reference).

syntactical structure on the form side. When producing, first lexical constructions are applied, mapping some of the conceptual substructure of an IRL network to words. Second, particular cognitive operators and variable links might be conveyed using grammatical markers, endings or word order. Successively syntactical structure is build, that partially encodes the structure of the IRL network. For instance, for the network in Figure 1, first lexical constructions map the categories `block` and `left` to words, followed by mapping the `get-speaker` operation onto the pronoun “me”, followed by adding the marker indicating the role of “me” as landmark and so forth. In parsing the process is reversed. By successively applying constructions, conceptual structure is inferred from syntactical.

4. Experimental Setup

For our setup we equipped robots with mechanisms for composing semantic structure and for verbalizing, parsing and reconstructing such structure and released them into an office environment in which they roam around freely (see Figure 2). When they encounter each other, they play a *language game* (?) in which the task of one agent is to draw the attention of the interlocutor to an object in the environment. In order to understand the influence of the world onto the particular language system developed by agents in the course of consecutive interactions, we setup the world such that the scenes contain multiple target objects, which are indistinguishable in size, color and form. Consequently, the only difference between target objects is their spatial position. To understand speakers’ choices for particular reference systems we spread wooden card boxes that are augmented

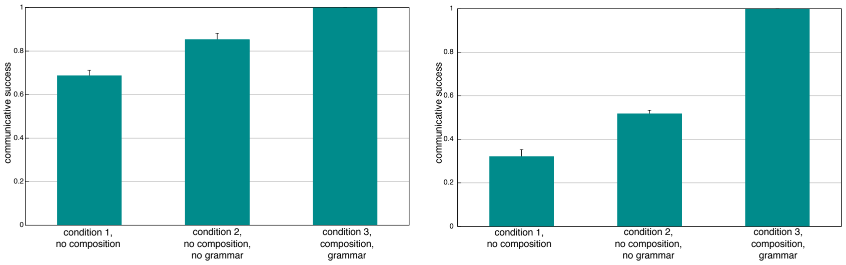

Figure 3. Experimental results I, three experimental conditions, two different sets of scenes. Left: scenes where robots are standing next to each other in all scenes, giving them a similar scene perception. Right: Difficult spatial setups with robots facing each other, looking at the scene from very different angles. Bars show the average communicative success over 20000 games. The benefit of allowing agents to compose complex meanings is evident. However, allowing for different frames of reference also leads to ambiguity, which can be resolved by introducing grammar.

by easily recognizable two dimensional barcodes in the environment. These objects function as reference objects besides the present robots. To run repeatable experiments, we recorded almost 400 scenes, differing in spatial layout. Some feature global reference systems, others only allocentric landmarks. Furthermore, the concrete spatial position of interlocutors is manipulated, such that in some scene agents face each other, whereas in others they have a very similar view on the scene.

5. Experimental Results I

We first investigated the general power of freely composing conceptual structure utilizing the reference systems, points and spatial categories. However, as in natural language, allowing for different ways of conceptualizing without clearly marking them leads to ambiguity. We thus also tested how language and, in particular, grammar can help disambiguate between the different ways of conceptualization by introducing a hand-crafted grammar, reminiscent of English distinctions between absolute, relative and intrinsic frames of reference. There are three experimental conditions: 1) agents are only given spatial categories like `left` together with cognitive operations to categorize the environment and ways to verbalize and parse categories; 2) agents are additionally given cognitive operations to conceptualize the context. Following the findings in human spatial language they are given different conceptualization strategies, i.e. absolute, intrinsic and relative frames of reference. Additionally agents are equipped with lexical items to denote the particular point of reference, but not the specific construal operation used. 3) agents are given a hand-crafted grammar, that marks complete conceptual semantic networks. Results (see Figure 3) show that indeed agents are better of using reference points, but, even more so, should mark the particular conceptual operation used.

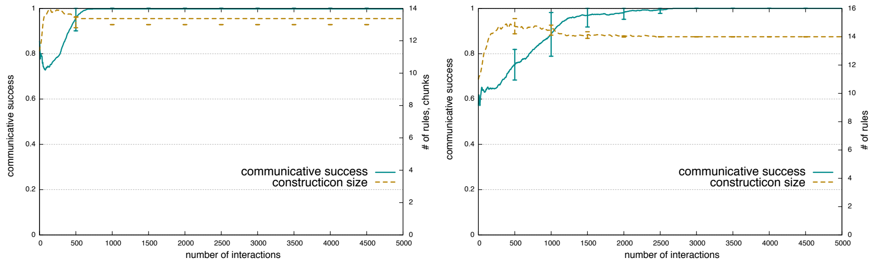

Figure 4. Experimental results II. These graphs show learning for population of 10 agents (10 runs avg) equipped with marker invention and alignment strategies. Left: simple spatial configurations. Right: difficult spatial setup (same as in Figure 3). Agents reach 100% communicative success in both cases (going from the middle bar in Figure 3 to the right bar). For simple spatial setups (left) convergence is much quicker, and the population invents two markers, since there are no global landmarks (one marker each for relative and intrinsic frames of reference). For the difficult spatial setups with global reference frames (right) convergence takes longer, mostly due to the interfering relative frames of reference and homonymy that is created when hearers reconstruct wrong networks.

6. Experimental Results II

Given the promising results of scenario I, which suggest clear communicative advantages for grammar developments in spatial settings, we went on to explore how such a grammar can evolve. In this scenario, we equipped agents with grammar invention and adoption mechanisms, as well as alignment strategies. Moreover, agents were given lexical items to denote categories and reference points. However, in contrast to the third experimental condition of scenario I, they were not given a target grammar but only diagnostic and repair strategies, as well as an alignment mechanism called *lateral inhibition* (?).

Agents that are producing, diagnose ambiguity in their utterance when re-entering (?) – parsing utterance themselves before passing it to the hearer – that do not clearly distinguish between objects in the environment. In such cases agents can repair the shortfall by introducing marking constructions, which symbolize particular semantic structure, i.e. the cognitive conceptualization operation used (intrinsic, relative, absolute). In the process new markers are introduced in the population. Whenever an agent perceives a marker that he does not know, a diagnostic detecting missing items kicks in and triggers an adoption mechanism, that associates the marker with the meaning the hearer inferred. This can of course go wrong. As both agents at the end of the interaction have established the referent of the utterance the speaker produced, but they do not necessarily share the conceptual structure that discriminates it. The same problem occurs in human communication where the referent of the phrase “to the left of the train station” might accidentally coincide with conceptualizing the train station as a relative or intrinsic frame of reference. So agents face the problem of homonymy, which is resolved by laterally inhibiting competitors – punishing constructions that either express the same conceptual structure or use the same marker. However, agents align in spite of this problem (Figure 4), and incorrect mappings die out.

7. Conclusion

We have demonstrated how phenomena in human spatial language can be modeled using an open-ended, rich, conceptual system, that allows agents to flexibly combine cognitive operations into large conceptual structures and how such structures can be verbalized. Furthermore, we have presented a first attempt at in silico analysis of the subtle interplay of frames of reference choice and categorization strategy in spatial language. We, moreover, hypothesized learning mechanisms that can try to resolve naturally arising ambiguities in spatial language showing how and why agents evolve spatial language.