Cognitive Systems Research 12 (2011) 175–185

# How groups develop a specialized domain vocabulary: A cognitive multi-agent model

Action editor: Andrew Howes

David Reitter, Christian Lebiere

*Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*

Available online 3 August 2010

**Abstract**

We simulate the evolution of a domain vocabulary in small communities. Empirical data show that human communicators can evolve graphical languages quickly in a constrained task (Pictionary), and that communities converge towards a common language. We propose that simulations of such cultural evolution incorporate properties of human memory (cue-based retrieval, learning, decay). A cognitive model is described that encodes abstract concepts with small sets of concrete, related concepts (directing), and that also decodes such signs (matching). Learning captures conventionalized signs. Relatedness of concepts is characterized by a mixture of shared and individual knowledge, which we sample from a text corpus. Simulations show vocabulary convergence of agent communities of varied structure, but idiosyncrasy in vocabularies of each dyad of models. Convergence is weakened when agents do not alternate between encoding and decoding, predicting the necessity of bi-directional communication. Convergence is improved by explicit feedback about communicative success. We hypothesize that humans seek out subtle clues to gauge success in order to guide their vocabulary acquisition.

*Keywords:* ACT-R; Alignment; Language Evolution; Microevolution; Cognitive Architectures; Multi-Agent Simulation

## 1. Introduction

Languages evolve: like biological systems, they undergo mutation and selection as they are passed on between speakers and generations. Human communication as well as its biological analog evolve under environmental constraints. Fitness of a set of linguistic devices is, thus, also a function of the cognitive facilities. In this paper, we assume that the acquisition and retention of linguistic facts in memory is a crucial factor determining how languages are developed by communities. We use a cognitive architecture to provide an independently validated model of human memory to simulate the evolutionary process that produces a crucial part of the communication system: the vocabulary.

Recent models of dialogue describe how interlocutors develop representation systems in order to communicate; such systems can, for instance, be observed using referring expressions that identify locations in a maze. Experiments have shown that referring expressions converge on a common standard (Garrod and Doherty, 1994). Pickering and Garrod's (2004) Interactive Alignment Model suggests that explicit negotiation and separate models of the interlocutor's mental state aren't necessary, as long as each speaker is coherent and adapts to his interlocutors, as speakers are known to do on even simple, linguistic levels (lexical, syntactic). This shifts the weight of the task from a sophisticated reasoning device to the simpler learning mechanism of the individual.

Some evolutionary models see the transmission of cultural information as a directed process, in which information is passed only from the older to the younger generation. Other models explain the emergence of language as a continuous process within generations. This process may be modeled as convergence towards the bias set by innate learning and processing

Preprint submitted to *Cognitive Systems Research*.
doi:10.1016/j.cogsys.2010.06.005
*Email addresses:* reitter@cmu.edu (David Reitter), cl@cmu.edu (Christian Lebiere)

systems of the individual, but it can also be seen as the result of ongoing changes that interact with the cultural environment of the collaborating language users. There, meaning-symbol connections spread between collaborating agents and ultimately converge on a predominant one. It is the dichotomy between individual and community-based learning that motivated the experiments by Garrod, Fay, Lee, Oberlander, and Macleod (2007) and Fay, Garrod, Roberts, and Swoboda (2010), which serve as the basis for the model presented here.

In the society of cognitive agents in Fay's study and in our experiments, agents adapt their communication system collaboratively to the environmentally shaped and cognitively constrained needs of each individual. With our model, we aim to use a cognitive framework – specifically a memory model – to reflect processes in the individual that give rise to emergent convergence and learning within the community. By this, we acknowledge the fact that cultural evolution is constrained by individual learning; each agent learns according to their cognitive faculty (cf., Christiansen and Chater, 2008). The possibility of cultural language evolution between generations has been supported by computational simulations (e.g., Kirby and Hurford, 2002; Brighton, Smith, and Kirby, 2005). Kirby and Hurford's (2002) Iterated Learning model of language evolution describes vertical development of a language system by feeding developed linguistic signs or conventions back into another agent. Thus, it abstracts away from the processes within the community that forms a generation, yet does not rely only on emergence through biological evolution of the system that processes language.

The individual language faculty as a result of biological evolution and adaptation to cultural language has been the focus of psycholinguistic models proposing specialized mechanisms (the Chomskian viewpoint). While syntactic theory has long relied on production rule systems, more recent lexicalist approaches (Jackendoff, 1975) also integrate well with theories of general cognition (ACT-R: Anderson et al., 2004; SOAR: Laird and Rosenbloom 1987). In this sense, the model presented here reflects the development of a common vocabulary, which we see as prototypical for that of the lexicon, the central component of a language.

Indeed, the multi-agent model discussed in the present paper sees part of the linguistic process as an instantiation of general cognition: the composition and retrieval of signs follows general cognitive mechanisms. Adaptation according to experience is determined by human learning behavior. Simulation in validated cognitive frameworks allows us to constrain the learning process by the bounds of human memory.

Griffiths and Kalish (2007), for instance, model language evolution through iteration among rational learners in a Bayesian framework; the purpose of the present project is to tie the simulation of language evolution to a concrete experiment and a more process-oriented cognitive architecture than the Bayesian framework. *ACT-R*'s learning mechanisms add a notion of recency (decay) to the Bayesian view. Work on language processing has modeled the relationship to ACT-R memory retrieval, both for language comprehension (Budiu and Anderson, 2002; Lewis and Vasishth, 2005; Stocco and Crescentini, 2005; Ball, Heiberg, and Silber, 2007) and for language production (Reitter, 2008).

We introduce a cognitive model that simulates a participant in the experiment; multiple models interact as a community of participants. The purpose of this paper is to observe how a compositional vocabulary is created between collaborating agents in a computational cognitive simulation. Like Smith, Brighton, and Kirby (2003), we represent meaning-signal mappings using associations between memory items to create compositional signs, but we augment this model of pre-existing knowledge with one of explicitly encoded and retrievable domain knowledge. Other simulations have shown that cultural evolution leads to compositional languages (Kirby and Hurford, 2002).

We will show that the model demonstrates learning behavior similar to the empirical data. We assume these agents share a common reference system initially, display cooperative behavior and adopt mixed roles as communicators. Therefore, we explore different scenarios that test the necessity of our preconditions, in particular the fact that each agent can be both on the sending and the receiving end of the communications. The underlying question is whether dialogue (producing *and* comprehending language) is necessary for participants to establish joint communication. In search for factors that influence community convergence, we also examine the effect of initial common ground between agents and the role of the structure of the network that describes each agent's knowledge. Specifically, we present results suggesting that the specific power-law distribution found in ontologies is beneficial to the within-community convergence.

## 2. The Task

The Pictionary experiment (Garrod et al., 2007) involves two participants, a *director*, who is to draw a given meaning from a list of concepts known to both participants, and a *matcher*, who is to guess the meaning. Director and matcher do not communicate other than through the drawing shared via screens of networked computers; the matcher is able to draw as well, for instance to request clarification of a part of the picture. Each trial ends when the matcher decides to guess a concept. Garrod et al.'s set of concepts is divided into five broad categories (e.g., actor, building); the concepts within each are easily confusable (e.g., drama, soap opera). Each game involves several trials, one for each concept on the list, in randomized order. The director is not informed of the guess made by the matcher, and neither participant receives feedback about whether the guess was correct. Participants switch roles after each trial. Participants play many games so that the emergence of consistent drawings can be observed.

We implement the experiment in a form applied by Fay et al. (2010); Fay, Garrod, and Roberts (2008), where 16 concepts (plus 4 additional distractors) were used in a design with two conditions. In the *isolated pair* condition, participants were split into the same pairs throughout. They played seven rounds of six games each with the same partner. Each game consisted of 16 trials, one for each target concept (in random order). In the *community* condition, participants changed partners after

each round. Each community consisted of eight participants. The pattern of pairings was designed so that after the first round, four sub-communities existed, and after the second round, two sub-communities. After round four, the largest separation between partners was 2 (i.e., each agent has interacted via another one with every other agent); it was 1 after round seven. Fay et al. evaluated the iconicity of drawings, showing that isolated pairs developed more idiosyncratic signs, while the signs emerging within communities were more metaphoric (i.e. deducible) and easier to understand for new (fictitious) members of the language community. As idiosyncrasy increases with each drawing-recognition cycle, but resets (to some degree) when communication partners change, communities may end up evolving similar idiosyncrasy once every pair of participants played the same number of games.

The simplest measure and the one crucial for the evaluation of models like ours is *identification accuracy*. Fay et al. found that their participants generally converged quickly to a common meaning system. Convergence reached a ceiling of around 95% in both community and isolated-pair conditions. Changing interaction partners from round to round, as in the community condition, reduced accuracy during the initial rounds; however, the community reached similarly good ID accuracy as the isolated pairs after just a few rounds. We will use the development of ID accuracy as one way to evaluate the model.

## 3. The Model

### 3.1. Architecture

ACT-R (Anderson, 2007) is an architecture for specifying cognitive models, one of whose major components is memory. ACT-R's memory associates symbolic chunks of information (sets of feature-value pairs) with subsymbolic activation values. Learning occurs through the creation of chunks, which are then reinforced through repeated presentation, and forgotten through decay over time. The symbolic information stored in chunks is available for explicit reasoning, while the subsymbolic information moderates retrieval, both in speed and in retrieval probability. The assumption of rationality in ACT-R implies that retrievability is governed by the expectation to make use of a piece of information at a later point. Important to our application, retrieval is further aided by contextual cues. When other chunks are in use (e.g., *parliament*), they support the retrieval of related chunks (*building*).

The properties of memory retrieval are governed by the *activation* of a chunk $i$ that is to be retrieved. Three components of activation determine retrieval time and probability, all of which are relevant to our model: *base-level activation*, *spreading activation* and *transient noise*:

$$A_i = \log \sum_{j=1}^{n} t_j^{-d} + \sum_{j}^{cues} S_{ji} + \epsilon$$

Base-level activation (the first term of the sun) is predictive of retrieval probability independent of the concurrent context. It is determined by the frequency and recency of use of the particular chunk, with $t_j$ indicating the time elapsed since use $j$ of the chunk. $d$ indicates a base-level decay parameter, usually 0.5. Retrieval is contextualized by cues available through spreading activation (second term). It is proportional to the strengths of association ($S_{ji}$) of each cue with the target chunk. While the base-level term can be seen as a prior, spreading activation models the conditional probability of retrieval given the available cues. Finally, $\epsilon$ is noise, sampled from a logistic distribution shaped by canonical parameters, so that retrieval follows a softmax (a.k.a. Boltzmann) selection. $A_i$ must surpass a minimum *retrieval threshold* for chunk $i$ to be successfully retrieved.

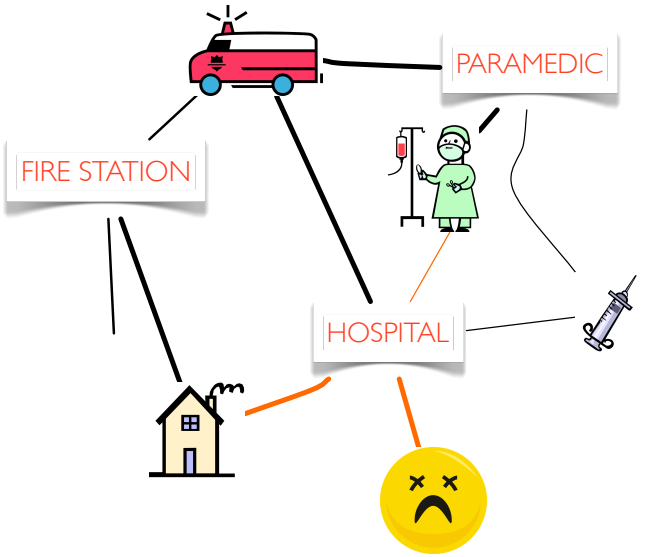



Figure 1: Example of a small ontology with abstract concepts (spelled-out words) and concrete ones (drawings).

### 3.2. Maintaining a communication system

A single ACT-R model implements the *director* and *matcher* roles. As a director, the model establishes new combinations of drawings to represent given target concepts. As a matcher, the model makes guesses. In each role, the model revises its internal mappings between drawings and target concepts. Table 1 gives an example of the process. In the simulations reported here, the directing model conveys the drawings to the recognizing model without actually producing a drawing; this recognition step is not assumed to be a source of error. The model is copied to instantiate a community of 64 agents, reflecting the subjects that took part in the Pictionary experiments.

The simplest form of representing a communication system in ACT-R memory *chunks* is as a set of signs. Each sign pairs a concept with a set of drawings. Competing signs can be used to assign multiple drawings for one concept, this would create *synonyms*; multiple concepts can also be associated with the same drawings, creating *homonyms* and ambiguity. Drawings, concepts, and signs are represented as ACT-R chunks.

To reflect semantic relationships, we need to introduce a subsymbolic notion of relatedness. We use ACT-R's spreading

| **Director** | **Matcher** |
|---|---|
| Fails to retrieve domain sign for *Paramedic*. | |
| Retrieves related concept: | |
| ⇒ component drawings syringe, doctor, emergency-vehicle | |
| Draws components syringe, doctor, emergency-vehicle | Requests related concept |
| | with cues syringe, doctor, emergency-vehicle (SDE) |
| | ⇒ concept *Hospital* |
| | Guesses Hospital |
| Learns domain sign *Paramedic-SDE* | Learns domain sign *Hospital-SDE* |
| Retrieves domain sign for target concept *Hospital* | |
| ⇒ component drawings sad, house, doctor (SHD) | |
| Verifies that *Hospital* is retrieved | |
| when drawings sad, house, doctor are activated | |
| Draws components sad, house and doctor (SHD) | Requests related concept |
| | with cues sad, house, doctor |
| | ⇒ concept *Hospital* |
| | Guesses *Hospital* |
| | Verification: |
| | Requests domain sign for *Hospital* |
| | ⇒ domain concept *Hospital-SDE* |
| | sad, house, doctor spread stronger activation |
| | to Hospital than do syringe, doctor, emergency-vehicle |
| Learns domain sign *Hospital-SHD* | thus, learns domain sign *Hospital-SHD* |

Table 1: A protocol of two model instances, conveying two target concepts. Refer to Figure 1 for the assumed, common knowledge base. In the first trial, the models fail to communicate the target concept *Paramedic* through three related drawings. In the second trial, they then successfully communicating concept *Hospital* via different drawings. The Matcher first adopts ambiguous drawings as a domain sign for *Hospital*, then revises it to reflect the better ones.

activation mechanism and weights between concepts to reflect relatedness. Spreading activation facilitates retrieval of a chunk if the current context offers cues related to the chunk. Relatedness is expressed as a value in log-odds space ($S_{ji}$ *values*).

When the model is faced with the task of drawing a given concept such as *Russell Crowe* (one of the concepts in the experiment) or *Hospital* (as in Figure 1) that has no canonical form as a drawing, the model is unable to actually draw Russell Crowe or Hospital directly. Then, a related but drawable concept (*drawing*) is retrieved from declarative memory (such as Syringe in the example). Similarly, two more concepts are retrieved, reflecting the desire of the communicator to come up with a distinctive rather than just fitting depiction of the target concept. The case of a model recognizing a novel combination of drawings is similar; the model retrieves the concept using the drawings as cues that spread activation, making the target concept the one that is most related to the drawings.

After directing or recognizing, the target or guessed concept, along with the component drawings, is stored symbolically in memory as a chunk for later reuse (*domain sign*). These signs differ from the pre-existing concepts in the network, although they also allow for the retrieval of suitable drawings given a concept, and for a concept given some drawings. When drawing or recognizing at a later stage, the memorized domain signs are preferred as a strategy over the retrieval of related concepts. The system of domain signs encodes what is agreed upon as a language system between two communicators; they will be reused readily during drawing when interacting with a new partner, but they will be of only limited use when attempting to recognize a drawing combination that adheres to somebody else's independently developed communication system.

### 3.3. Knowledge

Agents start out with shared world knowledge. Such knowledge comprises target items from the experiment, and concrete concepts representing the drawings. The important element of knowledge in the model is the relationship between target items and concrete concepts (drawings). The model abstracts away from the process of manually producing a drawing, and from the linkage between the drawing and the concept.

Knowledge is expressed as a network of concepts, connected by weighted links ($S_{ji}$). The distribution of link strengths is important in this context, as it determines how easily we can find drawing combinations that reliably express target concepts. Thus, the $S_{ji}$ were sampled randomly from an empirical distribution: log-odds derived from the frequencies of collocations found in text corpus data. In a corpus comprising several years worth of articles that appeared in the *Wall Street Journal* (WSJ), we extracted and counted pairs of nouns that co-occurred in the same sentence (e.g., "market", "plunge"). As expected, the frequencies of such collocations are distributed according to a power law. We found that the empirical log-odds resulting from these that form $S_{ji} = \log(P(j|i)/P(j|noti))$ (Anderson, 1993) ($j$ and $i$ being the events that $j$ and $i$ appear, respectively) can be approximated by a Generalized Inverse Gaussian-Poisson distribution (given in Baayen, 2001). Concepts can represent target concepts (abstract) as well as drawings (concrete). We randomly sample all link weights from the corpus; as a simplifying assumption, the concepts themselves receive equal base-level activation.

Such knowledge is, however, not fully shared between agents. Each agent has their own knowledge network resulting from life experience. This difference is essential to the difficulty of the task: if all agents came to the same conclusions about the strongest representation of target concepts, there would be little need to establish the domain language. We control the noise applied to the link strengths between concepts $j$ and $i$ for agent $A$ ($S^A_{ji}$) by combining the common ground $S_{ji}$ (shared between all agents) with a random sample from the empirical WSJ distribution $N^A_{ji}$ in a mixture model: $S^A_{ji} = (1-n)S_{ji} + nN^A_{ji}$. Then, $n$ [0–1] sets the proportion of noise. For Experiments 1 and 2, the noise coefficient is set to 0.2; the effect of noise is explored in Experiment 3.

### 3.4. Adaptation pressure

Notably, participants in the experiment converged to a common sign system fairly quickly. This happened even though there was no evident, strong pressure to do so. Agents received no explicit feedback about the quality of their guesses or drawings. The only weak clue to the success of a set of drawings was whether the partner made a guess quickly.

Invariably, the model will mistake a set of drawings for a reference to the wrong target. Lacking a feedback loop in this experiment, the model has no choice but to acquire even flawed domain signs and boost their activation upon repetition. Under these conditions, is there enough pressure to converge? It is difficult to see how interacting partners could ever agree on a working communication system, given that there is no benefit for a model in choosing the concept-drawing associations of its interaction partner. Still, ACT-R's declarative memory mechanism values frequency *and* recency. As a consequence, new concept-drawing mappings may override old ones, but are most successful if they are compatible with the majority of previously chosen mappings for that concept. In other words, agents may recognize drawings and assign the right concept, but the concept-drawing combination is not as likely to override more established, existing domain concepts that have seen frequent use. Thus, individual interactions may be successful task-wise, but they need not revise the established language system.

However, the model also leverages *consistency* as proposed in Grice's maxims of manner, "Avoid ambiguity" and "Avoid obscurity of expression" (Grice, 1975). In our context, these maxims postulate, e.g., that directors choose to draw combinations that are unique so that they won't be confused. To implement the maxims, the model assumes that a given set of drawings is associated with only one target concept, and, conversely, that a given target concept is associated with only one set of three drawings. Suppose, for example (Table 1), that the model associates concept $B$ with drawings $1, 2, 3$ (short: B-123). Later on, it comes across drawings $3, 4, 5$ as another good way to express $B$. In fact $3, 4, 5$ serve as convincingly stronger cues to retrieve $B$ than do $1, 2, 3$. Thus, the model not only correctly recognizes $B$, but also learns the new preferred combination B-345. In the following rounds, B-345 will likely shadow the alternative in a winner-take-all paradigm, since B-345 is newer than B-123 and, thus, has stronger activation due to activation decay (noise and reinforcement may keep B-123 as a winner for longer). The decay mechanism counteracts the creation of synonyms.

In evolving the domain language, the model will avoid creating homonyms as well. Suppose a concept $C$ is to be drawn, and 345 are retrieved as closely related and highly active drawings. Here, the model attempts to verify that 345 cannot be understood as any other concept than $C$. As the most strongly active concept for 345 is $B$, these drawings are ruled out to express $C$. With this mechanism, the model is able to cheaply modify the system of signs without extensive reasoning about the optimal combination every time a concept is added. Oliphant and Batali (1997) call this mechanism *Learning by Obverting*: Such a learning would "send for each meaning the signal that is most likely to be interpreted as that meaning."

### 3.5. Model

*Directing.* The model is given a target concept $A$ to convey. It uses *domain signs* and general knowledge to decide about three drawings with which to convey the concept. Domain knowledge is explicitly accessible and overrides subsymbolically derived compositions. After that decision has been made, the composed concept is committed to declarative memory as a new or reinforced domain sign. As a consequence, the model acts with consistency: once a combination has first been used to convey a concept, the model will be more likely to use it. The director proceeds with the following algorithm.

1. Retrieve a domain sign for $A$ of the form $A - \alpha\beta\gamma$.
   - *Verification:* If successful, retrieve a domain sign $B$ for the same three drawings $\alpha\beta\gamma$ ($B - \alpha\beta\gamma$). Only if $A = B$, accept the domain sign $A - \alpha\beta\gamma$ and continue with step 3; otherwise back to 1 (for a different domain sign).
2. If no acceptable domain sign is found, use subsymbolic knowledge to combine concepts to express related target meanings. Using the target meaning as cue, retrieve three drawings $\alpha\beta\gamma$. The most active drawings are retrieved preferentially.
3. Draw $\alpha\beta\gamma$.
4. Learn $A - \alpha\beta\gamma$ (note *use* of chunk, see Section 3.1).

*Matching.* Recognizing a drawing takes place in a similar fashion: domain knowledge is preferred over associative guesses. The model is given three drawings $\alpha\beta\gamma$. It proceeds with the following algorithm.

1. Attempt to retrieve a domain sign for $\alpha\beta\gamma$, resulting in $C - \alpha\beta\gamma$.
   - *Verification:* If successful, retrieve a domain sign of the form $C - \delta\epsilon\zeta$. Only if $\alpha, \beta, \gamma = \delta, \epsilon, \zeta$, accept the domain sign $C - \alpha\beta\gamma$ and continue with step 3.
2. If no acceptable domain sign is found, retrieve a concept $C$ using cues $\alpha\beta\gamma$ (spreading activation).
3. Enter $C$ as the model's guess.
4. Learn $C - \alpha\beta\gamma$ (note *use* of chunk, see Section 3.1).

### 3.6. Underspecification and Accountable Modeling

The modeling method used to design the language convergence model follows what we call *accountable modeling*: we abstract away from portions of the model that are difficult to evaluate, and focus on the aspects that yield predictions and explanations of the language convergence data. Working within the ACT-R theory, we formulate the model with a new toolbox implementation called *ACT-UP* (Reitter and Lebiere, 2010). ACT-UP reflects ACT-R, but makes individual cognitive functions rather than architectural modules directly available for combination by the modeler. The ACT-UP toolbox approach allows modelers to underspecify models by implementing a computational algorithm to cover elements that would neither introduce nondeterminism nor carry explanatory weight in this particular model.

The component of the ACT-R theory that is essential to this model are cue-based declarative memory retrieval, where cues spread activation to target chunks. We do not rely on specific procedural rules or other forms of parallelism or reinforcement learning.

### 3.7. Subsymbolic Parameters

In the following, we briefly discuss chosen values for ACT-R's subsymbolic parameters. Learning in declarative memory is governed by a decay (bll), kept at its default of 0.5. The chunk activation constant (BLC) is 2.4, and the retrieval threshold (RT, an activation below which a chunk is not retrieved) is −1. Transient noise is within the range used by other models at 0.2.

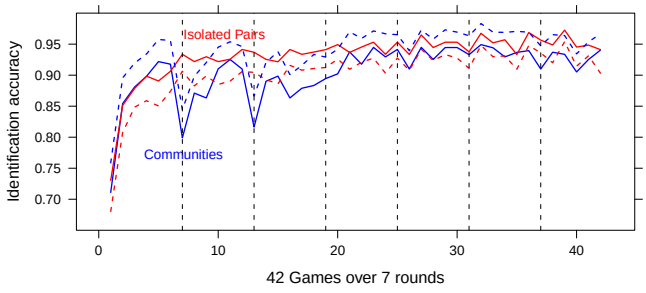
## 4. Simulation 1: Learning and Convergence

In the first simulation, we evaluate whether the model exhibits similar learning and convergence behavior, and whether there are differences in learning between the isolated-pair and community condition, as observed in Fay et al.'s experiment. The model uses the same number of concepts, trials and simulated participants as in the experiment. 100 repetitions of the simulation were run, each with a different, randomly sampled ontology structure; the same 100 ontologies were used for all simulations.

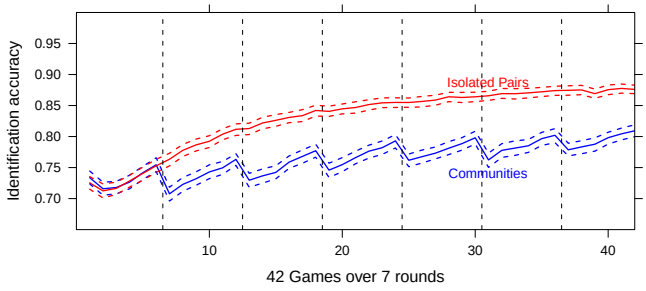
### 4.1. Results

As shown in Figure 2(b), the learning behavior differs in the two conditions. *Isolated pairs* and *Community pairs* show a learning effect. Their respective communication systems converge. However, unlike isolated pairs, community pairs display lower ID accuracy after the 7th game (game 1 of round 2), i.e., after switching partners. These effects are also present in the empirical data. The overall ceiling and the gain in accuracy, however, are lower in the simulation than empirically (see Figure 3a), and convergence in the model appears to be relatively shallow.

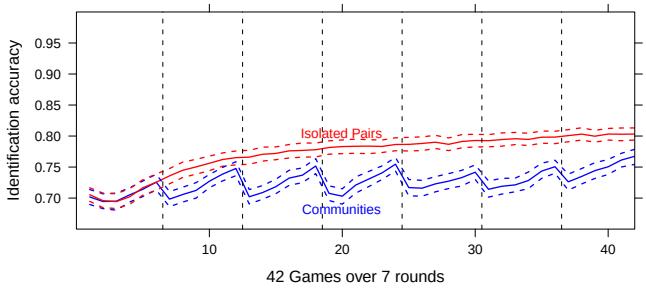
A linear mixed-effects model was fitted to the simulation's predictions to test for some of the key empirical effects. The linear regression model treated round, game and condition (isolated pairs vs. communities) as independent variables and



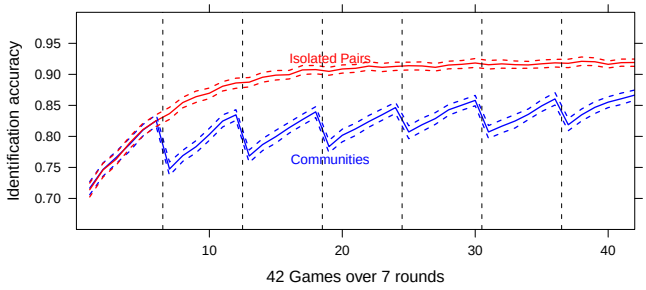

(a) Empirical




(b) Model




(c) Model: No role-switching




(d) Model: explicit feedback

Figure 2: Identification accuracy for isolated pairs and communities: human data (Fay, p.c.) and model runs. 95% confidence intervals (one-tailed for Fig. a). Partner switching (in community condition only) is indicated by dashed lines.

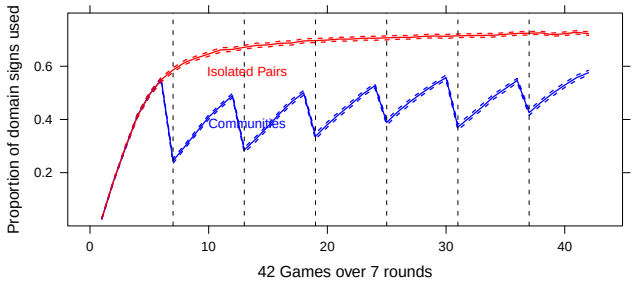



Figure 3: Proportion of domain signs used in concept recognition of the 7 rounds of the game (regardless of whether the sign was guessed correctly).

predicted log-transformed ID accuracy. A random intercept grouped by repetition was also fitted. The regression showed expected effects for round ($\beta_R = 0.015$, values: 1–7) and game ($\beta_G = 0.008$, values: 1–6), indicating improving accuracy with each game and round. An interaction of round and game ($\beta_{G:R} = -0.001$) showed that the convergence leveled off in later rounds (as expected). There was a main effect of condition ($\beta_{IP} = 0.081$), and an interaction of condition (isolated pairs) and round in the predicted direction ($\beta_{IP:R} = 0.014$), showing faster convergence within isolated pairs than within communities. (Intercept $\beta_0 = -0.273$. All $p < 0.0005$. All $\beta$ in log space. Independent variables were centered. $p$-values obtained via Markov-Chain Monte-Carlo (MCMC) sampling; fitted parameters appeared normally distributed.) The fit of the model (means by round, game and condition) with the empirical data yielded a Root Mean Square Error (RMSE) of 0.13. Correlation is 0.68.

To understand the relation between the development in identification accuracy and the learning mechanism of the model, we contrasted identification via domain signs and via cue-based retrieval (Section 3.2). Figure 3 shows the strong acquisition of domain signs in the first round; domain signs are maintained in isolated pairs, while switching partners causes domain sign use to deteriorate. Notably, the second dyad fails to establish the same level of domain sign use.

### 4.2. Discussion

The results demonstrate, first, that agents converge both when retaining partners and when interacting with changing partners. Second, the results show that partner switching results in a setback in performance. This means that different dyads indeed converge on different signs for the same concepts. Community agents recover from the loss in performance (as human subjects do) and continue to optimize their communication systems. Notably, the setback appears to be smaller for rounds 3 through 7, i.e., through repeated partner switching, agents arrive at a vocabulary that is less susceptible to disruption. While Figure 2(b) suggests an effect of condition on the ceiling that is achieved, more simulations would have to be run to derive a prediction (empirical data is not available beyond the 42 games). Overall, the model's results are qualitatively similar in many ways to the human subjects. Convergence in the community condition can be observed by a rise in performance during game 1 in each round, that is, the game played between agents that have not interacted before. In both empirical and simulated data, this rise is present. However, the magnitude of this effect is not reflected by the model.

Domain signs are established by the dyads. However, switching partners prevents the model from permanently fixing the language system that it developed initially. While communities eventually achieve similar levels of identification accuracy, that is not primarily owed to a recovery in domain sign use, but, we assume, also to strengthened base-level activations of the target concepts.

With the strong difference in domain sign use, the model displays a trait that could explain Fay et al.'s (2008)'s finding: in their study, human raters found community-evolved signs to be less iconic than the signs established just between dyads, which were highly optimized, abstract and less penetrable to newcomers.

## 5. Simulation 2: Director and Matcher roles

Garrod et al. (2007) compared the performance of their participants in a comparable Pictionary task when a single director remained in that role throughout the experiment (single director, SD condition), vs. when participants swapped roles after each round (double director, DD condition). Identification accuracy was slightly higher for the role-switching, double-director condition than in the single-director condition (significantly so only in the final rounds 5 and 6). This experiment is comparable to the *isolated pairs* condition in our model. Our model can not only simulate the role-switching conditions, but also predict contrasts between isolated pairs and communities. The general question here is whether unidirectional communication would be sufficient to develop a community language. So, in this experiment, agents did not switch roles after every concept conveyed, i.e. they remained either director or matcher throughout the game. Otherwise, this simulation mirrored Simulation 1.

### 5.1. Results

Identification accuracy for isolated pairs converged to a lower level than in Simulation 1. Communities also failed to achieve the same level of accuracy when director and matcher roles were not swapped (Figure 2(c)).

A linear mixed-effects model was fitted to the data obtained from Simulations 1 and 2, with condition (Isolated Pairs vs. Communities) and role swapping as independent variables and predicted log-transformed ID accuracy; a random effect grouped repeated measures as in Simulation 1. Role swapping led to reliably higher ID accuracy ($\beta_R = 0.051$). Isolated pairs showed reliably higher ID accuracy than Communities ($\beta_{IP} = 0.062, p < 0.0005$). A reliable, positive interaction was found between role swapping and the isolated pairs condition ($\beta_{IP:R} = 0.020$). (Intercept $\beta_0 = -0.324$. All $p < 0.0005$. All $\beta$ in log space; estimates appear normal. $p$-values via MCMC sampling.) [1]

[1] This model is constructed to fit a hypothetical situation. For comparison

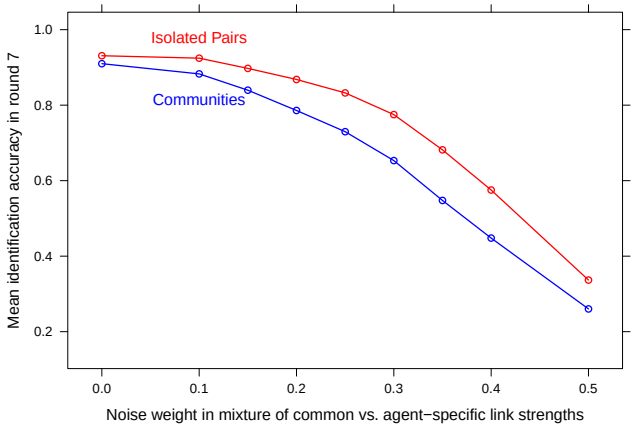



Figure 4: Mean identification accuracy at round 7 is reduced with noise between the knowledge bases of each agent. Bootstrapped 95% confidence intervals.

### 5.2. Discussion

This experiment showed that turn-taking aids in the development of a common community vocabulary in the case of the no-feedback, single-guess naming game. For the isolated pairs condition, this parallels Garrod et al.'s findings. Without turn-taking, the gap between communities and isolated pairs widens when uni-directional communication is used. It should be noted that, unlike Fay et al.'s experiments and our simulation, Garrod et al.'s study involved feedback about the guesses, potentially allowing better convergence overall. Simulation 5 will investigate the effect of explicit feedback on accuracy.

## 6. Simulation 3: Noise in Common Ground

A vital assumption of the compositional semantics in this model is that the agents start out with some common knowledge. For instance, both director and matcher need to accept that ambulances and buildings are strongly related to the concept *hospital*. However, the strength of each link between the same two concepts may differ between any two agents. This error does not necessarily preclude the matcher from making the right inference. The model allows us to test the role of inter-subject variation, and to predict the results of a lower overlap between the knowledge bases of each agent. This simulation was run repeatedly, varying only the level of variation in link strengths between the agent's ontologies. We measure the mean ID accuracy during the last, seventh round. Otherwise, this Simulation in the same as Simulation 1.

### 6.1. Results and Discussion

Figure 4 shows that mean identification accuracy (7th round, all games) decreases with increased levels of noise in the subsymbolic knowledge state between agents. The model appears to deal reasonably well with discrepancies between agent knowledge at levels of up to 0.3 (coefficient in the noise mixture) for both isolated pairs and communities configurations; at higher noise levels, performance drops more quickly. The same generally holds true when taking all rounds into account. (At high noise levels, the initial acquisition of domain signs still works, but agents fail to converge further beyond the initial game or beyond a lower ceiling.) Further work should reveal whether further learning cycles can make up for the effect, i.e., medium noise levels lead to slower convergence and the failure to converge here is due to the limited number of games.

## 7. Simulation 4: Ontology Structure

Does the structure of relationships between ontological concepts assist human communities in language convergence? In this task, human participants as well as cognitive models established new meanings of the concepts by combination. Accuracy of retrieval of target concepts given the combination of drawings depends on the ambiguity of the drawing-concept relations; in other words, it depends on how clearly the drawings identify the right target concept.

Commonly, the frequency of co-occurrences of concepts in text collections follow power-laws (we take co-occurrence to be an indicator of relatedness). It would not be surprising if the mechanisms of language convergence had evolved to benefit from the network topology in the ontology. During the initial development of the model, we noticed informally that the knowledge structure described in Section 3.3 was necessary to establish positive, reliable and strong convergence among the partipants; attempts with different distributions did not yield the desired effects.

In this experiment, we investigate the model's predictions w.r.t. ontology structure.

### 7.1. Method

We ran the same simulation as in Simulation 1, but replaced the ontology produced by sampling from the collocation frequencies of the Wall Street journal with a random one. Connection strengths ($S_{ji}$) were randomly sampled from a uniform distribution, whereas the parameters were set so that the mean connection strength equaled the mean of the log-odds of collocations from Simulation 1. All other parameters were the same. Generating this random weighted graph, we produced the hypothetical case where drawings and concepts in the knowledge base had different, but equiprobable connections strengths. Thus, target concepts could still be positively identified using their related drawings.

### 7.2. Results

The uniformly distributed ontology led to a severe drop in initial and also final accuracy. Figure 5 shows that final identification accuracy is near 0.20; compared to more than 0.85 when concept link weights are sampled from the Wall Street journal collocations (Figure 2(b)). The learning effect

with Simulation 1, we note that the fit of the model (means by round, game and condition) with the empirical data yielded an RMSE of 0.17. Correlation is 0.64.

in Simulation 1 led to a 30% reduction in error within the 7 rounds of convergence, while the present manipulation reduced the error by about 12%.

### 7.3. Discussion

The results provide preliminary support of an intriguing hypothesis: that the commonly found ontological structure provides an evolutionary environment that is particularly suited to the creation of sign systems through retrieval by association. Uniformly distributed link weights render the domain sign invention more difficult, as the semantic space becomes more ambiguous and concepts within it become less informative. It is possible that compositional learning strategies in our model and the cognitive framework represent mechanisms that have adapted to benefit from the typical distribution of relationships between ontological concepts.

There are ways of calibrating the alternative distribution for comparison to the power-law distribution obtained from the corpus data. Similarly, very different distributions can be considered; however, we still obtained similar drops in performance when sampling the weights from a uniform distribution constructed so that the mean odds before the log-transform were the same as the odds collocations in Simulation 1.

## 8. Simulation 5: Feedback

The game used in the experiments and simulations discussed here differs from the typical linguistic interaction between humans in one important aspect. Humans usually obtain some form of measure of success that determines whether the communication was received correctly. In the game, this measure of success is never explicit, and, at best, implied. For example, speakers may retroactively decide that a guess (of concept A) that they made earlier did not work, because a new sign much more clearly indicates concept A. (Each of the 20 concepts is only shown once within each game.) Also, the time it takes for a guesser to determine the meaning of a sign may be indicative of its qualities to the director (a sign that is recognized faster may be considered more reliable). It is likely that our model fails to capture these subtle signals that may be used by humans.

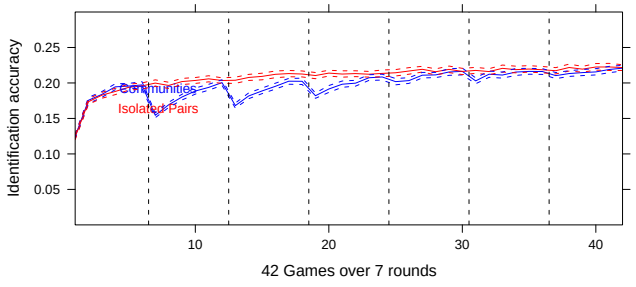



Figure 5: Convergence (model) when ontology link weights are sampled from a uniform distribution with the same mean as in the samples of collocations from the Wall Street Journal. Otherwise as in Figure 2(b).

To find out whether feedback makes an appreciable difference, we modified the simulations so that both models (director and matcher) receive feedback after each interaction. Director and matcher models learn or boost a sign in declarative memory only if the feedback was positive, i.e., the matcher guessed the sign correctly. Otherwise, this simulation parallels the model and parametrization used in Simulation 1.

### 8.1. Results

Our results (Figure 2(d)) suggest stronger increase in ID accuracy for both isolated pairs and communities than was observed in the interactions without feedback. The first game of each round in the community condition is indicative of convergence across the social network; the ID accuracy of these games improves steadily.

A linear mixed-effects model was fitted to the data obtained from Simulations 1 and 5, with condition (Isolated Pairs vs. Communities) and feedback as independent variables, log-transformed ID accuracy as response, and a random effect grouped repeated measures as in Simulation 1. Feedback led to reliably higher ID accuracy ($\beta_R = 0.066$). Isolated pairs showed reliably higher ID accuracy than Communities ($\beta_{IP} = 0.08, p < 0.0005$). No reliable interaction was found between feedback and the isolated pairs condition ($p > 0.4$). (Intercept $\beta_0 = -0.273$. All other $p < 0.0005$. All $\beta$ in log space; estimates appear normal. $p$-values via MCMC sampling.)

The fit of the model (means by round, game and condition) with the empirical data yielded an RMSE of 0.08. Correlation is 0.79.

### 8.2. Discussion

With feedback, developing a common language becomes much easier. We see convergence that surpasses the ceiling of Simulations 1 and 2, and that is closer to human performance. This is compatible with a view that human subjects make use of subtle cues to the success of proposed domain signs.

## 9. General Discussion

The model replicates several of the characteristics of the *communities* compared to the *isolated pairs* condition; specifically the setbacks after switching partners for the first few times and the ultimate convergence, despite very limited feedback. We also arrive at a clear prediction: bi-directionality is essential for linguistic convergence in communities. The model fails to explain several other characteristics of the data. While subjects gained most of their ID accuracy during round 1, the model shows a more gradual convergence towards functional vocabularies common to the interacting agents and reaches a lower performance level than do the human subjects. In informal experiments, we found that the overall gain from the initial to last game was not merely a matter of subsymbolic parameters; this motivated Simulation 5, which describes a hypothetical experiment where players in both conditions receive explicit feedback about the success of the communication.

At this point, we do not emphasize parameter optimization in order to achieve a better fit to the empirical data. We believe that adaptation rates and the convergence ceiling depend both on the difficulty of the task, the specific materials (concepts) and the higher-level reasoning tools employed to optimize the language system. The task in Fay et al.'s experiment structured the list of concepts into a tree (e.g., there were four actors), making the job of drawing and guessing easier. Rather than just drawing what seems most closely related to the target concept, the experimental design invites them to choose a component concept that best disambiguates the drawing in the light of competing concepts (a head and a movie screen may be descriptive of Robert De Niro, but they do not distinguish him from Brad Pitt). Neither specific differentiation nor the precise choice of materials are modeled. Thus, we may overestimate the difficulty of the task. As a further simplifying assumption, our model always produced three component drawings before a guess is made. Garrod et al.'s (2007) design had participants give one another feedback about whether a drawing was thought to be recognized. When a matcher recognized a drawing soon, then this could have been used by the director as a sign of the drawing's utility. Such cues are not leveraged at all by our model.

The model proposes a relatively mechanistic account of domain sign construction. We assume that subjects, when not paying attention to their communicative strategy or when under time pressure, construct signs by associative retrieval. They do not need to reason explicitly about the ambiguity of their signs, or the efficiency of communication. Our experiments do not exclude the possibility of implicit communication strategies that lead to more efficient communication systems; however, the simulations point out that these mechanisms are not strictly necessary in order to achieve an improvement in one-on-one communication, even between changing partners.

However, the speed and overall ceiling of the performance are clearly more limited than what would be necessary in order to convincingly explain the emergence of a common vocabulary in a community. Oliphant and Batali (1997) argue, using a statistical model of learning, that metacognition in the form of the verification steps we use here greatly helps convergence. Our findings seem to suggest that such a model, using ACT-R's learning mechanisms, can achieve at least some level of accuracy, even with limited means to evaluate the success of a single communication. At the same time, the simulations show that an adaptation-only account might not suffice to explain the strong, feedback-free community convergence.

As mentioned earlier, one possibility is that participants might make more sophisticated use of subtle feedback signals than our model does. Those signals can be both internal and external. External signals could include the time or number of drawings taken by the matcher to make a guess, or any requests for clarification from the matcher to the director. Our current model does not reflect any of these aspects of the experiment. Feedback signals can also be internal. In that regard, a fundamental metacognitive signal lies in our ability to recognize when a new sign combination is clearly better than the existing convention. This kind of insight is essential in choosing between competing signs in the absence of a clear external signal indicating which of the competing concepts is correct. Currently the ambiguity resolution ("verification") part of the model has no basis for making a choice one way or another and has to reject them both. This leads to significant difficulty in ensuring convergence in the presence of conflicting signs, as often happens due to variation in knowledge bases, but especially when switching partners in the community condition, which will almost always result in a clash of established domain signs. If such a metacognitive tie-breaking signal were available, a number of potential resolution strategies could be followed. One would be to mark the downgraded domain sign as obsolete. Thus, despite its high activation, the sign might still be retrieved but wouldn't be used any more. If one included an episodic record of the decision, it could link the obsolete sign to the new, better combination and thus use the strength of the conflict to resolve it. More sophisticated search strategies are also possible. For instance, when the matcher recognizes that a sign combination is more specific to a concept already identified, it could not only create a new domain sign for the better combination, but modify the outdated one to guess another concept. This search process could be very effective if one factored in the hierarchical distribution of concepts into a few small groups of a few items (4 groups of 4 concepts) assuming that group identification is unproblematic. This would result in adaptive clustering dynamics in which the concepts gradually claim areas of the combined multi-dimensional sign space for which they are most specific. This in turn suggests some limitations of this experiment in extending to real-world language learning. Often the space of possible vocabularies is not so neatly and tractably circumscribed that a global search strategy could be effective. It is for this reason that the model examined here has limited itself to what is plausible for natural language vocabulary convergence.

## 10. Conclusion

We have demonstrated the use of validated, cognitively plausible constraints to explain an emergent, evolutionary group process via multi-agent simulation. Subsymbolic and symbolic learning within a validated human memory framework can account for rapid adaptation of communication between dyads and for the slower acquisition of a domain language in small speaker communities despite very limited feedback about the success of each interaction. Bi-directional communication is predicted to be necessary for a common language system to emerge from communities. The effects are robust against some divergence in prior common ground between agents. However, the model does not yet account for the magnitude of the convergence effect, which suggests that humans make use of many more cues than the model in order to evaluate and promote communicative devices.

Our model of the horizontal emergence of a common language in multi-agent communities is a first step to a computational cognitive analysis of the learning processes involved in creating combined signs and acquiring links between them

and arbitrary concepts, in other words, the evolution of language. We have included only limited explicit metacognitive reasoning over the full set of domain signs in the model. The model is intended to reflect vocabulary acquisition for larger communication systems than the limited set of domain signs that was used in the design of the Pictionary study. For large problems such as natural languages, the model predicts slower convergence. Further work will strive to demonstrate robust convergence with realistic natural language examples, which will go well beyond the empirical data that served as basis for this study.

*Acknowledgments.* We thank Nicolas Fay and Simon Garrod for making their data and manuscripts available and Ion Juvina for comments. This work was funded by the Air Force Office of Scientific Research (FA 95500810356).

## References

Anderson, J. R., 1993. Rules of the Mind. Erlbaum, Hillsdale, NJ.

Anderson, J. R., 2007. How can the human mind occur in the physical universe? Oxford University Press, Oxford, UK.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., Quin, Y., 2004. An integrated theory of mind. Psychological Review 111, 1036–1060.

Baayen, R. H., 2001. Word Frequency Distributions. Kluwer, Dordrecht.

Ball, J., Heiberg, A., Silber, R., 2007. Toward a large-scale model of language comprehension in ACT-R 6. In: Lewis, R. L., Polk, T. A., Laird, J. E. (Eds.), Proceedings of the 8th International Conference on Cognitive Modeling. Ann Arbor, MI, pp. 163–168.

Brighton, H., Smith, K., Kirby, S., 2005. Language as an evolutionary system. Physics of Life Reviews 2 (3), 177–226.

Budiu, R., Anderson, J. R., 2002. Comprehending anaphoric metaphors. Memory & Cognition 30, 158–165.

Christiansen, M. H., Chater, N., 2008. Language as shaped by the brain. Behavioral and Brain Sciences 31 (5), 489–509.

Fay, N., Garrod, S., Roberts, L., 2008. The fitness and functionality of culturally evolved communication systems. Philosophical Transactions of the Royal Society B: Biological Sciences 363 (1509), 3553–3561.

Fay, N., Garrod, S., Roberts, L., Swoboda, N., 2010. The interactive evolution of human communication systems. Cognitive Science 34 (3), 351–386.

Garrod, S., Doherty, G. M., 1994. Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. Cognition 53, 181–215.

Garrod, S., Fay, N., Lee, J., Oberlander, J., Macleod, T., 2007. Foundations of representation: Where might graphical symbol systems come from? Cognitive Science 31 (6), 961–987.

Grice, H. P., 1975. Logic and conversation. In: Cole, P., Morgan, J. L. (Eds.), Speech Acts. Vol. 3. Academic Press, New York, pp. 41–58.

Griffiths, T. L., Kalish, M. L., 2007. Language evolution by iterated learning with Bayesian agents. Cognitive Science 31 (3), 441–480.

Jackendoff, R., 1975. Morphological and semantic regularities in the lexicon. Language 51 (3), 639–671.

Kirby, S., Hurford, J., 2002. The emergence of linguistic structure: An overview of the iterated learning model. In: Cangelosi, A., Parisi, D. (Eds.), Simulating the Evolution of Language. Springer Verlag, London, Ch. 6, pp. 121–148.

Laird, J. E., Rosenbloom, P. S., 1987. Soar: An architecture for general intelligence. Artificial Intelligence 33 (1), 1–64.

Lewis, R. L., Vasishth, S., May 2005. An activation-based model of sentence processing as skilled memory retrieval. Cognitive Science 29, 1–45.

Oliphant, M., Batali, J., 1997. Learning and the emergence of coordinated communication. The Newsletter of the Center for Research in Language 11 (1).

Pickering, M. J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences 27, 169–225.

Reitter, D., 2008. Context effects in language production: Models of syntactic priming in dialogue corpora. Ph.D. thesis, University of Edinburgh.

Reitter, D., Lebiere, C., 2010. Accountable modeling in ACT-UP, a scalable, rapid-prototyping ACT-R implementation. In: Proceedings of the 10th International Conference on Cognitive Modeling. Philadelphia, PA, pp. 199–204.

Smith, K., Brighton, H., Kirby, S., 2003. Complex systems in language evolution: the cultural emergence of compositional structure. Advances in Complex Systems 6 (4), 537–558.

Stocco, A., Crescentini, C., 2005. Syntactic comprehension in agrammatism: A computational model. Brain & Language 95 (1), 127–128.