Fall 12-15-2013

# A Multi-Agent, Connectionist Model of Metonymy Production

Robert J. Powers
*University of Connecticut - Storrs*, bopowers9@gmail.com

# A Multi-Agent, Connectionist Model of Metonymy Production

Robert J. Powers
University of Connecticut
December, 2013

Advised by Whitney Tabor

## Abstract

Metonymy, like metaphor, has received much attention in cognitive linguistics literature (Croft, 1993; Kövecses & Radden, 1998; Panther & Radden, 1999). Most experimental work focuses on comprehension. However, why a speaker would choose to produce a metonym in some cases and not others is not fully understood. Connectionist models are well-suited to deal with the partial-semantic/partial-syntactic information which is characteristic of metonymy. Moreover, these models can capture some of the complex, time-varying dynamics of language development. Here, a model is presented in which dyads of artificial agents (each a recurrent neural network based on Rogers, et al., 2004) were trained in a structured environment to develop a language and use it to coordinate in a simple task. The model was informed by an experiment with human subjects based on existing work (Clark & Wilkes-Gibbs, 1986; Selten & Warglien, 2007). While inconclusive, the study offers some insight into the strengths and weaknesses of this particular approach to modeling metonymy production.

*Keywords:* metonymy, connectionist, neural network, semantics, pragmatics

## Acknowledgments

I would like to thank Dr. Whitney Tabor for advising me on this project; Pyeong-Whan Cho for helpful discussions about the topic and about data analysis; The UConn Office of Undergraduate Research for generous SURF funding; members of SOLAB; and, last but not least, my family for their support.

This project used MikeNet, created by M.W. Harm, for building and running neural network simulations.

# A Multi-Agent, Connectionist Model of Metonymy Production

Metonymy has received plenty of attention in the field of cognitive linguistics (CL). Though originally overshadowed by the study of metaphor, by the late 1990s a growing consensus began to emerge that metonymy may actually be the more fundamental cognitive phenomenon (Kövecses & Radden, 1998; Panther & Radden, 1999). Despite this theoretical claim, there has been little psycholinguistics research on metonymy (Gibbs, 2007). This is partly because designing a task to reliably elicit or control for metonymy in natural language would be difficult, and also because there is a lack of models from which testable predictions might be made.

Metonymy can be notoriously hard to define in the first place, and this adds to its slipperiness (Barcelona, 2011). The most widely accepted view, and the one taken here, is stated in Panther & Radden (1999): "metonymy is a cognitive process where one conceptual entity, the *vehicle*, provides mental access to another conceptual entity, the *target*, within the same idealized cognitive model."

This definition can be illustrated with an example. Consider the statement, "Proust is tough to read" (Croft, 1993). A listener will likely have little trouble understanding that "Proust" refers not to Marcel Proust the person, but to something closely related to and salient about Proust—namely, his books. Thus, the vehicle in this case would be the concept of Proust, the author. The target, to which the vehicle provides mental access, is the set of books written by Proust.

Note that the vehicle provides mental access by virtue of how closely related the vehicle and target concepts are. Conceptual contiguity is a classic feature of metonymy. To describe this, many CL theorists appeal to domains. In the context of domains, metonymy can be seen as a process of domain highlighting (Croft, 1993). Barcelona states this more formally: “Domain highlighting consists in highlighting a secondary domain within the domain matrix constituted by a speaker’s encyclopedic knowledge of the meaning of a linguistic expression” (2011). Other authors sometimes use different terms other than “domain” (i.e. cognitive models), but the semantics of these distinctions do not concern the present study. To avoid equivocation, “domain” will be understood here to be closely related to the treatment in Chen (2011).

An additional complication that falls out of this definition of metonymy is that several different types are possible, depending on the relationship between different subdomains. Some examples are PRODUCT-FOR-PRODUCER (e.g. “He’s got a Picasso”), OBJECT-FOR-USER (e.g. “The sax has the flu today”), and PLACE-FOR-EVENT (e.g. “Watergate changed out politics”) (Lakoff & Johnson, 1980). Arguably one of the simplest types is PART-FOR-WHOLE, sometimes referred to as synecdoche. As this label implies, the vehicle in a PART-FOR-WHOLE is simply part of the target concept (e.g. using “wheels” to refer to a car).

Most experimental studies have focused on metonymy comprehension rather than production. One notable example, a study by Frisson and Pickering (1999), examined metonymy processing difficulty by tracking subjects’ eye movements. There have been some studies investigating production indirectly by examination of corpora (Handl, 2011), but to date there are no studies of which the author is aware in which metonymy can be reliably elicited in a controlled, laboratory setting. Such studies would be useful in addressing a simple question: why produce metonymic utterances at all? We currently have a stockpile of theories from CL (cf. Handl, 2011), but no methods to test them directly.

The present study represents an attempt at a “first-pass” model in the connectionist paradigm and a corresponding experimental task to fill the previously mentioned gap.

Chronologically, the experiment was designed and performed first so that some of the difficulties inherent in eliciting production from humans might be better understood. The design was based on some well-established work in psycholinguistics research relating to pragmatics and language emergence (Clarke & Wilkes-Gibbs, 1986; Selten & Warglien, 2007). These are discussed in more detail in the following sections.

**Common Ground**

A proper model of metonymy production should be able to replicate the conditions that license emergence of metonymic expressions in natural language. But speakers do not produce language in a vacuum; humans learn language from one another and share ideas through both linguistic and non-linguistic communication. Indeed, even differentiating production and comprehension is problematic, since a language user will often need to switch rapidly back and forth between these, perhaps even needing to be in both roles simultaneously. Importantly, it also seems that metonymy has a functional, referring role which requires some task or context (one in which the best referent needs to be decided upon). The most natural scenario for this is in ordinary conversation. The first desideratum for our model is that it captures this intuition.

Clark and Wilkes-Gibbs (1986) worked on a simple experimental paradigm to study precisely this kind of problem. They were interested in how language users collaborate to find *common ground*—an agreed upon set of meanings for linguistic expressions—while solving a task. This is relevant here for the following reason: if metonymy is produced because of the relationship of concepts in a particular conceptual domain, then interlocutors in both production and comprehension roles should have similar cognitive representations of that domain—or at least they should understand the organization of each other's domain. This is crucial, because we can imagine that if their domains structures were wildly different, then a metonym would probably not be interpreted correctly. So finding common ground can be understood for our purpose as the process of shaping domain representations so that they align well enough for metonymy to work.

The Clark and Wilkes-Gibbs study used two participants who were seated so that they could not see one another, but they could hear each other. Both were given a set of 12 images of ambiguous figures, composed of tangrams. One subject was designated as the *director* and the other was the *matcher.* On each trial, the director had the images placed in a certain order, while the matcher's images were shuffled. It was the director's task to describe the images so that the matcher could place them in the correct order. The experiment ended after six trials.

They recorded large quantities of data, as subjects were allowed to freely describe the items. One of the key findings was that subjects tended to reduce the complexity of their descriptions as the experiment progressed. This is because the images never changed, so the subjects were simply negotiating the best descriptions. Over time, superfluous lexical items (like provisional phrases) and descriptors that were not agreed upon were dropped.

One minor observation was something Clark and Wilkes-Gibbs called *narrowing*. They defined narrowing as a "refinement in perspective" in which "the focus of a perspective was narrowed to just one part of a figure" (1986). For example, "the guy in the sleeping bag" could be reduced at a later time to just "sleeping bag." Note that "sleeping bag" is peripheral in the original description, but it later comes to represent the whole configuration. Because narrowing resembles metonymy in this way (particularly PART-WHOLE-METONYMY), this paradigm seems like it may be a good starting point for a task designed to elicit metonymy.

## Language Emergence

In order to understand the fundamental question of how metonymy emerges, it may be necessary to abstract away from the complexities of an already fully developed language, and to recreate conditions in which a language is dynamically shaped by its users in the course of language development.

Prior work in experimental semiotics (ES) could be helpful in this regard. Galantucci (2005) and Selten & Warglien (2007) are two notable examples. Like the tangram studies, these were in the context of cooperative tasks. In Galantucci’s study, participants created and communicated with novel orthographic symbols. In Selten & Warglien (2007), participants communicated with a predetermined set of symbols. In both cases, the participants had to develop meanings for symbols on their own, effectively creating their own language.

The biggest downside of the original tangram approach for our present purposes is that most cases of metonymy in natural language seem to be conventionalized (Handl, 2011). They are already fixed, and we may not learn much about how they became lexicalized in the first place. Stripping away natural language semantics by incorporating some of the ES methodology would allow the lexicalization process to be more directly observed. In this case, selection of a language medium becomes an important concern. Also, some complications may be introduced in the analysis, since interpreting meaning in an artificial language, especially one with allowed ambiguity, is not a trivial task.

## Connectionist Networks

*Parallel Distributed Processing*, or PDP (Rumelhart & McClelland, 1986a) laid the foundation for a class of models known as connectionist models. Not all connectionist models are the same, but many are based on the premise that representations, rather than being atomic in nature, may be distributed over an entire network of units. The activation state of a connectionist network is the vector of its unit activations. Connectionist models, using assemblies of simple processing units, are often referred to as *artificial neural networks*, or simply *neural networks*.

These neural networks have some interesting properties. They are designed to have an element of biological plausibility, and they are relatively robust against damage. By adding recurrence, these networks can often take on additional properties that make them suitable to simulating language tasks, including tasks involving semantics. Specifically, recurrent models with the right processing dynamics can form attractors (see Hinton & Shallice, 1991; Plaut & Shallice, 1993; Harm & Seidenberg, 1999, 2004). In dynamical systems, attractors are points in a space towards which solutions move over time. In recurrent neural networks, repeated training can cause attractors to develop in the activation state space. This property was used, for instance, by Plaut & Shallice (1993) to create semantic attractors, corresponding to semantic representations (concepts) in studies on deep dyslexia.

The importance of attractor dynamics to the present study is that they may present a natural implementation of conceptual contiguity in domains, which is essential to metonymy (for a sense of this similarity, compare the diagrams of domains in Chen, 2011 to the diagrams of semantic attractors in Plaut & Shallice, 1993). If two attractors are close to one another in the semantic space, and the state of the system approaches this configuration, then noise may push the state into one of two possible end behaviors.

The very notion of distributed representations could also be useful in understanding the cognitive processing behind metonymy. It is easy to imagine, in the case of PART-FOR-WHOLE metonymy, that a subset of the features of an intended “whole” becomes activated, and the “part” corresponding to this subset gets named instead. Thus, the combination of distributed representations and attractor dynamics makes the connectionist approach an attractive one for the metonymy production problem.

## Operational Definition and Objectives

The CL definition of metonymy for natural language is inherently complicated. For all analyses in this paper, a clear, consistent definition of metonymy was needed that would be reasonable for a simple, artificial language. The Clarke & Wilkes-Gibbs (1986) definition of narrowing was taken to be the basis for the present operational definition of metonymy. To observe such narrowing would be a good indication that a simple form of PART-FOR-WHOLE metonymy was being used (because of the simplicity and compositional structure of PART-FOR-WHOLE metonymy, we will restrict our attention to only this type for the remainder of this paper). An operational definition of metonymy can therefore be stated as follows.

**Definition 1.** A word, $\omega$, is used metonymically if and only if it meets the following three criteria:

i) Word $\omega$ is used in conjunction with one or more other words at some time, $t$, to describe object $X$.

ii) Word $\omega$ is used *by itself* to describe $X$ at some time, $t'$, $t' > t$.

iii) There exists a consistent, one-to-one mapping from $\omega$ to some feature of $X$.

Note that this definition does not require metonymy to persist. In natural languages, it is surely the case that not all instances of linguistic innovation become lexicalized. What is of interest here is the emergence of metonymy as a function of language change. While this definition would certainly allow a metonym to become lexicalized, it would also allow metonymy to be used as a transient linguistic tool, created to fill a specific need at a particular point in time.

This definition ignores syntax, as the order (if it exists) of symbols is not taken into account. This is likely unrealistic in the case of natural language, especially if syntactic form and content are linked as in construction grammars (Fillmore, Kay, & O’Connor, 1988). However, the absence of syntax in this model is motivated by the primary concern of this study; we are interested here in the simplest conditions that may lead to the emergence of metonymy. If metonymy could occur without syntactic constraints, then this would indeed be the simplest case, and different syntactic constructions may then serve to augment the use of metonymy. Since this is the first study, to our knowledge, designed to address these questions in this way, the simplest case is preferred. An improvement in future studies would be to allow the role of syntax to be explored. Connectionist models could still apply, as there are recurrent neural network architectures which have been shown to be capable of learning syntactic structure (Elman, 1990).

For the present study, the main research question was this: what are the conditions that license the emergence of metonymy in a language as it changes over time? The objective was to construct a model utilizing connectionist principles that would be capable of simulating the production of metonymy as two agents learn and simultaneously shape the semantics of a shared language. The behavior of the model was evaluated by testing a specific hypothesis and comparing the results to human data. The null hypothesis was that the presence of contrastive features in the language users’ environment has no influence on the frequency of metonymical expressions—in other words, that contrastiveness does not constrain which lexical items are likely to function as metonyms. The alternative hypothesis is that contrastive features in the environment are more likely to be described metonymically in the language than non-contrastive ones.

This paper is organized in two main sections. Experiment I describes the preliminary, exploratory study with human subjects. Results from Experiment I were used not only to provide a baseline for statistical hypothesis testing, but also to inform the implementation of the model. Both the model and some analyses of its output are described under Experiment II. Finally, general discussion summarizes the results of both experiments.

## Experiment I

### Method

**Participants.** Sixty-six undergraduate students enrolled in psychology courses at the University of Connecticut participated in the experiment for course credit.

**Materials and Design.** Materials consisted of a two-dimensional, virtual environment and an alphabet of symbols. The environment was populated with *objects*, each object being composed of three *features*: shape, color, and pattern. Each feature could take on one of three values: shapes were either square, circle, or triangle; colors were red, blue, or yellow; and patterns were vertical lines, crosses, or curved lines. These nine features could be combined into a total of twenty-seven possible objects. See Figure 1.

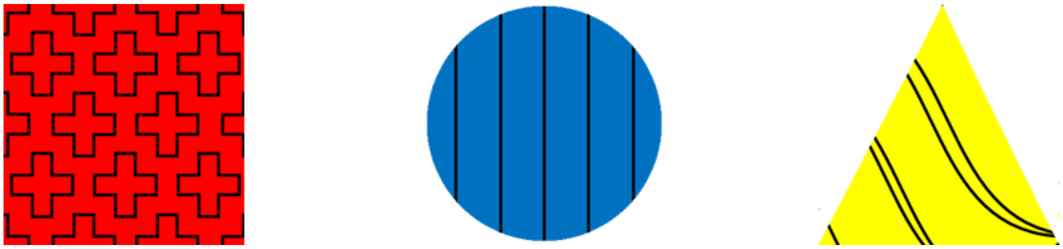

**Figure 1. Three of the possible 27 objects used in Experiment I.**

The task required participants to develop a language. Following previous work in ES (Galantucci, 2005; Selten & Warglien, 2007), we restricted the language to only an orthographic component. As in Selten & Warglien (2007), a simple alphabet was fixed before the present study began. This alphabet was designed around three base symbols, Ж, Ҕ, and ԃ, from the Cyrillic alphabet. These three characters were selected by the experimenter with the assumption that they were both unfamiliar to native English speakers, and also relatively dissimilar to one another. These base symbols were also given variants that were systematically related across the base forms. The first variant of each symbol was the same as the base form, but with an added diacritic line over the top of the symbol. The second variant was a 90° counter-clockwise rotation of the base form, with no other changes. These variants were created to introduce systematicity to the primitives of the language, thus potentially offsetting the predicted difficult of the task. These base symbols along with their variants are demonstrated in Figure 2.

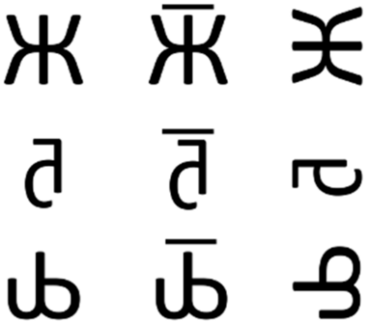

**Figure 2. The alphabet used in Experiment I. The first column contains the base symbols. Columns two and three display variants. In composing a message, only one symbol from each row was allowed.**

The experiment was organized in blocks. There were twelve blocks, each composed of six individual trials, for a total of seventy-two trials. A trial consisted of three objects being displayed in a two-dimensional plane on a computer screen. One of these objects was designated as the *target*, and the others *competitors*. For later analysis, three conditions (*all-contrastive, single-contrastive,* and *unconstrained*) were established based on the number of contrastive features between target and competitor objects. In the all-contrastive condition, no features were shared between any of the objects on screen, so that each feature dimension was contrastive. In single-contrastive, exactly two features were shared between the target and at least one of the competitors, and one feature was unshared. In the unconstrained condition, the number of shared features was equally likely to be zero, one, or two (not three, because duplicate objects were not allowed).

Randomization was involved in selecting which particular objects were displayed from trial to trial. A computer program, using a pseudo-random process, generated 81 potential trial combinations for each condition, making sure that any feature-sharing constraints were satisfied. Additionally, there was some counterbalancing: of the 81 potential trials in each condition, every possible object was a target exactly three times so that no single object would be over-represented in the trial sample space.

Finally, a virtual interface was built into the experiment program which allowed alphabet symbols to be manipulated and combined into strings. The interface was set up so that participants could use the computer mouse to drag symbols into a blank composition area. Symbols could be added to a message by clicking the desired symbol and dragging it into the composition area. Removing a symbol could be accomplished by dragging it outside of the composition area. Certain restrictions were imposed by the design of the interface; no symbol could be used more than once in a message, and the ordering of symbols was fixed and outside of the participants' control. Clicking on the symbol without dragging allowed one to cycle through its various forms. The interface can be seen in Figure 3.

**Procedure.** The experiment was run on multiple computers simultaneously in a single room. This setup allowed multiple subjects to be involved in each session. Each participant was seated at a computer running the experiment program, which was designed so that two programs could transmit information via internet connection. Unbeknownst to participants, experimenters randomly paired them together before the experiment began, and took care to situate each member of a dyad so that one could not see the other's screen. Each dyad was also assigned to one of the three conditions at the beginning of the experiment.

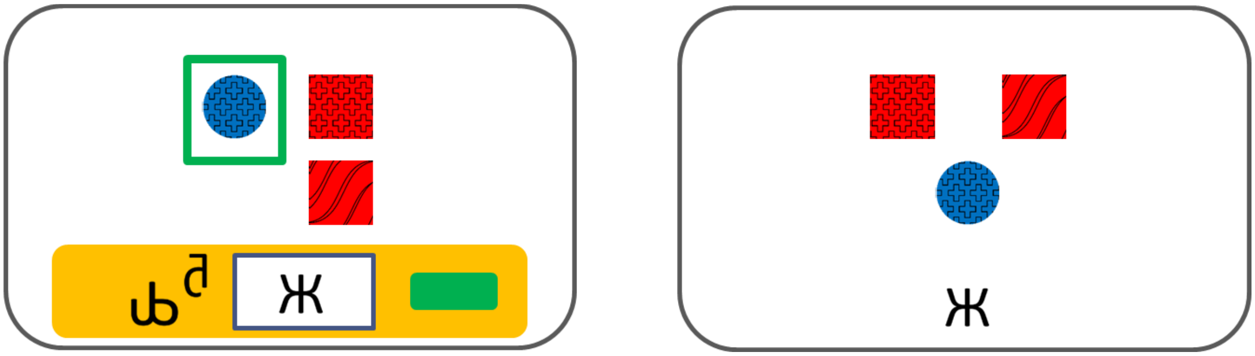

**Figure 3. The message composition interface in Experiment I. The image on the left is the producer's screen, and the image on the right belongs to the responder. The target object was highlighted for the producer (framed by a green box), and the producer used the message composition interface at the bottom of the screen. After clicking the green "send" button, the responder was able to see the same set of objects along with the message sent by the producer. The responder would click an object to end the trial. Feedback, given to both subjects, is not shown here.**

Participants were given both written and verbal instructions before the experiment began. They were told that they would be cooperating with an anonymous partner, and that they would be composing and sending messages back and forth to each other with the task of describing objects on their screens. Additionally, an initial practice session was included to familiarize participants with the message composition interface.

After completing the practice session, the first block was initiated. At the beginning of each block, participants were assigned to a role: either *producer*, or *responder* (corresponding to director and matcher respectively in Clark & Wilkes-Gibbs, 1986). The producer was tasked with composing and sending messages, and the responder would receive the producer's messages. Also, at the beginning of each block, participants had the opportunity to review the instructions relevant to their upcoming role.

Each trial began as follows. The responder's screen displayed only the message, "waiting for partner." The producer's screen was populated by the target and competitor objects with relative object positions randomized. The particular set of objects displayed was determined by drawing at random from the distribution of potential sets for the dyad's condition. For example, if the dyad was assigned to the all-contrastive condition, then any object was equally likely to be the target, and no features would be shared between any of the objects on screen.

The target object was indicated for the producer by the presence of a green rectangular border. The producer was instructed to compose a message, using the interface at the bottom of the screen, so that the responder would be able to identify the target correctly. After composing a message (no time limit was imposed), the producer could click a "send" button, and no further revision to the message was allowed. Next, the responder was shown the same objects—with relative positions also randomized—except that the target object was not indicated. Instead of seeing an interface, the responder was shown the producer's message, and was instructed to click on the object which the message described. After the responder clicked an object, both participants were given feedback as to whether or not the responder chose correctly.

At the start of every new block, participants switched roles. This was continued until all twelve blocks had been completed. The duration of the experiment varied, but no group took longer than the allotted forty-five minutes to complete the experiment.

## Results

By Definition 1 (parts i & ii), a necessary condition for metonymy was a *reduction* in message length, in which a single symbol used to describe an object had at some previous time been used as part of a longer description of the same object. Communication, in such circumstances, would presumably be most effective if only a single symbol were required to differentiate a target object from its competitors. It was found that, of the 2376 total trials across all dyads, 139 trials required at least two symbols to correctly identify the target object. These trials were discarded prior to analysis.

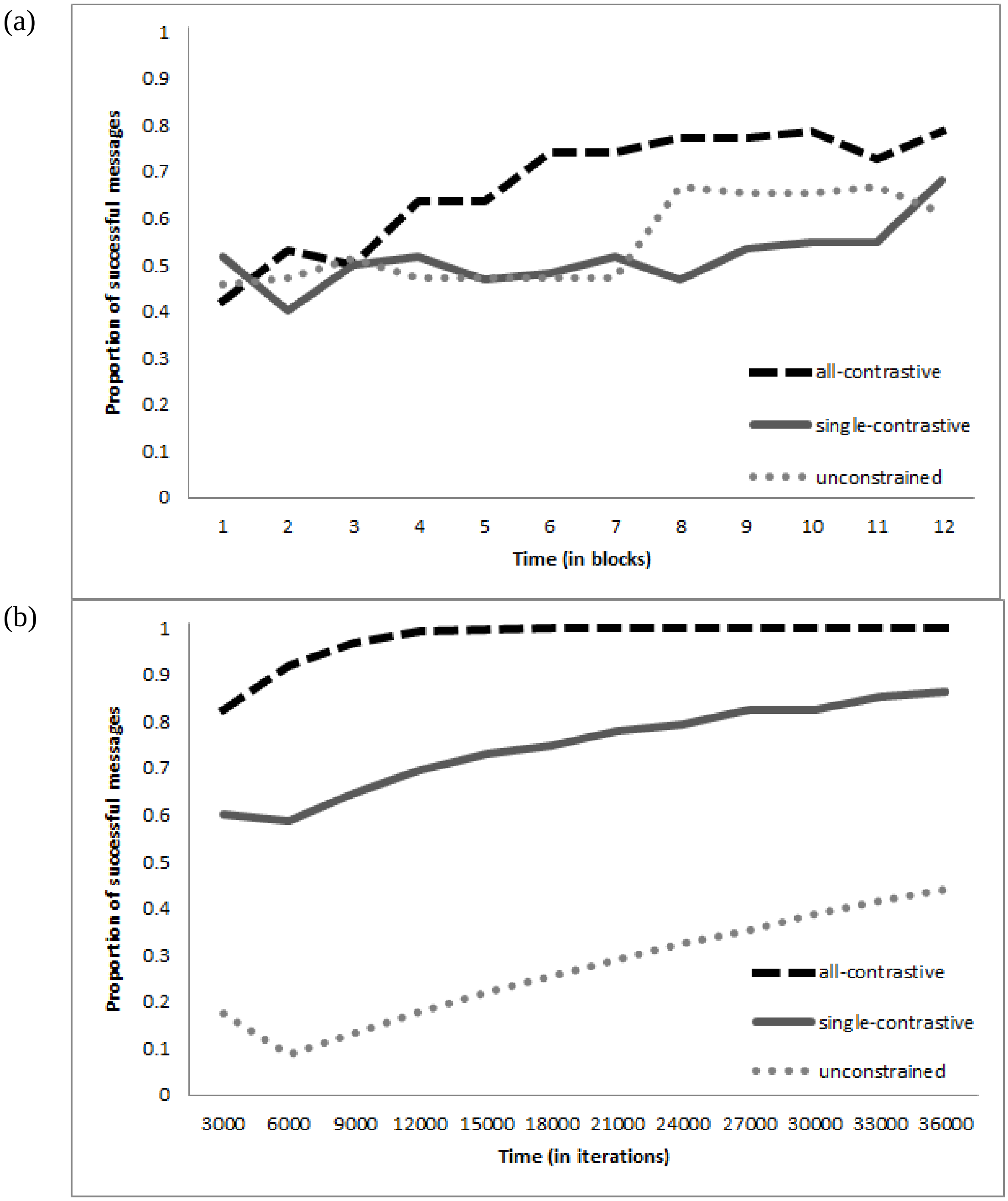

**Figure 4. Accuracy averaged by condition. (a) Human subjects. (b) Neural network model.**

Figure 4(a) displays average accuracy plotted over time for each of the three conditions. Here, accuracy refers to how well messages were successfully understood. If a responder always chose the correct objects based on the producer's messages, then accuracy, as a proportion of correct trials, would be counted as 1. If a responder never chose correctly, accuracy would be counted as 0.

Accuracy data needed to be considered, because we would probably not gain much insight about subtle language behaviors of subjects if they could not use their languages successfully. Note that each condition had a relatively small number of dyads (11 all-contrastive, 10 single-contrastive, and 12 unconstrained). In order to determine whether or not one condition performed better overall, an ANCOVA analysis was performed on dyad accuracy over the 12 blocks in each of the three conditions. It was found that all-contrastive performed significantly better than single-contrastive at a .05 significance level (estimated difference between intercepts 0.0183, CI [0.0019, 0.0346]). There was no significant difference in the performance of unconstrained compared to either of the other conditions. Despite the better performance on average of the all-contrastive group, it should be noted that it was still rare for any dyad to have two perfect blocks in a row (only four dyads achieved this). Therefore, we can conclude that performance was not very high overall.

Now what about metonymy? It is not enough to say that a participant who produced a reduction by Definition 1 (parts i & ii) was actually using metonymy, because there is a lack of information about what individual symbols *mean*. A measure was needed to describe the stability of a subject's mapping between symbols and features. We were looking, especially, for a one-to-one mapping between a symbol and a particular feature (part iii of Definition 1). Here, a D-prime analysis was used to determine, for each of the nine symbols, the consistency of mapping between that symbol and each of the nine possible features. This required 81 D-prime scores ($d'$) for each subject, where each $d'$ was estimated by calculating both the "hit" rate and "false alarm" rates for its symbol-feature pair (only trials in the second half of the experiment were considered so that early errors would be discounted), and then deriving the D-prime value itself by the formula $d' = Z(hit) - Z(false\ alarm)$. $Z$ values were taken from a standardized table.

The resulting $d'$s were in the interval [-6.8, 6.8]. A high value meant that the corresponding symbol was mapped strongly to the corresponding feature; in this case, the signal (correspondence between the symbol and feature) would be high, and the noise (correspondence between the symbol and other features) would be low. But what value is high enough? We chose a threshold of 3.4, so any symbol-feature mapping with $d' \geq 3.4$ was considered to be strong enough to be considered stable.

Here, a symbol that had a consistent one-to-one mapping to a particular feature is said to be *compositional.* A symbol that was both reduced (by Definition 1, i & ii) and compositional (by Definition 1, iii) would satisfy our operational definition metonymy (note: when testing a contrived, perfectly compositional language, it was found that $d'$s for the metonymic symbol were found to be unaffected by reduction).

By this measure, metonymy was only observed twice in the human data – too few to make any statistical claims. A more substantial picture emerged from the separate compositionality and reduction information extracted from each condition. Table 1 shows both the proportion of symbols per dyad that were reduced at least once, and the proportion of symbols that were found to be compositional by the D-prime analysis.

Table 1

*Proportion of Symbols, Compositional or Reduced*

| | | Condition | | | | | |
|---|---|---|---|---|---|---|---|
| | | all-contrastive | | single-contrastive | | unconstrained | |
| Type | Experiment | Mean | Std. Dev. | Mean | Std. Dev. | Mean | Std. Dev. |
| Compositional | Human | 0.39 | 0.08 | 0.18 | 0.13 | 0.21 | 0.13 |
| | Model | 0.01 | 0.02 | 0.01 | 0.03 | 0.02 | 0.04 |
| Reduced | Human | 0.01 | 0.02 | 0.03 | 0.05 | 0.02 | 0.04 |
| | Model | 0.12 | 0.12 | 0.09 | 0.09 | 0.13 | 0.14 |

*Note.* Standard deviation is estimated sample standard deviation.

Finally, a one-way ANOVA was used to test the alternate hypothesis—that manipulation of contrastive features would influence the frequency of metonyms (symbols that were both compositional and reduced). Since proportion of metonymic symbols was measured, an arcsine transformation was applied to the data. No significant effect was found, F(2, 30) = 0.75, p = .482.

**Discussion**

The task was found to be very difficult for human subjects. This can be seen in the low overall accuracy. Also, the low proportion of compositionality indicates that the languages created by participants did not tend toward one-to-one mappings between symbols and individual features. For these reasons, interpretation of the ANOVA results is probably not useful.

Several measures might be taken to improve accuracy, thereby reducing noise in the results. For one, the number of features may have been too large for participants to keep up with. Reducing the number of dimensions (from three to two, for example) could certainly make the task easier. Also, the experiment had a fixed number of blocks. It is possible that, given more time, most of the groups would have achieved high accuracy. An accuracy-based stop criterion might have allowed each group to create their language on a timescale that was optimal for them, although this would make analysis more difficult.

The experimenter’s choice of alphabet symbols was designed to push participants towards an optimal solution. There was an equivalence class of perfectly compositional solutions—each a one-to-one mapping of symbols to features—and the 3×3, base-variant alphabet structure, shown in Figure 2, was intended to make these solutions more obvious. Since no participants arrived at a perfectly compositional mapping, we can possibly infer that the alphabet structure was not helpful in driving participants towards the optimal solutions, even though some tendency for compositionality was observed. This may have something to do with the alphabet symbols themselves, since there was some tendency for participants to map structure of the orthographic symbols onto features in creative ways (e.g. using a more curvy symbol to describe “circle”, or a symbol that partly resembles the letter b for “blue”).

As this experiment was exploratory in nature, it is not surprising that the results lacked significance. It should be noted that narrowing was very infrequent in the original tangram studies, and so a low proportion of metonymic terms was expected here. Still, the nature of these preliminary results is clearly too uncertain for the few examples in these data to be truly considered metonymy. But it is a start. These data will be referenced in Experiment II, for comparison to the neural network model results.

## Experiment II

### Neural Network Model

**Architecture.** The present model was based on the PDP approach taken in Rogers, et al. (2004), but it was modified to simulate two artificial agents. This was accomplished by having two distinct, but isomorphic *modules* corresponding to the two *agents* (here, agent refers to a simulation of a human subject). Information could be transferred between the two modules via a set of shared language units. See Figure 5 for a schematic of the model.

Each module, representing an agent, is really just a collection of *groups* (here, “group” is used instead of the more traditional term, “layer,” but it simply denotes a set of functionally related processing units). The groups within a module were designed to be as close as possible to the Rogers, et al. (2004) model. A central *semantic* group is connected to a *visual* group, a *focus* group, and three *language* groups. The focus group was included in the present study to replicate Experiment I, and was not part of Rogers’ original model. Also, the language groups are divided here in a way that differs from that study. Rather than have classes of units representing different types of verbal propositions, three language groups were created here to mirror the three possible symbol slots in the human experiment.

Each visual group had 27 units corresponding to individual features of three objects. As in Experiment I, an object could be composed of exactly three out of nine possible features. Units were ordered in each visual group, so that the first set of nine units corresponded to object one, the second set to object two, and the third set to object three. In each trial, visual groups for both agents were clamped with the same bit vector, where an individual feature was present if that unit’s value was 1 (“on”) and not present if the value was 0 (“off”).

The function of the focus group would depend on the role (producer or responder) that its agent was in, since an agent could be either. When producing, the agent’s focus group, which had three units, was clamped with a bit vector input pattern for the duration of the trial. One unit would be clamped “on”, and the others “off”. The “on” clamp functioned in the same way as the rectangular green frame in Experiment I, only here it picked out a discrete-valued position (either one, two, or three) corresponding to one of the objects in the visual field. The responding agent’s focus group was the site of training.  Its output was compared to a target at the end of a trial. The target was identical to whatever clamp was being applied to the producer in that same trial. In other words, the responder had to learn to select the abstract object to which the producer was attending.

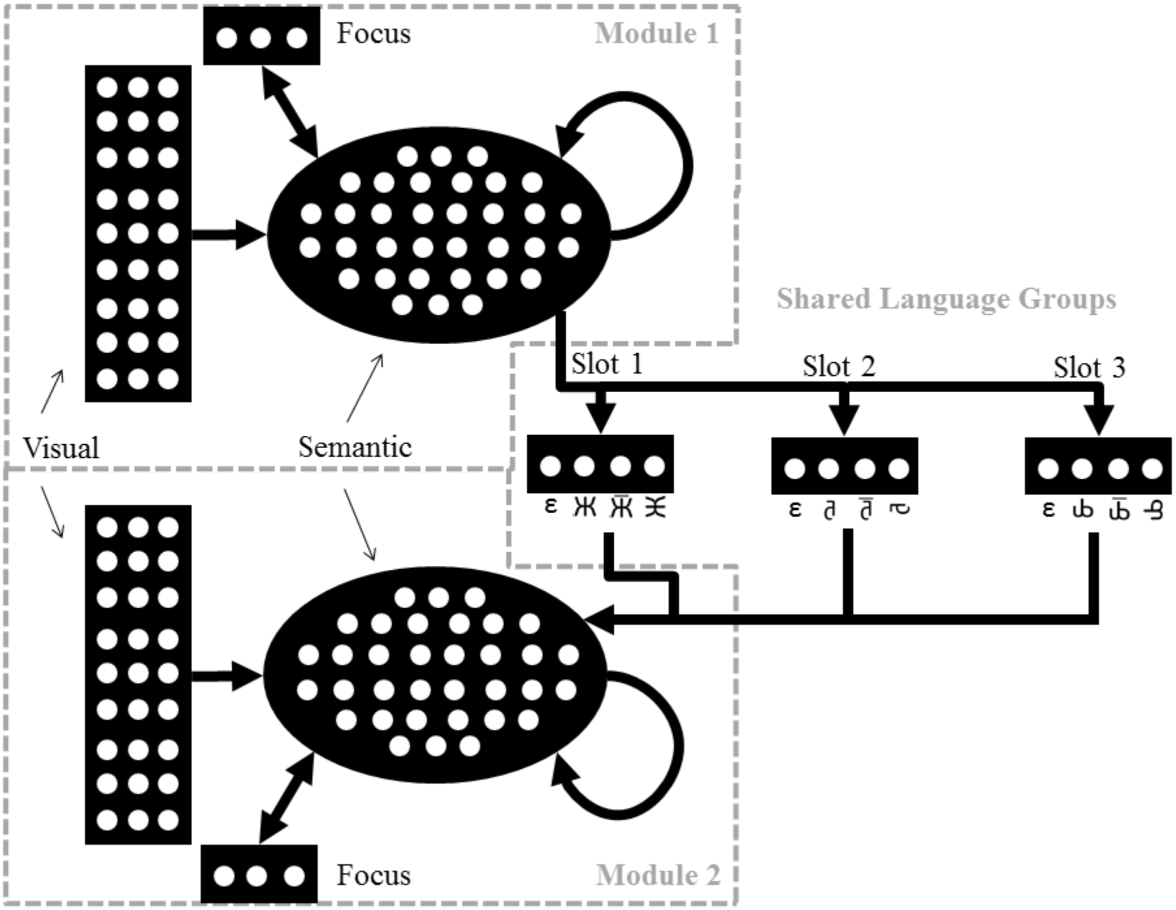

**Figure 5. Network architecture. Each agent is represented by a module, as indicated by the dotted boundaries. The modules are linked by a set of three shared language groups, corresponding to the message slots from Experiment I. One unit in each language group is reserved for the empty string, *ε*. Arrows to and from the language groups indicate that the top agent is producing a message, and the bottom agent is receiving it. Another set of arrows (not shown) with directions reversed would be used when roles were switched.**

As already mentioned, the three language groups were designed set up to mirror the three-slot symbol structure of the message interface in Experiment I. Each language group contained four units: one for each of the three possible symbol variants, and one to represent an empty string, $\varepsilon$. The $\varepsilon$ unit was required in order for reductions to be possible. The units were ordered, and the first unit in each language group was a-priori considered the $\varepsilon$ unit.

Groups were connected by *weights* that were either bidirectional or unidirectional. Here, "connected" means fully connected, such that each unit in the source group is connected to every unit in the destination group. Weights from the visual to semantic groups were unidirectional, because visual inputs were always clamped, and they did not require feedback from the semantic units. Between focus and semantic groups, there were bidirectional weights. There were two separate sets of weights between the modules; these weights, connected to the intermediate language groups, were either directed from module one to module two or vice versa depending on which module was simulating the producer role from Experiment I.

Recurrence was built into the model by creating weights from each semantic group to itself through that group's own set of cleanup units. Each semantic group was given 100 units, and each cleanup group had 20 units.

**Dynamics.** As in Rogers, et al. (2004), a fixed, untrainable bias of value $-2.0$ was applied to all groups in the model. This bias functioned to drive the unit activations downward in the absence of input.

The hyperbolic tangent function was used as the activation function for the semantic units, as it was found, in early tests, to speed convergence. A sigmoid function was not used for the language or focus groups because these groups were involved in making discrete choices. For instance, human participants in Experiment I were not allowed to make continuous adjustments to the symbols that constituted their messages, and so a continuous vector of outputs would not effectively simulate the choice of a particular symbol in one of the language groups.

To discretize the outputs of these groups, a special, two-part activation function was defined. First, this function used the *softmax* rule to compute the activations of the group's units so they all would add to one, effectively generating a probability distribution over symbol values (McCullagh & Nelder, 1989). In the next step, a simple greedy algorithm was used to select the unit with the highest activation, and this unit's activation was pushed up to a value of 1.0, while all other units were set to 0.

**Training.** The back-propagation through time (BPTT) algorithm with cross-entropy error was used here (Rumelhart, Hinton, & Williams, 1986b; same algorithm as Harm & Seidenberg, 1999). Each trial had a duration of 10 discrete time steps. The training procedure was online, and activations were reset to a default value at the beginning of each trial. Momentum was set to 0.

During initial tests, it was found that the discretized activation of the focus and language groups made learning convergence very hard for the model. A preconditioning phase was designed in which these groups used the continuous version of the softmax activation function for some fixed duration (5000 trials of precondition at a learning rate of 0.01 was found to be adequate). This preconditioning took up only a relatively small fraction of the total trials, and it facilitated convergence later on. This is not completely unreasonable, as it may actually be a way to account for non-linguistic biases that human subjects bring to the task, even without natural language semantics—preferences for certain colors or shapes, etc. After preconditioning, the learning rate was lowered to 0.001.

Visual input patterns were generated that matched the distribution of objects in Experiment I. Focus group input patterns were generated that corresponded to the original target objects in Experiment I.

In each trial, the input patterns were clamped for all 10 time steps. Note that, in Experiment I, responders did not get access to any visual information about objects or their partners' messages until after a message was sent. To replicate this, the semantic units of the responder were clamped to a value of 0, and the focus units clamped to a value of 0.33 for the first five time steps. This had the effect of "blinding" the responder while the initial processing dynamics of the producer were taking shape. The target pattern was not compared to the responder's focus group activations until the final three time steps.

### Simulation: Revisiting Experiment I

**Method.** The model was run in a similar fashion to the procedure of Experiment I. Ten simulations were run in each of the three conditions. For each simulation, a different random seed was used to simulate the effect of individual differences among subjects. After every six iterations, the roles of the two agents were reversed. Training continued until 36,000 trials had been completed. A fixed window was chosen to simulate the fixed duration of Experiment I.

**Results.** The same analyses from Experiment I were repeated here. Figure 4(b) shows accuracy by condition over the course of each simulation. Table 1 shows compositionality and reduction data.

Metonymic behavior was very rare in the model, as it was in the human data. There was only a single instance of a symbol which was both compositional and reduced. As expected, the one-way ANOVA demonstrated no effect of feature contrastiveness on metonymic frequency, F(2, 27) = 1, p = .381.

**Discussion.** The high accuracy of the all-contrastive condition is worth noting. Within less than 15,000 trials, these networks converged to a solution with 100% accuracy. What may be misleading is the low accuracy of the unconstrained condition. This is misleading because it may be tempting to infer that the networks could not learn a good solution. However, the upward trend in accuracy is an indication that the network may have performed optimally if given enough time (much like the human subjects).

One variable that was not manipulated here was the number of hidden units. The number of hidden units (100) was used simply because initial tests found it was adequate to lead to convergence. However, varying the hidden units would certainly lead to different performance, and it would need to be explored further.

## General Discussion

These results may hold some clues regarding how the problem might be explored in future experiments and model implementations. One interesting thing to note is that the all-contrastive condition seems to have the highest average accuracy for both humans and networks. This is perhaps not surprising, as a clean mapping from symbols to features might be easier in this condition. Unfortunately, conclusions about the accuracy data in the human experiment are not reliable given the noisy nature of the human dyad performance in the face of a difficult task. But it is clear in the case of the networks that consistently contrastive features made the task quite easy. Unpredictably contrastive features made the task extremely difficult at first, with accuracy below chance. Even in this case, the model can still converge slowly.

The most striking result is that compositionality and reduction profiles of the humans and networks are almost polar opposites of each other. Humans tended to develop at least some compositional structure to their languages, but with very little narrowing. Networks, on the other hand, often utilized message length reduction; but we cannot classify this behavior as metonymic, because the symbol-feature mappings did not turn out to be one-to-one. Indeed, the net $d'$ gain for networks was negative, indicating that they tended not to settle into a compositional system (for humans the net $d'$ was split almost evenly, with approximately half of the subjects positive and half negative).

This highlights a potential problem with this model. It might be difficult to understand what kind of system the network is coming up with to map symbols to features. For future studies, it should be noted that the network design here did not randomize visual positions of objects as was done in the human experiment. This detail could turn out to explain why compositionality was so low for the networks. In theory, the responder network may have been able to learn to perform well by "drowning out" the visual input, because all that the task required was for the focus units of the responder to match those of the producer. If this happened, then messages from producer to responder might be uncorrelated with features in the environment in the systematic way that was expected. This could be addressed in future research.

Another drawback of this particular model is that, despite its recurrent properties, the entire system may have a fixed point attractor. This means that once an optimal solution is found, there will be no further change of organization in the system, which may be at odds with our notion of what metonymy is. Alternative architectures with more varied dynamics may need to be explored to address this concern.

## References

Barcelona, A. (2011). Reviewing the properties and prototype structure of metonymy. In R. Benczes, A. Barcelona, F. J. Ruiz de Mendoza Ibáñez (Eds.) *Defining Metonymy in Cognitive Linguistics: Towards a Consensus View* (pp. 7-57). Amsterdam & Philadelphia: John Benjamins.

Chen, X. (2011). Metonymic matrix domains and multiple formations in indirect speech acts. In R. Benczes, A. Barcelona, F. J. Ruiz de Mendoza Ibáñez (Eds.) *Defining Metonymy in Cognitive Linguistics: Towards a Consensus View* (pp. 249-268). Amsterdam & Philadelphia: John Benjamins.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* , 22, 1-39.

Croft, W. (1993). The role of domains in the interpretation of metaphors and metonymies. *Cognitive Linguistics*, 4, 325-70.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Fillmore, C., Kay, P., & O'Connor, C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language* 64: 501–538.

Frisson, S., & Pickering, M. J. (1999). The processing of metonymy: Evidence from eye-movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 25, 1366-1383.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science, 29(5),* 737-767.

Gibbs, R. W. (2007). Experimental tests of figurative meaning construction. In G. Radden, K. Köpcke, T. Berg, & P. Siemund (Eds.) *Aspects of Meaning Construction* (pp.19-32). Amsterdam & Philadelphia: John Benjamins.

Handl, S. (2011). Salience and the conventionality of metonymies. In S. Handl, & H. Schmid (Eds.) *Windows to the Mind: Metaphor, Metonymy, and Conceptual Blending* (pp. 85-112). Berlin & New York: De Gruyter Mouton.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106,* 491-528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*, 662-720.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98*, 74-95.

Kövecses, Z. & Radden, G. (1998). Metonymy: Developing a cognitive linguistic view. *Cognitive Linguistics* 9(1): 37-77.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By.* Chicago: University of Chicago Press.

McCullagh, P., & Nelder, J. A. (1989) *Generalized Linear Models* (2nd ed., chap. 5). London: Chapman & Hall.

Panther, K., & Radden, G. (Eds.). (1999). *Metonymy in Language and Thought.* Amsterdam & Philadelphia: John Benjamins.

Plaut, D. C. and Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10,* 377-500.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., and Patterson, K. (2004). The structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, 111, 205-235.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.). (1986) *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations.* Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In *Mit Press Computational Models of Cognition and Perception Series* (pp. 318-362). Cambridge, MA: MIT Press.

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences, U.S.A.* 104, 7361-7366.