# Function Words and Context Variability*

Shane Steinert-Threlkeld
S.N.M.Steinert-Threlkeld@uva.nl

Draft of 30 October 2018. Comments Welcome!

> *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), emphasis mine.

The poem excerpted in the epigraph has often been called a 'nonsense poem'. After all: what does it mean? What is a slithy tove? What does it mean to be brillig or mimsy? Calling it nonsense, however, overlooks the amount of meaning we can extract from the emphasized words: minimally, a scene in the past is being described, which took place somewhere called a 'wabe'. The emphasized words are what are known as *function words*: they provide the 'grammatical glue' among the *content words*, which are indeed nonsense in this excerpt.

The distinction between these two types of expression occupies a central place in modern theoretical linguistics. Rightfully so: every natural language exhibits a distinction between function and content words. 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 steps towards answering these questions by presenting a simple model of trial-and-error language learning in which a division of signals into function and content words emerges. In the next section, I briefly but more explicitly introduce the distinction. In Section 2, I argue that a necessary condition for the emergence of the distinction is the presence of *non-trivial composition* (in a sense to be made precise). I present three case studies in which only trivial composition emerges and a mathematical result that diagnoses why that is the case. In Section 3, I introduce a new type of signaling game – the Extremity Game – in which the objects of communication vary from play to play. Amidst such variation, a distinction between function and content words could be useful. Section 4 reports an

*Acknowledgments to be added. This work was supported by funding from the European Research Council under the European Unions Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. STG 716230 CoSaQ.

experiment, in which artificial neural networks are trained by reinforcement learning to communicate in the Extremity Game. The emerging languages are analyzed: when the agents can pay attention to perceptually salient features of the context, they learn a system with complex signals that we can interpret as a gradable adjective plus a superlative morpheme (a prime example of a functional item). Section 5 concludes.

# 1 Functional and Lexical Categories

Modern theoretical syntax distinguishes between two broad types of syntactic categories: lexical and functional.[1] The former broadly correspond to the major parts of speech: nouns, verbs, adjectives, and adverbs. The latter are a bit more varied, but include:[2]

- Prepositions: 'in', 'above', 'from', 'to', ...
- Determiners: 'a', 'the', 'every', 'some', 'many', ...
- Conjunctions: 'and', 'or', ...
- Complementizers: 'that', 'if', 'whether', ...
- Tense:
  - Auxiliaries: 'have', 'is', 'was', ...
  - Modals: 'will', 'would', 'can', 'might', 'ought', ...

Exactly characterizing the distinction remains tricky. After observing that lexical categories have 'contentful' meaning, while functional categories have 'grammatical' meaning,[3] it is usually observed that the former constitute *open classes* and the latter *closed classes*. Roughly: one can very readily introduce new nouns and verbs to a language as needed. By contrast, trying to introduce a new preposition 'belove' meaning partially above and partially below would be quite difficult. Kaplan (1978) famously tried to introduce a new expression 'dthat', which rigidly referred to the satisfier of a description. Though he ably demonstrated the use of such a tool, that it never caught on can be partially attributed to the fact that demonstratives belong to a closed class. For the purposes of this paper, these distinctions suffice to point to the intended contrast.

Before proceeding, it's worth highlighting that the field of semantics – the scientific study of linguistic meaning – roughly divides itself along the lexical/functional line as well. The tradition descending from Montague via Partee and many others, usually called formal semantics, studies specifically compositional semantics. A survey of the textbooks in this field[4] shows that the major expressions studied come exactly from the functional categories.

[1] See, for example, pp. 43-46 of the textbook Carnie (2006).

[2] This list is incomplete and meant to be illustrative only. There are some debates about exactly which category certain expressions belong to, but they are orthogonal to present concerns.

[3] See, e.g., Carnie (2006) and Rizzi and Cinque (2016). Muysken (2008) is a thorough overview of functional categories.

[4] For example, Heim and Kratzer (1998) and Jacobson (2014).

Lexical semantics – the study of the meanings of basic expressions – studies at length the meanings of individual expressions and groups thereof in the lexical categories.[5] Seen in this light, explaining the emergence of the distinction between functional and lexical categories occupies a central role in the broader explanation of the emergence of compositionality.

## 2 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. After presenting this argument, I will analyze three case studies from the literature on the evolution of compositionality which exhibit only trivial composition. The reasons for this are then made precise in the form of a triviality result: 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.[6] 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:[7]

(1) A communication system is *trivially compositional* just in case complex expressions 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.[8] 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 $B$s, and raptors in the canopy elicit a sequence of $A$s followed by a sequence of $B$s. While the full details do not concern us,[9] Schlenker, Chemla, Schel, et al. (2016a) 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

[5] See, for example, Levin and Rappaport Hovav (2005).
[6] Frege (1923), Janssen (1997), Pagin and Westerståhl (2010a), and Pagin and Westerståhl (2010b).
[7] For this use, see Schlenker, Chemla, Schel, et al. (2016b) and Zuberbühler (2018).
[8] Cäsar et al. (2013) and Schlenker, Chemla, Schel, et al. (2016a).
[9] See Steinert-Threlkeld (2016b) for some reservations about the full analysis.

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).[10]

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.[11] They have two basic alarm calls: an eagle call *hok* and a general alert *krak*.[12] Moreover, both calls combine with what appears to be a suffix *-oo*, which has the effect of weakening the severity of the calls. Schlenker, Chemla, Schel, et al. (2016a) 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 emergence of function words. Recall the characterization thereof as 'grammatical glue': they precisely 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.

Before proceeding, we note that the presence of non-trivial composition does not suffice for the presence of function words. To see this, consider subsective adjectives.[13] 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.

[10] Berthet et al. (2018) argue that the proper semantics for Titi calls is not in fact trivially compositional. Nevertheless, the presentation just given illustrates what such a system would look like.

[11] Ouattara, Lemasson, and Zuberbühler (2009) and Schlenker, Chemla, Arnold, et al. (2014).

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

[13] Partee (1995).

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.

## 2.1 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) and Barrett (2009) studies a generalization of signaling games[14] with multiple senders. In the simplest case, there are four states of nature and two sender, 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.[15] 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. Such a 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.[16] 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.

[14] Lewis (1969) and Skyrms (2010).

[15] See, e.g., Lewis (1988).

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

## 2.2 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.

It turns out that under these assumptions, there's a mathematical sense in which optimal communication will be trivially compositional. This is captured in the following result:

(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.

$$\text{If } g \circ f = \mathrm{id}_X, \text{ then for all } \overrightarrow{m}, \{g(\overrightarrow{m})\} = \bigcap_i f_i^{-1}(\overrightarrow{m})$$ [17]

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 $X$. Because $\mathrm{id}_X$ is the identity 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) and Steinert-Threlkeld (2016a) as well as Barrett, Skyrms, and Cochran (2018) drop (A3). In these models, not every message is a

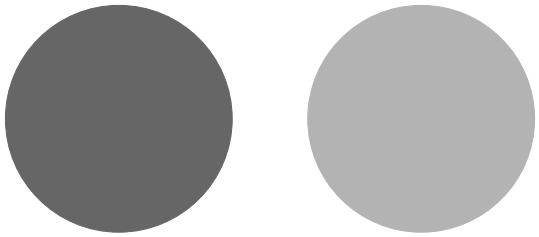
[17] *Proof:* Note first that $g$ must be a surjection and $f$ an injection. Without the former, there would be an $x \in X$ that is not $g(\overrightarrow{m})$ for any $\overrightarrow{m}$, and so $g \circ f \neq \mathrm{id}_X$. Without the latter, distinct points in $X$ would get mapped to the same 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 $f$. □

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.

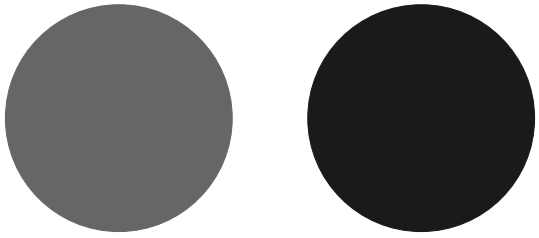
In the remainder, I will develop a model which maintains (A2) and (A3) but drops the assumption (A1) of a *fixed* set of states that the agents communicate about. That is: the context in which the agents are communicating will vary. Against that backdrop, there will be a role for function words to play.

# 3 A Signaling Game with Varying Contexts

The variant on the signaling game that I will use to illustrate the emergence of function words will have the agents talking about varying sets of objects with multiple *gradable properties*. To get a feel for the kind of task involved, consider the following adaptation of an example from Graff (2000).[18] Suppose that we are both looking at the following two circles, drawn on top of a table.



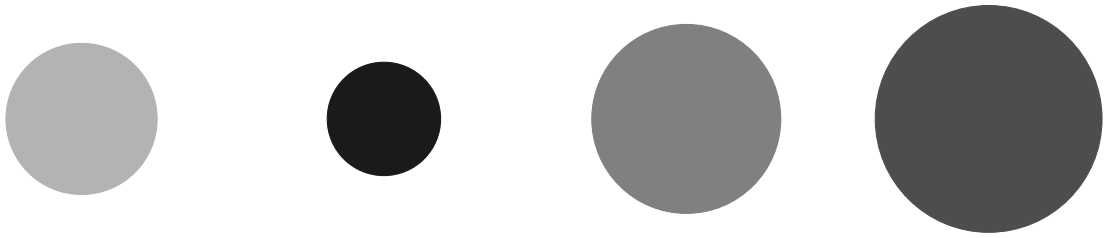
For whatever reason, you need me to put something on the left circle. You might say "put it on the *darker circle*". By contrast, suppose that you had the same communicative needs, but now the circles on the table looked as follows.



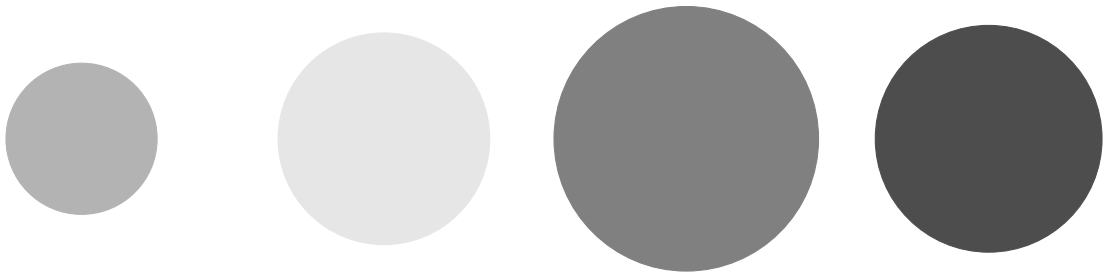
Now, to tell me to put it on the left circle, you might say "put it on the *lighter circle*", or a bit more circuitously "put it on the less dark one".

The target referent of your communication – the circle on the left – has exactly the same size and shade in both contexts. But in one context, it's dark*er* than the other circle, while in the other context, it's light*er*. (For the purposes of illustration, we can assume that you could not refer to the circles by their spatial position or demonstratively. If you'd like: your friend is looking at a picture on a screen that may have been scrambled.) Finally, we can imagine similar situations with more than one gradable property. Suppose, again, that you need to communicate about the leftmost circle in the following array.

[18] See Syrett, Kennedy, and Lidz (2010) for a study using similar contexts with children.



Here, it's natural to call the leftmost circle "the lightest one". Now consider the following context.



Here, you are likely to refer to the circle on the left as "the smallest one". These situations have the following structure: in each context, each object has two very salient gradable properties: a size (radius) and a darkness. These dimensions distinguish the target object: it has either the largest or the smallest value in one of those dimensions. By drawing attention to that fact, one can successfully refer to it. Moreover, you can do so in a very economical way: with labels for the properties and morphemes like the superlative *-est* (and its corresponding negative counterpart, 'the least'), successful communication is ensured. This is done without talking about specific degrees of size or of lightness and in a way in which an object with exactly the same degrees on all relevant properties will be referred to in different ways in different contexts.

I will convert communicative scenarios like the above into a type of signaling game – called the Extremity Game – with a few helper definitions. Following the literature on gradable adjectives,[19] 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 $s$ ($o = \arg\max_{o' \in c} s(o')$).

At its most 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 $M$.

[19] See, for instance, Kennedy and McNally (2005) and Kennedy (2007) and the references therein.

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 the subset with the highest and lowest degrees.

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

Now, for the crucial observation: 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'). The players of an Extremity Game will be able to successfully communicate if they can learn to associate each message in $M_S$ with a distinct scale and the two signals in $M_P$ with something akin to *-est* and *least*. As advertised, this setup meets two of the three assumptions in the limitative result (5) – (A2) optimal communication is correct identification of a target object and (A3) messages come from a product space – but drops (A1): because the context varies from play to play of the game, there is no fixed set of objects about which the agents communicate.[20]

# 4 Experiment

The goal is to show how a simple semantic system like 8 could emerge via a simple dynamics among agents playing an Extremity Game. In particular, we will use *reinforcement learning*:[21] agents make choices, receive some reward (in our case, for successful communication of the target object in context), and adjust their behavior so that they are more likely to make the corresponding choices in the future.

While most approaches to reinforcement learning in signaling games use a variant of a simple algorithm called Roth-Erev learning,[22] such an algorithm will not suffice for present purposes. On this approach, choices are reinforced entirely independently of one another. Two factors of the present setup require a stronger method. On the practical side, there is a combinatorial explosion that comes from having variable contexts with multiple objects that have multiple gradable properties: there are so many contexts that most of them will not be seen often enough for such an algorithm to be effective. On the conceptual side, if

[20] While the agents in an intuitive sense communicate 'about' a fixed set of objects – all objects with $|S|$ gradable properties – each communicative exchange concerns a different subset thereof.

[21] Sutton and Barto (2018)

[22] Roth and Erev (1995)

choices are reinforced entirely independently, there will be no pressure for signals to emerge that group objects based on the degrees of various properties and their relative position on scales in context.

To overcome this limitation, I will use a type of agent with a built-in capacity for stimulus generalization: artificial neural networks.[23] This choice was made because such networks provide a simple, widely used, and somewhat biologically plausible model that has the capacity to generalize. Other approaches to stimulus generalization in learning in signaling games use a method called *spill-over*.[24] In that framework, not only are the actual choices reinforced, but so too are *similar* choices in similar choice points. Exactly how reinforcement works thus depends on definitions of similarity between choices and between states. While some domains provide natural such definitions,[25] it is not immediately obvious how to define how similar one context-target pair is to another in an Extremity Game. Neural networks will learn to treat certain pairs as similar and others not, without the theorist having to hard-wire a definition of similarity into the learning model.[26]

## 4.1 Methods

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, the simplest in a family of methods known as policy gradient methods.[27] The intuition behind this algorithm is just as before. Consider the sender. The sender is a policy that takes as input a context and a target and outputs a probability distribution over messages (in this case, two distributions: one over $M_S$ and one over $M_P$). The sender's policy is parameterized by the weights and biases that connect the neurons in the network. Thanks to what is known as the policy gradient theorem, modern variants of stochastic gradient descent can be used to adjust the weights and biases in a way that is guaranteed to make positively reinforced actions sampled from the policy more likely in the future.

We varied the number of dimensions (i.e. gradable properties) between 1 and 3, and ran 10 trials for each. We trained for five-, twenty-, and fifty-thousand mini-batches respectively, where each mini-batch was size 64. In other words, the agents play 64 games in between each update of their policies; this reduces the variance in learning. We also experimented with two different neural architectures for the receiver – called Basic and Attentional – for reasons that will become clear in what follows. We 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.

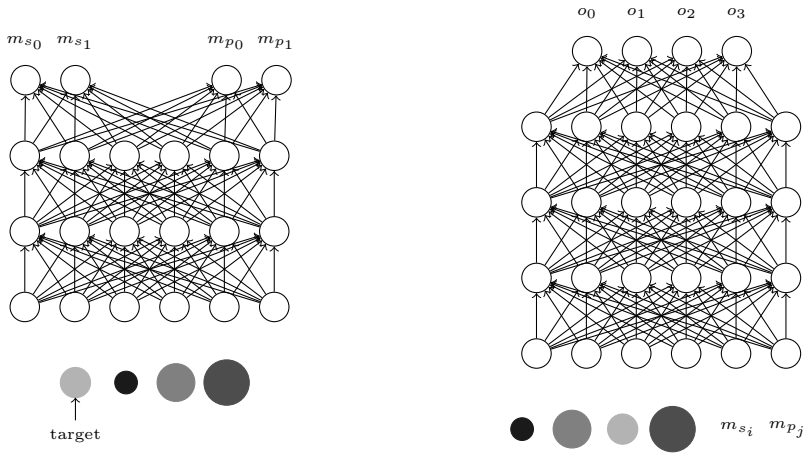
[23] Nielsen (2015) and Goodfellow, Bengio, and Courville (2016)

[24] See O'Connor (2014). The name 'spill-over' comes from Franke (2016).

[25] For instance, if the goal is to choose a point on a line, the distance between the true point and the guessed point is very natural.

[26] See Lazaridou, Peysakhovich, and Baroni (2017) for a similar approach, which inspired the present one. Their contexts consist of two natural images, one of which is the target. The sender chooses one signal from a fixed-sized vocabulary to send to the receiver. While they are interested in whether natural concepts emerge in such a setting, I am focused on less natural input but more complex communication structures in order to explore the emergence of functional vocabulary.

[27] Williams (1992). See chapter 13 of Sutton and Barto (2018) for a modern introduction.




(a) Basic Sender (b) Basic Receiver

Figure 1: Schematic depictions of basic network architectures. The input is on the bottom, followed by a sequence of hidden layers to output layers on the top. The output neurons produce probabilities of choosing the action written above them.

Complete details of the network architectures and training set-up are included in an Appendix. The code and data can be found at https://github.com/shanest/function-words-context.

## 4.2 Results: Basic Receiver

The learning curves over training for each trial of each dimension, with the Basic Receiver, are plotted in Figure 2. As can be seen, in the one- and two-property cases, the agents learn to communicate nearly perfectly in a relatively short amount of training steps. By contrast, in the three-dimension case, the agents do not regularly achieve a high degree of communicative success after 50000 mini-batches. The mean success rates on 5000 new games at the end of training time are reported in Table 1.

In the one-dimensional case, the context consists of two objects that have a property to different degrees. The successful communication protocol that the agents learn to use reliably sends one signal when the target has the lower degree and the other signal when the target has the higher degree.

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

Table 1: Accuracies on novel games.

In the two-dimensional case, things are not quite as aligned with expectations. Figure 3 shows a typical communication protocol that emerges in the two dimensional case. The colored bars correspond to the particular signals sent. The left column corresponds to $M_S$ and the right column to $M_P$The colored bars correspond to the particular signals sent. The left column corresponds to $M_S$ and the right column to $M_P$. In the top row, the $x$-axis corresponds to the 'true' dimension of the target object (i.e. the dimension for which the target had an extreme value in context). In the bottom row, the $x$-axis corresponds to the 'true' polarity of the target object (i.e. whether it had the true property to the least or highest degree).

The bottom-left cell shows an interesting pattern: the message from $M_S$ sent always


Figure 2: Learning curves for basic sender and receiver.

corresponds to the true polarity (minimum or maximum). This is because one message is always sent when the true polarity is 0 (minimum) and the other when the polarity is 1 (maximum). Unfortunately, that the top row shows no such separation implies that no signal is being used to communicate the 'true' dimension. The equal heights of all the bars in the top row imply that the two messages in $M_S$ (left column) and in $M_P$ (right column) are used an equal number of times when the true dimension is 1 and when the true dimension is 0.

In fact, closer inspection reveals the following: the learned communication 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 the 3-dimensional case as well. Figure 4 shows an example learned system. The bottom-right cell shows that the agents do use $M_P$ to distinguish the true direction of the target. But the top-left cell shows that the agents do not associate different signals in $M_S$ with different dimensions: rather, they separate targets in the way just described.

These results show that basic senders and receivers do not, under the REINFORCE algorithm, learn to communicate in accord with the toy semantics in (8). One might think that one of the messages still looks like a superlative morpheme, since it reliably correlates

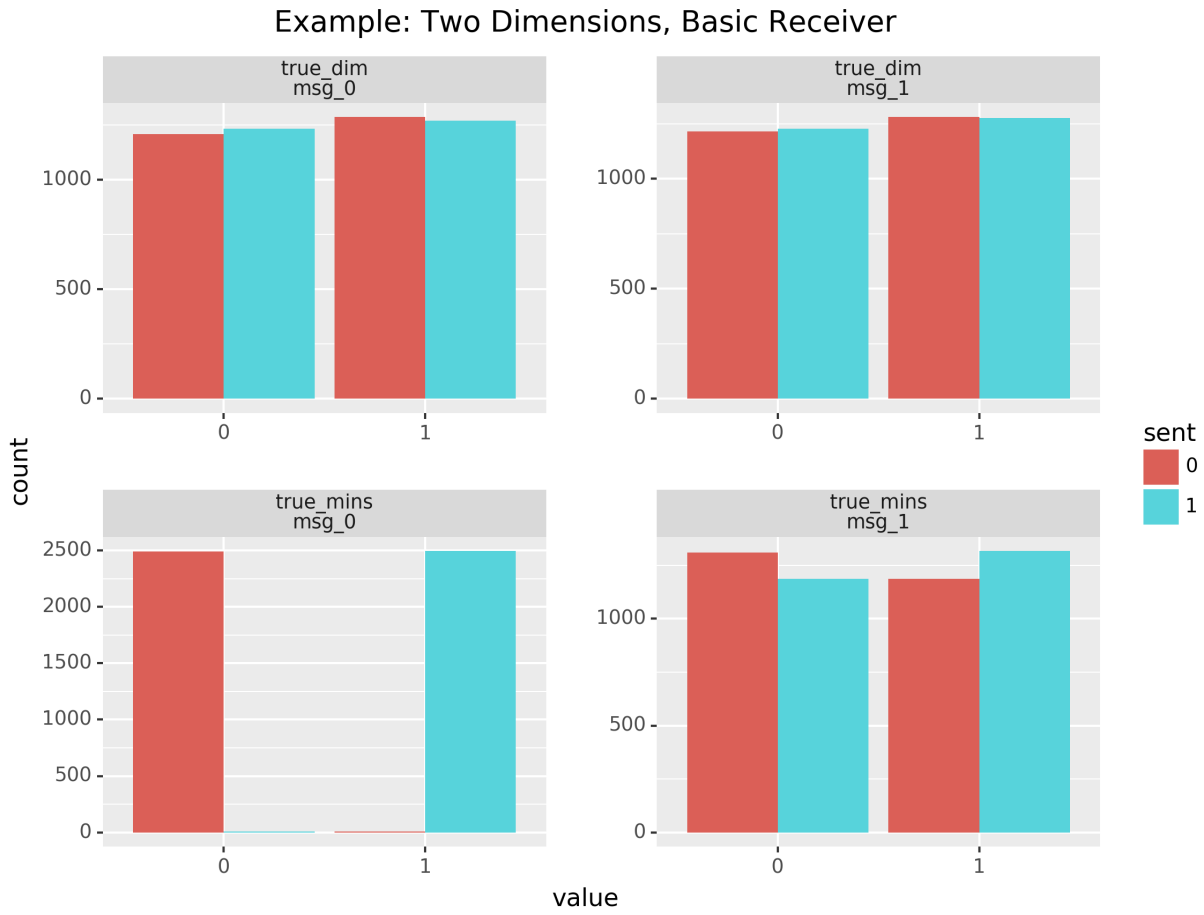



Figure 3: Example communication system with basic receiver and two dimensions.

with the true direction of the target object. While this is indeed very interesting and does show that the networks are clustering objects on the basis of their direction (for example, they never separate on dimension and group together based on direction), given that they do not use the other signal to communicate the true dimension, it does not look like there's non-trivial modification of one linguistic item by another.

## 4.3 Results: Attentional Receiver

Intuitively, the networks are not learning to use a signal to group objects together based on dimension. 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 a dimension will actually be farther from each other in Euclidean space than from other objects. Because of this, it could be that the agents use maximally different signals for the two types of target objects.

To help the agents learn to communicate based on the dimension, I will use what is known as an *attention mechanism* in machine learning.[28] Intuitively, a neural network can

[28]See, for instance, Mnih et al. (2014) and Xu et al. (2015).

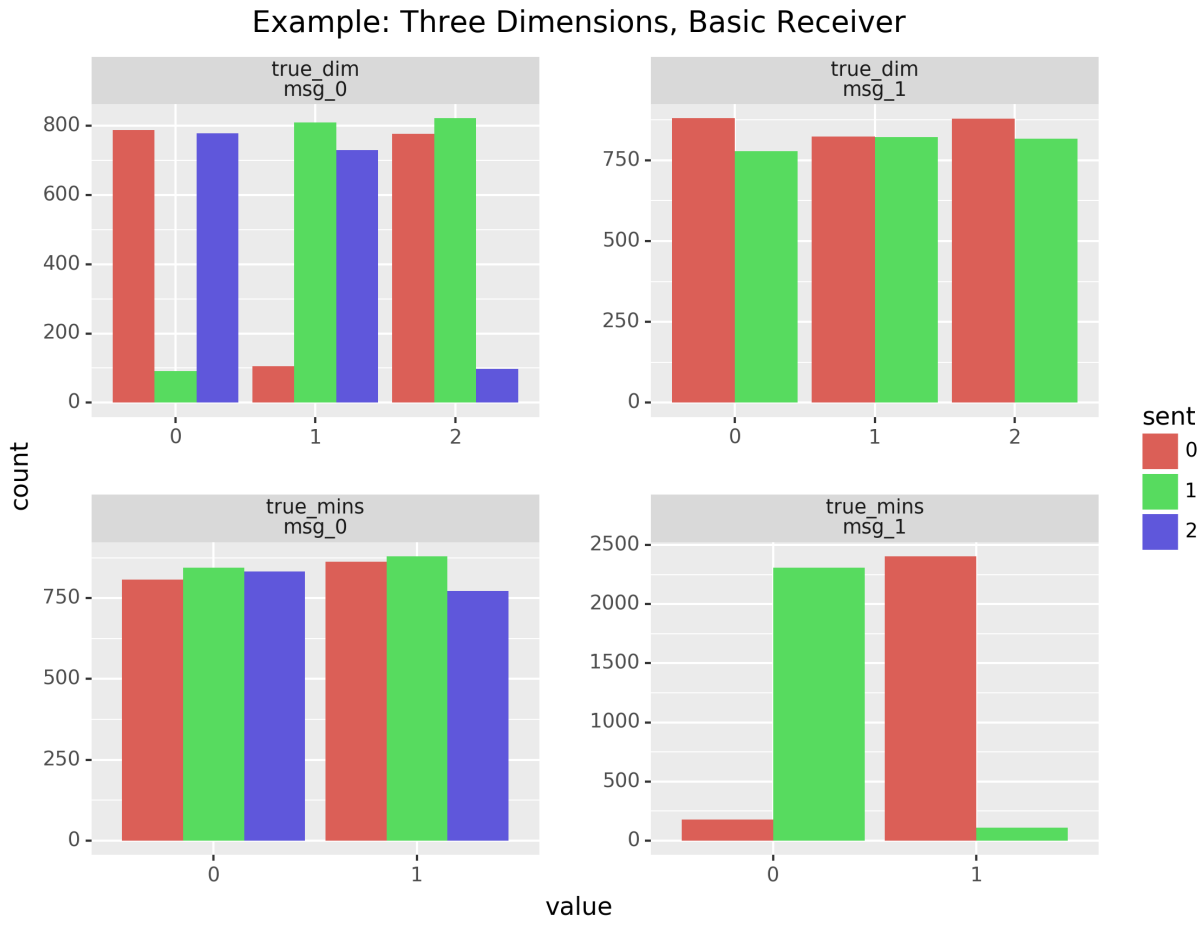



Figure 4: Example communication system with basic receiver and three dimensions.

learn to pay more or less attention to different portions of its input. The network (or a sub-component thereof) computes a weighting of the input positions which is then used to filter the actual input. The weight can be 'hard' – selecting a sub-region of the input – or 'soft' – re-weighting the input so that different nodes are more or less attended to than in the raw input.

One can think of attention as reflecting something like perceptual salience: the network can learn to focus its attention on salient features of its input, since those features are likely to help it solve its task. For instance, a neural image caption generator with attention will likely focus its attention on well-defined objects in an input image. These salient objects are likely to help it generate a plausible caption.

Attentional Receivers, as I will develop them, implement a hard attention mechanism 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 in Section 3, to attend only to the size or the shade of the circles. Figure 5 depicts this architecture.

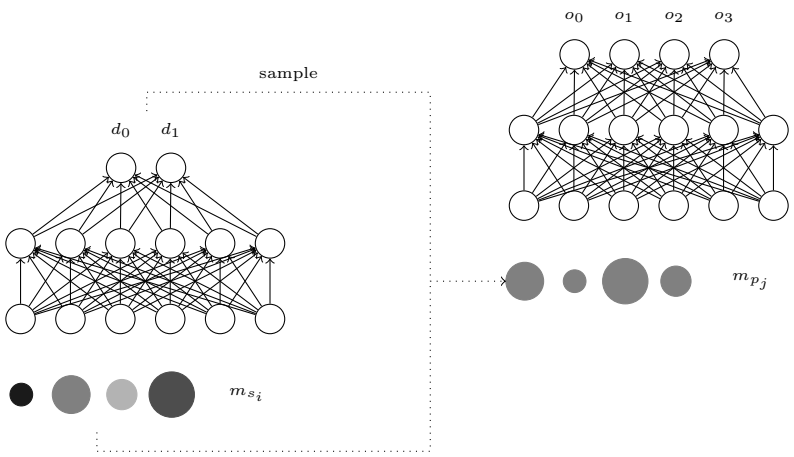



Figure 5: Attentional Receiver architecture, schematically. The receiver first chooses a dimension to attend to, then chooses a target based only on that dimension. In this schematic, the chosen dimension is size; differences in shading have been washed out by the attention mechanism.

The learning curves over training for each trial of each dimension – but with a Basic Sender and Attentional Receiver – are plotted in Figure 6. The mean success rates on 5000 new games at the end of training time are reported in Table 2. As before, in the one- and two-property cases, the agents learn to communicate nearly perfectly in a relatively short amount of training steps. In all cases, it appears that learning is a bit slower than with basic receivers. This makes perfect sense: an attentional receiver has to learn two types of choices to make, as opposed to just one. In the three dimensional case, the attentional receiver achieves a high-degree of accuracy more frequently than the basic receiver, but also gets stuck in sub-optimal states more frequently.

The resulting communication protocols behave exactly like the toy semantics in (8). Figure 7 shows an example protocol in two dimensions. Here, the top-left cell shows that the choice of signal from $M_S$ reliably communicates the true dimension: when the dimension is 0, the sender chooses $m_{s_0}$ and when the dimension is 1, the sender chooses $m_{s_1}$. Similarly, the bottom-right cell shows that the choice of signal from $M_P$ signal reliably communicates the true direction (i.e. whether the target has the relevant property to the largest or smallest degree). Figure 8 shows an example learned communication system in three dimensions. Again, in complex signals, one signal communicates a dimension, and the other communicates whether the target has the most or least degree on the corresponding scale.

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

Table 2: Accuracies on novel games.

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


Figure 6: Learning curves for basic sender and attentional receiver.

modification of one linguistic item by another. Thus, when the receiver knows to use one of the signals to attend to a particular dimension in context, the two agents can learn to use their signals in a non-trivially compositional way.

# 5 Conclusion

Let us take stock. After introducing the distinction between functional and lexical categories, I argued that there are in principle reasons why many extant models of the evolution of compositionality cannot explain the emergence of function words: given their assumptions, they can only explain trivial composition; but non-trivial composition is a necessary precondition for the presence of function words. I then introduced a signaling game with variable contexts consisting of multiple objects with varying 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.

Much work remains to be done. One would like neural architectures that make fewer assumptions about what aspects of the input the receiver pays attention to. A first step in this direction will be to use a soft, as opposed to hard, attention mechanism. A more

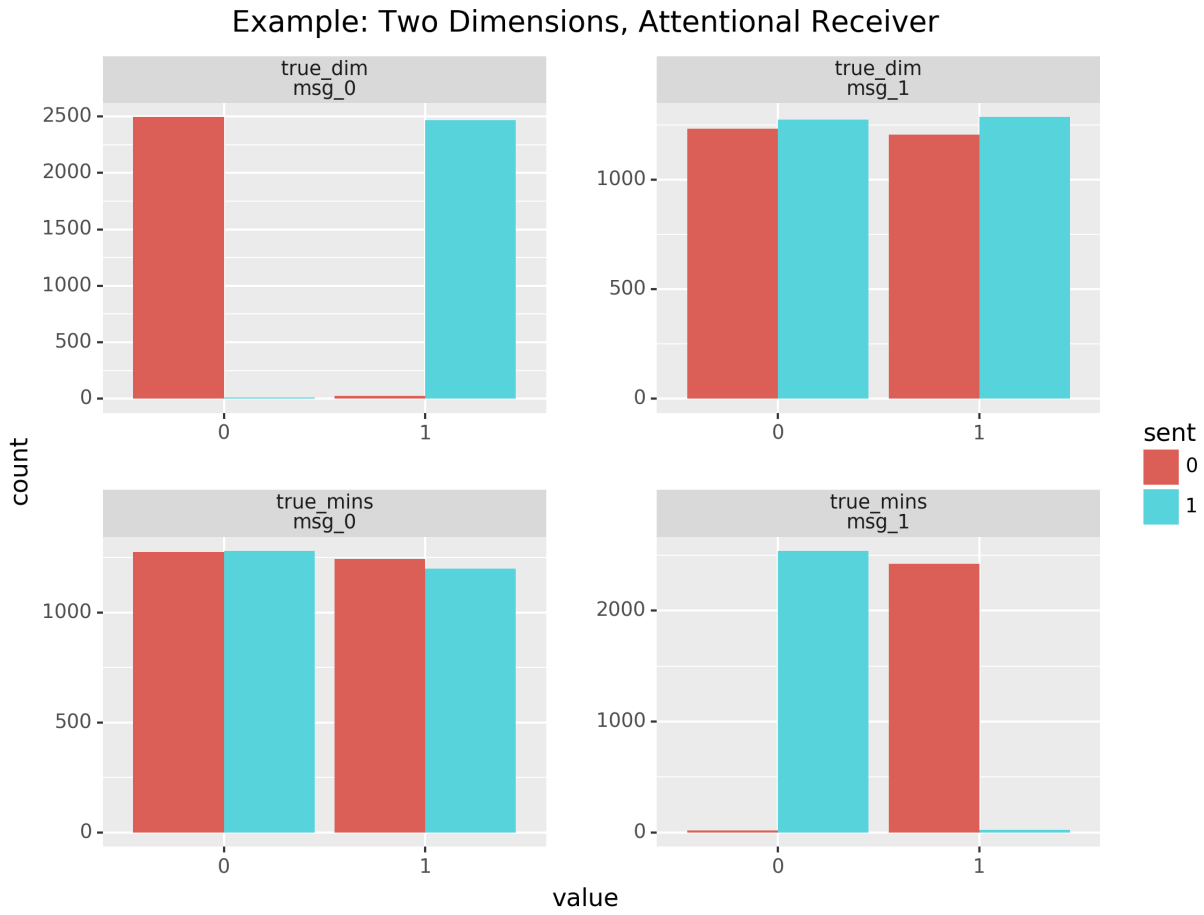



Figure 7: Example communication system with attentional receiver and two dimensions.

thorough hyper-parameter search may also generate more reliable learning results in the higher-dimensional setting. One can also generalize the input so that the networks also have to discover *which dimensions* are relevant for being able to successfully refer to objects across contexts, instead of having it built into the current definition of context. More generally, one would like communication systems like those exhibited here to emerge in the very general setting of communicating by a sequence of symbols with costs for things like vocabulary size and length of messages. All of these exciting avenues remain to be pursued in future work.

# References

Barrett, Jeffrey A (2007). "Dynamic Partitioning and the Conventionality of Kinds". In: *Philosophy of Science* 74, pp. 527–546.

— (2009). "The Evolution of Coding in Signaling Games". In: *Theory and Decision* 67.2, pp. 223–237. DOI: 10.1007/s11238-007-9064-0.

Barrett, Jeffrey A, Brian Skyrms, and Calvin Cochran (2018). "Hierarchical Models for the Evolution of Compositional Language". In: *26th Philosophy of Science Association Biennnial Meeting*.

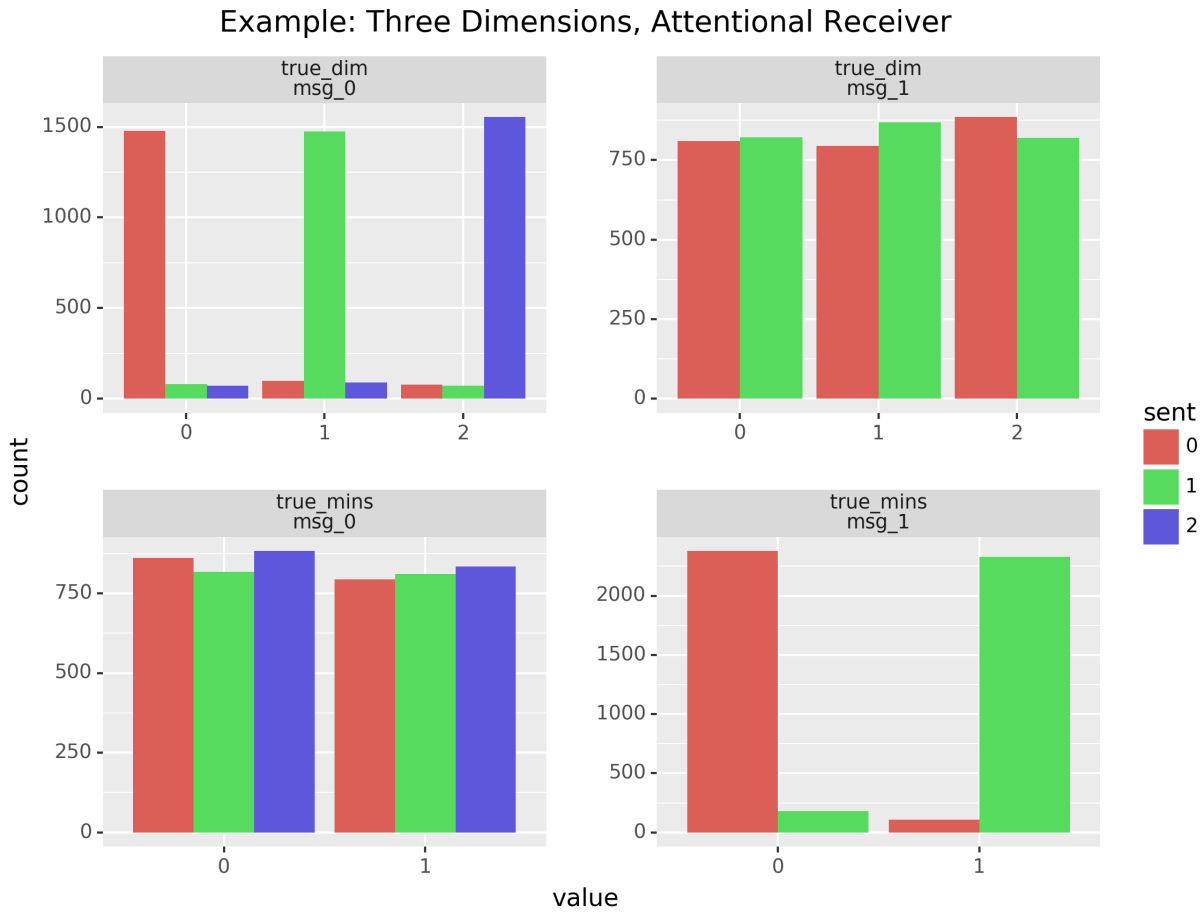



Figure 8: Example communication system with attentional receiver and three dimensions.


Berthet, Mélissa et al. (2018). "Titi monkey alarm sequences: when combining creates meaning". In: *26th Philosophy of Science Association Biennnial Meeting*.

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

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

Cäsar, Cristiane et al. (2013). "Titi monkey call sequences vary with predator location and type". In: *Biology Letters* 9.20130535, pp. 2–5. DOI: 10.1098/rsbl.2013.0535.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter (2016). "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)". In: *International Conference of Learning Representations*. URL: http://arxiv.org/abs/1511.07289.

Franke, Michael (2016). "The Evolution of Compositionality in Signaling Games". In: *Journal of Logic, Language and Information*. DOI: 10.1007/s10849-015-9232-5.

Frege, Gottlob (1923). "Logische Untersuchungen. Dritter Teil: Gedankengefüge ("Compound Thoughts")". In: *Beiträge zur Philosophie des deutschen Idealismus* III, pp. 36–51. DOI: 10.1093/mind/LI.202.200.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. The MIT Press. URL: https://www.deeplearningbook.org/.

Graff, Delia (2000). "Shifting Sands: An Interest-Relative Theory of Vagueness". In: *Philosophical Topics* 28.1, pp. 45–81.

Heim, Irene and Angelika Kratzer (1998). *Semantics in Generative Grammar*. Blackwell Textbooks in Linguistics. Wiley-Blackwell.

Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167.

Jacobson, Pauline (2014). *Compositional Semantics: An Introduction to the Syntax/Semantics Interface*. Oxford Textbooks in Linguistics. Oxford Univeristy Press.

Janssen, Theo M V (1997). "Compositionality". In: *Handbook of Logic and Language*. Ed. by Johan van Benthem and Alice ter Meulen. Elsevier Science. Chap. 7, pp. 417–473. DOI: 10.1016/B978-044481714-3/50011-4.

Kaplan, David (1978). "Dthat". In: *Syntax and Semantics*. Ed. by Peter Cole. Vol. 9. New York: Academic Press, pp. 212–233.

Kennedy, Christopher (2007). "Vagueness and grammar: the semantics of relative and absolute gradable adjectives". In: *Linguistics and Philosophy* 30, pp. 1–45. DOI: 10.1007/s10988-006-9008-0.

Kennedy, Christopher and Louise McNally (2005). "Scale Structure, Degree Modification, and the Semantics of Gradable Predicates". In: *Language* 81.2, pp. 345–381.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *International Conference of Learning Representations (ICLR)*. arXiv: 1412.6980. URL: https://arxiv.org/abs/1412.6980.

Lazaridou, Angeliki, Alexander Peysakhovich, and Marco Baroni (2017). "Multi-Agent Cooperation and the Emergence of (Natural) Language". In: *International Conference of Learning Representations (ICLR2017)*. arXiv: 1612.07182. URL: http://arxiv.org/abs/1612.07182.

Levin, Beth and Malka Rappaport Hovav (2005). *Argument Realization*. Cambridge University Press.

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

— (1988). "Relevant Implication". In: *Theoria* 54.3, pp. 161–174. DOI: 10.1111/j.1755-2567.1988.tb00716.x.

Mnih, Volodymyr et al. (2014). "Recurrent Models of Visual Attention". In: pp. 1–12. arXiv: 1406.6247. URL: http://arxiv.org/abs/1406.6247.

Mordatch, Igor and Pieter Abbeel (2018). "Emergence of Grounded Compositional Language in Multi-Agent Populations". In: *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*. URL: http://arxiv.org/abs/1703.04908.

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

Nielsen, Michael A (2015). *Neural Networks and Deep Learning*. Determination Press. URL: http://neuralnetworksanddeeplearning.com/.

Nowak, Martin A and David C Krakauer (1999). "The evolution of language". In: *Proceedings of the National Academy of Sciences* 96, pp. 8028–8033.

O'Connor, Cailin (2014). "Evolving Perceptual Categories". In: *Philosophy of Science* 81.5, pp. 840–851.

Ouattara, Karim, Alban Lemasson, and Klaus Zuberbühler (2009). "Campbell's monkeys concatenate vocalizations into context-specific call sequences." In: *Proceedings of the National Academy of Sciences* 106.51, pp. 22026–22031. DOI: 10.1073/pnas.0908118106.

Pagin, Peter and Dag Westerståhl (2010a). "Compositionality I: Definitions and Variants." In: *Philosophy Compass* 5.3, pp. 250–264. DOI: 10.1111/j.1747-9991.2009.00228.x.

— (2010b). "Compositionality II: Arguments and Problems." In: *Philosophy Compass* 5.3, pp. 265–282. DOI: 10.1111/j.1747-9991.2009.00229.x.

Partee, Barbara Hall (1995). "Lexical Semantics and Compositionality". In: *Invitation to Cognitive Science, Part 1: Language*. Ed. by Lila Gleitman and Mark Liberman. Cambridge: MIT Press. Chap. 11, pp. 311–360.

Rizzi, Luigi and Guglielmo Cinque (2016). "Functional Categories and Syntactic Theory". In: *Annual Review of Linguistics* 2.1, pp. 139–163. DOI: 10.1146/annurev-linguistics-011415-040827.

Roth, Alvin E. and Ido Erev (1995). "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term". In: *Games and Economic Behavior* 8, pp. 164–212.

Schlenker, Philippe, Emmanuel Chemla, Kate Arnold, et al. (2014). "Monkey semantics: two 'dialects' of Campbell's monkey alarm calls". In: *Linguistics and Philosophy* 37, pp. 439–501. DOI: 10.1007/s10988-014-9155-7.

Schlenker, Philippe, Emmanuel Chemla, Anne M Schel, et al. (2016a). "Formal monkey linguistics". In: *Theoretical Linguistics* 42.1-2, pp. 1–90. DOI: 10.1515/tl-2016-0001.

Schlenker, Philippe, Emmanuel Chemla, Anne M Schel, et al. (2016b). "Formal monkey linguistics: The debate". In: *Theoretical Linguistics* 42.1-2, pp. 173–201. DOI: 10.1515/tl-2016-0010.

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

Steinert-Threlkeld, Shane (2014). "Learning to Use Function Words in Signaling Games". In: *Proceedings of Information Dynamics in Artificial Societies (IDAS-14)*. Ed. by Emiliano Lorini and Laurent Perrussel.

— (2016a). "Compositional Signaling in a Complex World". In: *Journal of Logic, Language and Information* 25.3, pp. 379–397. DOI: 10.1007/s10849-016-9236-9.

— (2016b). "Compositionality and competition in monkey alert calls". In: *Theoretical Linguistics* 42.1-2, pp. 159–171. DOI: 10.1515/tl-2016-0009.

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

Syrett, K., C. Kennedy, and J. Lidz (2010). "Meaning and Context in Children's Understanding of Gradable Adjectives". In: *Journal of Semantics* 27.1, pp. 1–35. DOI: 10.1093/jos/ffp011.

Williams, Ronald J (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine Learning* 8.3-4, pp. 229–256.

Xu, Kelvin et al. (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *International Conference on Machine Learning (ICML 32)*. Ed. by Francis Bach and David Blei, pp. 2048–2057. arXiv: 1502.03044. URL: https://arxiv.org/abs/1502.03044.

Zuberbühler, Klaus (2018). "Combinatorial capacities in primates". In: *Current Opinion in Behavioral Sciences* 21, pp. 161–169. DOI: 10.1016/j.cobeha.2018.03.015.

# A Full Experiment Details

For each number of dimensions $n$, a context has $2n$ objects. Each object is specified by $n$ real numbers, chosen uniformly at random from the interval $(0, 2)$ at steps of 0.1. The values are uniformly subtracted by 1 to center them around 0.

The sender thus has $2n^2$ input nodes. As a convention, the first object for the sender is always the target. It has two hidden layers of 64 nodes each, with exponential linear activation.[29] The final hidden layer is then passed through two linear layers, with output sizes $|M_S|$ and 2, respectively. These are batch normalized[30] and fed into a softmax, to generate distributions over $M_S$ and $M_P$.

The Basic Receiver receives the context, but with the objects in a random order compared to the sender, and two signals sampled from the sender's output distributions, encoded as one-hot vectors. It then has three rectified linear hidden layers of 64, 64, and 32 units respectively. Then a final linear layer with $2n$ output nodes (one for each target object) is passed through batch normalization and softmax to generate a distribution.

The Attentional Receiver passes the context and a messaged from $M_S$ sampled from the sender through one exponential linear layer of 64 units, before batch normalization and softmax of size $n$, one for each dimension. A sample is taken from this distribution. The corresponding scalar values for each object along the dimension, together with a message sampled from the sender's distribution over $M_P$ are passed through exponential linear layers of size 64 and 32, before batch normalization and softmax produce a distribution over target objects.

We trained using the REINFORCE algorithm, with mini-batches of size 64, and the Adam optimizer[31] with learning rate $5 \cdot 10^{-4}$. For $n = 1, 2, 3$ dimensions, and each type of receiver, we ran 10 trials of 5000, 20000, and 50000 mini-batches of training. After training, the trained networks then played 5000 versions of the game; the signals chosen, the target chosen, whether it was correct, and what the 'true' dimension and direction (min/max) for identifying the target in context were recorded.

Everything was implemented in PyTorch. The code and data are available at https://github.com/shanest/function-words-context.

[29] Clevert, Unterthiner, and Hochreiter (2016)
[30] Ioffe and Szegedy (2015)
[31] Kingma and Ba (2015)