Towards the Emergence of Non-trivial Compositionality

Shane Steinert-Threlkeld*

Department of Linguistics, University of Washington

shanest@uw.edu

Forthcoming in *Philosophy of Science*

Abstract

All natural languages exhibit a distinction between content words (nouns, verbs, etc.) and function words (determiners, auxiliaries, tenses, etc.). Yet surprisingly little has been said about the emergence of this universal architectural feature of human language. This paper argues that the existence of this distinction requires the presence of *non-trivial compositionality* and identifies assumptions that have previously been made in the literature that provably guarantee only trivial composition. It then presents a signaling game with variable contexts and shows how the distinction can emerge via reinforcement learning.

*Thanks to Jeff Barrett, Emmanuel Chemla, Meica Magnani, Iris van de Pol, and Jakub Szymanik as well as the audience at the Symposium on Evolutionary Models of Compositional Communication at PSA2018 and an anonymous referee for this journal for helpful comments and discussion. The author has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. STG 716230 CoSaQ.

Twas brillig, *and the* slithy toves
Did gyre *and* gimble *in the* wabe;
All mimsy *were the* borogoves,
And the mome raths outgrabe.

Excerpt from 'Jabberwocky' in Carroll (1871).

The poem excerpted in the epigraph has often been called a 'nonsense poem'. But it is not entirely so. While the *content* words (unemphasized: nouns, verbs, adjectives) are nonsense, the *function* words (emphasized: determiners, tense, auxiliaries, conjunctions, etc.) are not. The structure that they provide greatly aids our interpretation.

The distinction between these two types of expression occupies a central place in modern linguistics (Carnie, 2006; Muysken, 2008; Rizzi and Cinque, 2016). Rightfully so: every natural language exhibits a distinction between content and function words. The former provide the content of sentences and fall into what are called 'open classes' (it is easy to introduce a new noun, for instance) while the latter provide the 'grammatical glue' of complex expressions and fall into 'closed classes' (it is difficult to introduce a new determiner, for instance).

Yet surprisingly little has been said about the emergence of this universal architectural feature of natural languages. Why have human languages evolved to exhibit this division of labor between content and function words? How could such a distinction have emerged in the first place?

This paper takes preliminary steps towards answering these questions by showing how a communication system with such a distinction can emerge via a process of reinforcement learning. A crucial innovation consists in having the learning agents play a signaling game

across variable contexts which contain multiple objects that possess multiple perceptually salient gradable properties. In the next section, I define a principle of non-trivial compositionality and argue that it is essential for the emergence of function words, while also diagnosing why existing approaches have failed to account for its emergence. After that, I will introduce the new signaling game. Section 3 presents experimental results. I conclude with future directions in Section 4.

1 Non-Trivial Composition

In this section, I build on the foregoing remarks in order to argue for the following claim: for a communication system to have function words, there must exist *non-trivial composition* (in a sense to be made precise) of complex signals. I will then analyze three case studies from the literature on the evolution of compositionality which exhibit only trivial composition. The reasons for this are then made mathematically precise: given the assumptions about optimal communication often made, the resulting systems must be trivially compositional.

The principle of compositionality says that the meaning of a complex expression is determined by the meanings of the parts and how they are put together (Frege, 1923; Janssen, 1997; Pagin and Westerståhl, 2010a,b). Natural languages are compositional: whence the ability of competent speakers to produce and comprehend a potentially infinite set of novel expressions. A language can, however, be compositional without exhibiting the rich flexibility that human languages do. We will use the following definition:[1]

(1) A communication system is *trivially compositional* just in case complex expressions

[1]For this usage, see Schlenker et al. (2016); Zuberbühler (2018).

are always interpreted by intersection (generalized conjunction) of the meanings of the parts of the expression.

The force of this definition can be brought out by an example: Titi monkey calls (Cäsar et al., 2013; Schlenker et al., 2016). In a series of predator-model experiments, it was found that raptors in the canopy elicit sequences of A calls, cats on the ground elicit sequences of B calls, cats in the canopy elicit one A followed by a sequence of Bs, and raptors on the ground elicit a sequence of As followed by a sequence of Bs. While the full details do not concern us,[2] Schlenker et al. (2016) argue that the best analysis of this call system involves the following semantics, interacting with some plausible pragmatic principles:

(2) Compositional semantics of Titi alarm calls: where t is a time,

a. $[\![B]\!]^t = 1$ iff there is a noteworthy event at t

b. $[\![A]\!]^t = 1$ iff there is a serious non-ground alert at t

c. $[\![wS]\!]^t = 1$ iff $[\![w]\!]^t = 1$ and $[\![S]\!]^{t+1} = 1$

[where w is a call and S a sequence of calls]

The crucial feature of this semantics concerns the rule (2c) for interpreting complex expressions (sequences of calls). It says that a sequence of calls is interpreted by first evaluating the beginning of the sequence at time t, then evaluating the rest of the sequence at time $t + 1$, and conjoining the results. This clause results in the following: each call in the sequence contributes to the meaning of the whole *independently* of the other calls, with the complete meaning resulting from conjunction. It thus constitutes a paradigm of the definition of trivial compositionality in (1).[3]

[2]See Steinert-Threlkeld (2016b) for some reservations about the full analysis.

[3]Berthet et al. (2018) argue that the proper semantics for Titi calls is not in fact trivially

In other words, non-trivial compositionality involves non-conjunctive modification of one linguistic item by another. Examples of such systems can also be found in communication systems much simpler than human language. In particular, Campbell's monkeys have been argued to exhibit it (Ouattara et al., 2009; Schlenker et al., 2014). They have two basic alarm calls: an eagle call *hok* and a general alert *krak*.[4] Moreover, both calls combine with what appears to be a suffix *-oo*, which has the effect of weakening the severity of the calls. Schlenker et al. (2016) propose the following semantics:

(3) $[\![R\text{-}oo]\!]^t = 1$ iff at t the sender is alert to a disturbance that licenses R but that is not strong among such disturbances.

This is non-trivial: *-oo* does not contribute independent meaning that is then conjoined with the contribution of *hok* or *krak*. Rather, it combines with one of the latter calls to modify the normal meaning of that call.

Here is the simple argument for the claim that non-trivial composition is necessary for the existence of function words. Recall the characterization thereof as 'grammatical glue': such words do not contribute independent content to a sentence, but structure that provided by the content words. In a trivially compositional communication system, each expression contributes independent meaning to the complex expressions containing it. Therefore, none of the expressions therein are function words.

Note that the presence of non-trivial composition does not suffice for the presence of

compositional. Nevertheless, the presentation just given illustrates what such a system would look like.

[4]The possibly different meaning of *krak* in different habitats of Campbell's monkeys is the subject of the aforementioned papers. I follow Schlenker et al. (2016) in giving it a general meaning.

function words. To see this, consider subsective adjectives (Partee, 1995). These are adjectives like 'skillful', which have the property that for every noun, a 'skillful N' is an N, but is not 'skillful' in any sense independent from the noun. For example:

(4) a. Jakub is a skillful rock climber.

b. Jakub is a cook.

c. Therefore, Jakub is a skillful cook.

The inference pattern in (4) is not valid: Jakub can be skillful at one thing but not at another. If 'skillful' contributed its meaning independently of the noun it combines with, the inference would be valid: Jakub would be a climber, a cook, and skillful; therefore, a skillful cook. But 'skillful' is still a content word. One could imagine a very simple language whose only complex expressions were of the form 'Adj N', but which had subsective adjectives. This language would be non-trivially compositional but would have no function words.

Now, I will present three case studies of prominent models purporting to explain aspects of the evolution of compositional communication. Each of them, however, will turn out to exhibit only trivial composition. After presenting the case studies, I identify common underlying assumptions and then prove a mathematical fact demonstrating that under those assumptions, the resulting communication systems must be trivially compositional. In light of the foregoing, none of these extant approaches can explain the emergence of the distinction between function and content words.

1.1 Three Études

Nowak and Krakauer (1999) apply mathematical models of natural selection to the

evolution of language, providing conditions under which a 'grammatical' language will evolve from a non-compositional one. In their model, states are object-action pairs, loosely modeling events. They compare two types of languages: one in which each object-action pair has an independent label, and another in which each object has a corresponding expression, each action has a corresponding expression, and the agents communicate by sending the corresponding pair of expressions to communicate about an object-action pair. While the results they obtain are indeed interesting, it should be clear from this brief exposition that the type of language that they consider exhibits only trivial composition: each component of a complex expression contributes its bit of meaning (either an object or an action) independently of the other.

Barrett (2007, 2009) studies a generalization of signaling games with multiple senders (Lewis, 1969; Skyrms, 2010). In the simplest case, there are four states of nature and two senders, each of whom can send one of two signals to one receiver. The senders, but not the receiver, know which state obtains. Simulations show that a simple form of reinforcement learning leads these agents to a situation of perfect communication. Given the nature of the setup, the resulting systems look as follows. One sender partitions the four states into two sets of two, one for each signal. The other sender sends its two signals in an *orthogonal* partition (Lewis, 1988). One can imagine the states as a two-by-two square, with one sender indicating the row and the other the column of the true state. While these agents have certainly learned to solve a non-trivial coordination problem, their communication system again exhibits only trivial composition, since the meaning of each sender's signal is independent of the other's and the receiver interprets the sequence by intersecting the two.

Finally, Mordatch and Abbeel (2018) study the emergence of communication in a

multi-agent setting where each agent has a private goal that it wants to achieve.[5] The agents—which are in this case recurrent neural networks—communicate about a world with various colored landmarks in it. Each agent additionally has a color and its own perspective from its position (i.e. no agents share a frame of reference). The goals consist of getting an agent to perform an action (going to or looking at) at one of the landmarks. With appropriate costs for maintaining large lexicons, the agents learn to send sequences of signals with separate signals for which agent, which action, and which landmark. These three types of signals have independent meanings, which are combined by conjunction.

1.2 A Limitative Result

There is in fact an underlying reason that these systems exhibit only trivial composition. Although the three cases just illustrated come from different theoretical frameworks, they all share the same following assumptions:

(A1) Agents communicate about a fixed set of states. (Object/action pairs, separate points of a state space, and agent/landmark/action tuples, respectively.)

(A2) Optimal communication consists in correctly identifying the true member of the state space.

(A3) Messages are fixed-length sequences of signals from fixed sets.

Under these assumptions, optimal communication will be trivially compositional:

[5]The set of goals is assumed to be consistent, i.e. all of the goals are simultaneously realizable.

(5) Let X and $\{M_i\}_{i\in I}$ be any sets, and f, g two functions of the following type:

$$X \xrightarrow{f} \prod_i M_i \xrightarrow{g} X$$

Define $f_i^{-1}(\overrightarrow{m}) := \{x \in X : f(x)_i = \overrightarrow{m}_i\}$. Then the following holds.

$$\text{If } g \circ f = \mathrm{id}_X, \text{ then for all } \overrightarrow{m} \in f[X], \{g(\overrightarrow{m})\} = \bigcap_i f_i^{-1}(\overrightarrow{m})\ ^6$$

Here, X represents the fixed set of states about which the agents communicate. Note that the structure of this set does not matter. $\prod_i M_i$ is the set of possible sequences of signals, with each M_i being the signals available to be sent in position i of a sequence. f is a sender function: a function from states to sequences of signals. This can capture a single sender, or multiple acting either independently or in concert. g is a receiver function: it decodes the sequence of signals to one of the states X.[7] Because id_X is the identity

[6] *Proof:* Note first that g must be a surjection and f an injection. Without the former, there would be an $x \in X$ that is not $g(\overrightarrow{m})$ for any $\overrightarrow{m}$, and so $g \circ f \neq \mathrm{id}_X$. Without the latter, distinct points in X would get mapped to the same point in X by $g \circ f$. Now, suppose there were an $\overrightarrow{m}$ such that $\{g(\overrightarrow{m})\} \neq \bigcap_i f_i^{-1}(\overrightarrow{m})$. This can hold only if $\bigcap_i f_i^{-1}(\overrightarrow{m})$ contains more than one element, since $g(\overrightarrow{m})$ has to belong to the intersection. This entails that there is another point $x \neq g(\overrightarrow{m})$ for which $f(x) = \overrightarrow{m}$, contradicting the injectivity of f. □

[7] I have focused on deterministic, not probabilistic, senders/receivers because of assumption (A2). Even if f, g could in principle be stochastic (i.e. have probability distributions as their range), this assumption would entail that optimal agents assign all their mass to a single point and so are effectively deterministic. Nevertheless, pursuing stochastic gener-

function on X, mapping each point to itself, that $g \circ f = \mathrm{id}_X$ means that optimal communication has been achieved, in the sense that the receiver always recovers the true state from X. Under that assumption, the result says that the receiver interprets a complex message (a sequence) by *intersecting* the independent meanings of each signal in the sequence (represented by $f_i^{-1}(\overrightarrow{m})$).

This result identifies three assumptions that cannot all be maintained if one wants to model the emergence of non-trivial composition, which I have just argued is a necessary step for explaining the emergence of function words. Not every approach makes all three of these assumptions. In particular, Steinert-Threlkeld (2014, 2016a) as well as Barrett et al. (2018) drop (A3). In these models, not every message is a sequence of the same length. In the former, one sender can choose whether or not to prefix a set of signals with an additional signal. In the latter, two senders choose *whether or not* to send a signal, so messages can be either of length one or two. In either case, the message space is a union, not a product (i.e. not of the form $\prod_i M_i$ for any sets M_i), and so the limitative result does not apply.

In the remainder, I will develop a model which maintains (A2) and (A3) but drops the assumption (A1) of a *fixed* set of states that the agents communicate about. That is: the context in which the agents are communicating will vary. Against that backdrop, there will be a role for non-trivial composition to play.

---

alizations of this result, likely using conditional independence, is a worthwhile endeavor. Thanks to an anonymous referee for discussion.

2 A Signaling Game with Varying Contexts

I will introduce a type of signaling game (Lewis, 1969; Skyrms, 2010)—called the Extremity Game—with a few helper definitions. Following the literature on gradable adjectives (Kennedy and McNally, 2005; Kennedy, 2007), I will assume that objects have some number of gradable properties, where each property has a corresponding *scale*. A scale in turn is a set of *degrees*, totally ordered with respect to a dimension. For example, the size of a circle corresponds to its radius, with degrees being positive real numbers (i.e. $\mathbb{R}^+$). For the degree of an object o on a scale s, I will write $s(o)$. Given a set S of scales, I will define a context as follows.

(6) A *context* c over scales S is a set of objects such that: for each $o \in c$, there is a scale $s \in S$ such that either o has the least degree on s ($o = \arg\min_{o' \in c} s(o')$) or the highest degree on s ($o = \arg\max_{o' \in c} s(o')$).

In its general form, the game takes place between a sender and a receiver in the following way.

(7) Extremity Game, in general:

a. Nature chooses a context c and a target object $o \in c$.

b. The sender sees c and o and sends a message m from some set of messages M.

c. The receiver sees c and m and chooses an object o' from c.

d. The play is successful (and the two agents equally rewarded) if and only if $o' = o$.

To fully specify a game, one must say what the messages M available are and how the agents make their choices. I will specify the former now and the latter in the next section. The set of available messages will be inspired by the semantics for gradable adjectives.

There, it is assumed that adjectives map objects (of type e) on to their degree on the corresponding scale (of type d). Morphemes like *-est* and *least* then map a contextually specified set of objects to those with the highest and lowest degrees, respectively.

(8) Toy semantics for a gradable adjective and superlative morphemes.

a. $[\![\text{size}]\!] = \lambda x.s_{\text{size}}(x)$

b. $[\![\text{-est}]\!]^c = \lambda P_{\langle e,d\rangle}.\lambda x_e.x \in c \text{ and } \forall x' \in c, P(x) \succeq P(x')$

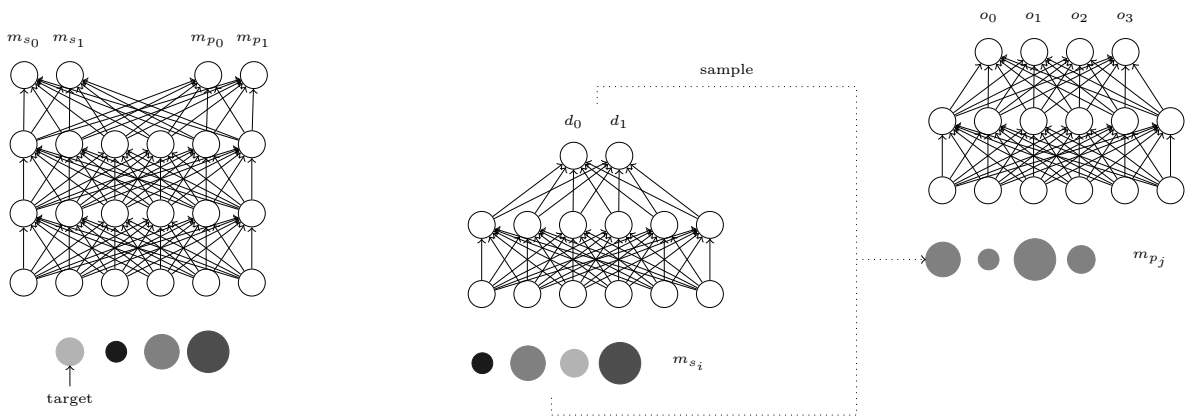
c. $[\![\text{least}]\!]^c = \lambda P_{\langle e,d\rangle}.\lambda x_e.x \in c \text{ and } \forall x' \in c, P(x) \preceq P(x')$

In contexts as defined in (6), having one expression for each scale and the morphemes *-est* and *least* will suffice to uniquely pick out each object in the context. I will assume, then, that the set of messages $M = M_S \times M_P$ where M_S is a set of size $|S|$ (i.e. there are as many messages in M_S as there are gradable properties for each object) and M_P is a set of size two (P for 'polarity').

3 Experiment

A trial of our experiment will consist of some number of iterations of playing an Extremity Game as in (7). The sender and receiver are each neural networks, schematically depicted in Figure 1. They are trained using the REINFORCE algorithm (Williams, 1992; Sutton and Barto, 2018). This algorithm applies standard reinforcement learning logic—successful choices become more likely and unsuccessful ones less likely—to neural networks.

There are two types of receiver: Basic and Attentional. The Basic one is a multilayer perceptron, taking the context and the signals, outputting a distribution over target objects, from which a sample is taken to determine the reward.



(a) Basic Sender (b) Attentional Receiver

Figure 1: Schematic depictions of network architectures.

The Attentional Receiver uses an *attention mechanism* (Mnih et al., 2014; Xu et al., 2015) to focus on a perceptually salient dimension. They implement *hard* attention in the following sense. First, they receive as input the context c and the message m_{s_i} from M_S chosen by the sender. On this basis, the receiver *chooses a dimension to attend to*: the input is filtered so that the agent only sees the objects according to one dimension (e.g. size or lightness). Then, the agent uses this attended-to dimension and the message from M_P chosen by the sender to choose a target object. This attention mechanism reflects the perceptual salience of the gradable properties of the objects: it is very natural, for instance, in the contexts depicted in Figure 1, to attend only to the size or the shade of the circles. Visual attention in humans exhibits a similar sensitivity to saliency along perceptual dimensions like color and shade (Theeuwes, 1992; Nothdurft, 1993, 2000).

I varied the number of dimensions (i.e. gradable properties) between 1 and 3, and ran 10 trials for each (for five-, twenty-, and fifty-thousand mini-batches respectively, where each mini-batch was 64 iterations). I recorded the rolling accuracy over 10 training steps, as well as the accuracy and detailed properties about contexts and signals used on 5000

new games at the end of training. The code and data can be found at `https://github.com/shanest/function-words-context`.

Results: Basic Receiver Mean communicative success per number of dimensions on 5000 novel games is provided in Table 1. In one and two dimensions, the agents reliably learned to communicate effectively. In three dimensions, they usually get stuck in sub-optimal protocols.

dims	mean	std
1	0.975	0.006
2	0.985	0.003
3	0.731	0.062

Table 1: Success on test, Basic Receiver.

Inspection of the learned communication protocols also show that they do not learn to treat either of the signals as a function word. The learned systems are always 'maximally' separating in the following sense: for any two contexts c, c' and targets o, o', if $o = \arg\min_c s_d(o)$ and $o' = \arg\max_{c'} s_d(o)$ for the same dimension d, then the sender's message for o in c differs from its message for o' in c' in both syntactic positions. This holds true for both the 2- and 3-dimensional cases. In such a system, the agents are not grouping context/target pairs according to the dimension along which the target can be singled out as maximal or minimal.

This could be for roughly the following reason: in expectation, target objects that differ only in whether they are the minimum/maximum in context on the same dimension will actually be farther from each other in Euclidean space than from other objects. So the sender could be using maximally different signals for the two types of target objects to help the receiver distinguish them.

Results: Attentional Receiver Mean communicative success per number of dimensions on 5000 novel games is provided in Table 2. In one and two dimensions, the agents reliably learned to communicate effectively. In three dimensions, we find a lower mean and higher variance. Visual inspection shows that many trials wind up near 88% communicative success, while others get stuck in very sub-optimal communication protocols.

Analyzing the resulting communication protocols yields promising results. Figure 2 shows an example learned communication system for a three-dimension trial. These are bar plots, showing the frequency with which the sender made various choices on the test games. The left column corresponds to M_S, and the right to M_P. The top row corresponds to the true dimension of the target object in context, and the bottom row to the true polarity of the target object. The top-left corner in each case shows that the different signals in M_S are being used to reliably communicate the dimension. The bottom-right corner in each figure shows that the signals in M_P are reliably being used to communicate the polarity of the object.

dims	mean	std
1	0.959	0.005
2	0.964	0.005
3	0.697	0.144

Table 2: Success on test, Attentional Receiver.

When the agents are communicating in this way, the signals that communicate direction can be interpreted as function words. The signals in M_S reliably communicate a bit of 'content': a dimension. The signals in M_P reliably signal whether the target has the greatest/lowest degree *along that dimension* of all the objects in the context. This is non-trivial modification of one linguistic item by another. The resulting communication protocols behave exactly like the toy semantics in (8).

4 Conclusion

Every natural language divides the lexicon into content and function words. The latter provide the 'grammatical glue' that enables robust forms of compositional communication to arise. Most existing approaches to the evolution of compositionality do not explain the emergence of function words. In this paper, I provided a diagnosis for this situation and introduced a signaling game with variable contexts consisting of multiple objects with multiple gradable properties. Simple reinforcement learning by neural networks—in particular with the ability to pay attention to certain perceptually salient aspects of the input—in this game can generate expressions that are appropriately characterized as function and as content words.

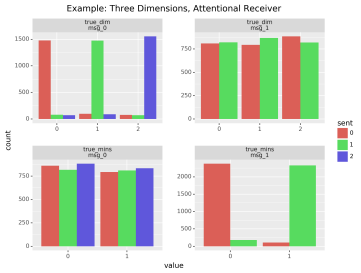


Figure 2: Example communication system with three dimensions.

Because this was a preliminary study, much work remains to be done. In particular, the model presented here builds in many assumptions that could possibly be relaxed in the end. For example, the syntactic role of the two signals—that those from M_P can see and act on those from M_S—has been built in. In a more general model, the specialization of signals into these syntactic roles could emerge. More generally, one would like communication systems like those exhibited here to emerge in the very general setting of communicating by a sequence of symbol, with costs for things like vocabulary size and length of messages. While attention to salient dimensions sufficed in the present case, exactly which costs and biases must be added to generate non-trivial composition in a general setting—and whether a unified explanation of all forms of non-trivial composition

can be given—remains for future work.

References

Barrett, J. A. (2007). Dynamic Partitioning and the Conventionality of Kinds. *Philosophy of Science 74*, 527–546.

Barrett, J. A. (2009). The Evolution of Coding in Signaling Games. *Theory and Decision 67*(2), 223–237.

Barrett, J. A., B. Skyrms, and C. Cochran (2018). Hierarchical Models for the Evolution of Compositional Language. In *26th Philosophy of Science Association Biennnial Meeting*.

Berthet, M., G. Mesbahi, A. Pajot, C. Cäsar, C. Neumann, and K. Zuberbühler (2018). Titi monkey alarm sequences: when combining creates meaning. In *26th Philosophy of Science Association Biennnial Meeting*.

Carnie, A. (2006). *Syntax: A Generative Introduction* (Second ed.). Oxford: Blackwell Publishing.

Carroll, L. (1871). *Through the Looking-Glass, and What Alice Found There*. Macmillan.

Cäsar, C., K. Zuberbühler, R. J. Young, and R. W. Byrne (2013). Titi monkey call sequences vary with predator location and type. *Biology Letters 9*(20130535), 2–5.

Frege, G. (1923). Logische Untersuchungen. Dritter Teil: Gedankengefüge ("Compound Thoughts"). *Beiträge zur Philosophie des deutschen Idealismus III*, 36–51.

Janssen, T. M. V. (1997). Compositionality. In J. van Benthem and A. ter Meulen (Eds.), *Handbook of Logic and Language*, Chapter 7, pp. 417–473. Elsevier Science.

Kennedy, C. (2007). Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy 30*, 1–45.

Kennedy, C. and L. McNally (2005). Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language 81*(2), 345–381.

Lewis, D. (1969). *Convention*. Blackwell.

Lewis, D. (1988). Relevant Implication. *Theoria 54*(3), 161–174.

Mnih, V., N. Heess, A. Graves, and K. Kavukcuoglu (2014). Recurrent Models of Visual Attention. pp. 1–12.

Mordatch, I. and P. Abbeel (2018). Emergence of Grounded Compositional Language in Multi-Agent Populations. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*.

Muysken, P. (2008). *Functional Categories*. Cambridge: Cambridge University Press.

Nothdurft, H. C. (1993). Saliency effects across dimensions in visual search. *Vision Research 33*(5-6), 839–844.

Nothdurft, H. C. (2000). Salience from feature contrast: additivity across dimensions. *Vision Research 40*(10-12), 1183–1201.

Nowak, M. A. and D. C. Krakauer (1999). The evolution of language. *Proceedings of the National Academy of Sciences 96*, 8028–8033.

Ouattara, K., A. Lemasson, and K. Zuberbühler (2009). Campbell's monkeys concatenate vocalizations into context-specific call sequences. *Proceedings of the National Academy of Sciences 106*(51), 22026–22031.

Pagin, P. and D. Westerståhl (2010a). Compositionality I: Definitions and Variants. *Philosophy Compass 5*(3), 250–264.

Pagin, P. and D. Westerståhl (2010b). Compositionality II: Arguments and Problems. *Philosophy Compass 5*(3), 265–282.

Partee, B. H. (1995). Lexical Semantics and Compositionality. In L. Gleitman and M. Liberman (Eds.), *Invitation to Cognitive Science, Part 1: Language*, Chapter 11, pp. 311–360. Cambridge: MIT Press.

Rizzi, L. and G. Cinque (2016). Functional Categories and Syntactic Theory. *Annual Review of Linguistics 2*(1), 139–163.

Schlenker, P., E. Chemla, K. Arnold, A. Lemasson, K. Ouattara, S. Keenan, C. Stephan, R. Ryder, and K. Zuberbühler (2014). Monkey semantics: two 'dialects' of Campbell's monkey alarm calls. *Linguistics and Philosophy 37*, 439–501.

Schlenker, P., E. Chemla, A. M. Schel, J. Fuller, J.-P. Gautier, J. Kuhn, D. Veselinović, K. Arnold, C. Cäsar, S. Keenan, A. Lemasson, K. Ouattara, R. Ryder, and K. Zuberbühler (2016). Formal monkey linguistics. *Theoretical Linguistics 42*(1-2), 1–90.

Schlenker, P., E. Chemla, A. M. Schel, J. Fuller, J. P. Gautier, J. Kuhn, D. Veselinović, K. Arnold, C. Cäsar, S. Keenan, A. Lemasson, K. Ouattara, R. Ryder, and

K. Zuberbühler (2016). Formal monkey linguistics: The debate. *Theoretical Linguistics 42*(1-2), 173–201.

Skyrms, B. (2010). *Signals: Evolution, Learning, and Information.* Oxford University Press.

Steinert-Threlkeld, S. (2014). Learning to Use Function Words in Signaling Games. In E. Lorini and L. Perrussel (Eds.), *Proceedings of Information Dynamics in Artificial Societies (IDAS-14).*

Steinert-Threlkeld, S. (2016a). Compositional Signaling in a Complex World. *Journal of Logic, Language and Information 25*(3), 379–397.

Steinert-Threlkeld, S. (2016b). Compositionality and competition in monkey alert calls. *Theoretical Linguistics 42*(1-2), 159–171.

Sutton, R. S. and A. G. Barto (2018). *Reinforcement learning: an introduction.* (Second Edi ed.). The MIT Press.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics 51*(6), 599–606.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning 8*(3-4), 229–256.

Xu, K., J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In F. Bach and D. Blei (Eds.), *International Conference on Machine Learning (ICML 32)*, pp. 2048–2057.

Zuberbühler, K. (2018). Combinatorial capacities in primates. *Current Opinion in Behavioral Sciences 21*, 161–169.