A Recurrent Network Approach to Modeling Linguistic Interaction

Rick Dale (rdale@ucmerced.edu)
Cognitive and Information Sciences, University of California, Merced

Riccardo Fusaroli (fusaroli@dac.au.dk), Kristian Tylén (kristian@dac.au.dk)
Interacting Minds Centre & Center for Semiotics, Aarhus University

Joanna Rączaszek-Leonardi (joanna.leonardi@gmail.com)
Institute of Psychology, Polish Academy of Sciences, Warsaw, Poland

Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University
Interacting Minds Centre, Aarhus University

Abstract

What capacities enable linguistic interaction? While several proposals have been advanced, little progress has been made in comparing and articulating them within an integrative framework. In this paper, we take initial steps towards a connectionist framework designed to systematically compare different cognitive models of social interactions. The framework we propose couples two simple-recurrent network systems (Chang, 2002) to explore the computational underpinnings of interaction, and apply this modeling framework to predict the semantic structure derived from transcripts of an experimental joint decision task (Bahrami et al., 2010; Fusaroli et al., 2012). In an exploratory application of this framework, we find (i) that the coupled network approach is capable of learning from noisy naturalistic input but (ii) that integration of production and comprehension does not increase the network performance. We end by discussing the value of looking to traditional parallel distributed processing as flexible models for exploring computational mechanisms of conversation.

Keywords: language; interaction; neural networks; production; comprehension

Introduction

What capacities enable linguistic interaction? There are a large number of extant theoretical proposals. A glance at the literature reveals a host of proposed mechanisms that support conversation and other sorts of interactive tasks. Some of these are specific to social or linguistic cognition, such as mirroring and simulation (Oberman & Ramachandran, 2007), mind or intention reading (Tomasello, Carpenter, Call, Behne, & Moll, 2005), linguistic alignment (Garrod & Pickering, 2004), and use of common ground (Clark, 1996). Others have drawn on domain-general cognitive processes, including memory resonance of social identity (Horton & Gerrig, 2005), perceptuomotor entrainment (Shockley, Richardson, & Dale, 2009), synergies (Fusaroli, Rączaszek-Leonardi, & Tylén, 2014), one-bit information integration (Brennan, Galati, & Kuhlen, 2010), coupled oscillatory systems (Wilson & Wilson, 2005), executive control (Brown-Schmidt, 2009), brain-to-brain coupling (Hasson, Ghazanfar, Galantucci, Garrod, & Keysers, 2012), and situated processes (Bjørndahl, Fusaroli, Østergaard, & Tylén, 2014).

Many of these proposals are individually supported by rigorous experimentation or corpus analysis. However, language happens in the "here-and-now" (Christiansen & Chater, in press) and thus must satisfy a plurality of constraints at the same time: from the perceptuomotor level all the way "up" to social discourse and pragmatics (Abney et al., 2014; Fusaroli et al., 2014; Louwerse, Dale, Bard, & Jeuniaux, 2012). Accordingly, there remains a need to systematically compare and articulate the contributions of the suggested mechanisms in an integrative model of interactional language performance (Dale, Fusaroli, Duran, & Richardson, 2013).

In this paper, we propose a computational framework that enables flexible combination and comparison of different cognitive constraints. We show that coupled simple-recurrent networks are capable of learning sequential structure from latent-semantic analysis (LSA) representation of wordforms in interactive transcripts. As a first case study, we use a traditional neural-network approach to test the role of production-comprehension integration during natural language performance (MacDonald, 2013; Pickering & Garrod, 2014).

Production, Comprehension, and Prediction

We look to production-comprehension integration to illustrate this framework. The relationship between production and comprehension is a key factor in most theories of language processing. In research on conversational or task-based interaction, these two systems are granted considerable and often distinct attention. Does language production vary more as a function of internal constraints of the speaker, or more in response to the needs of his or her listener (for some review, among many, see Brennan & Hanna, 2009; Ferreira & Bock, 2006; Jaeger, 2013)? A prominent recent theory takes these systems to be deeply intertwined. Pickering and Garrod (2014; see also MacDonald, 2013) have argued that an integration of production and comprehension is critical in understanding the mechanistic underpinnings of interaction.

Experimental and neuroimaging work suggests simultaneous involvement of both aspects of language processing during linguistic interactions (e.g, Silbert, Honey, Simony, Poeppel, & Hasson, 2014). However, explicit cognitive modeling can more directly reveal the extent to which, for example, prediction and understanding are improved as a function of

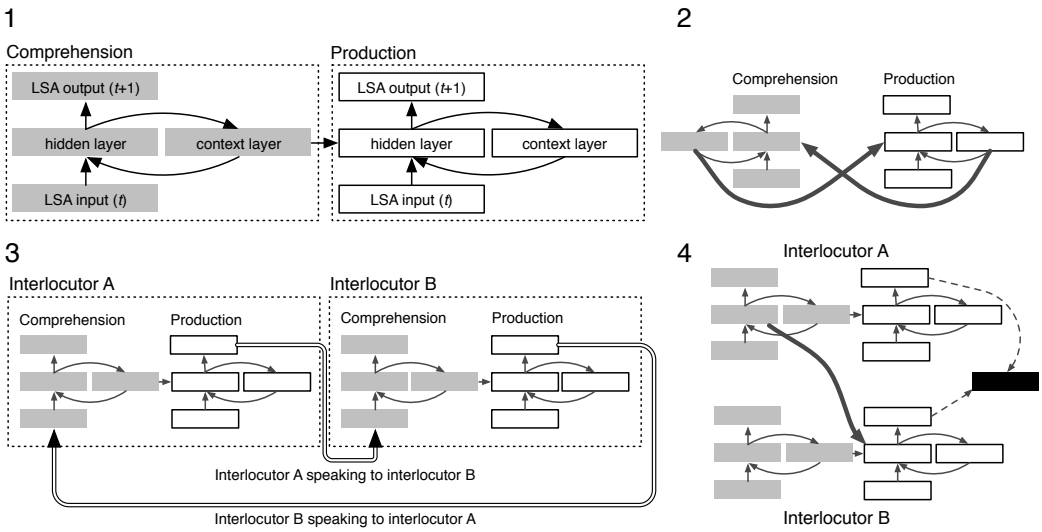

Figure 1: (1) A representation of two coupled simple-recurrent networks (SRN) inspired by Chang (2002). A conversant is modeled as a two-SRN agent. A pair of coupled subnetworks is referred to as an *agent network*. (2) In the original Chang (2002) model, production did not influence comprehension. We model the complete integration of production-comprehension by having these two subnetworks share internal states. (3) A conversation can be modeled as a coupling between two such "nets of nets," leading to a second-order recurrent network. Each agent receives input from the other, and shares the hidden states of its comprehension subnetwork with the input layer of its production when it is its turn. We refer to this second-order network as a *dyad network*. (4) This framework can be parameterized to investigate, for example, the effect of explicitly externally shared information between interlocutors, akin to emerging common ground (black box with dotted lines), or the extent to which one network is facilitated by having access to the "internal state" of another network (thick solid line).

tighter integration. In what follows, we describe one way to model human interaction using parallel distributed processing (PDP). Inspired by a predictive approach to language, we adapt the models of Elman (1990) and Chang (2002) to couple neural networks into two interacting systems, and show that such a model can be parameterized in various ways to test computational claims.

Higher-Order Recurrent Dynamics

We draw inspiration from the successful PDP model of Elman (1990) and adapted by Chang (2002) to investigate sentence processing in a single cognitive agent. The architecture of this simple-recurrent network (SRN) is shown in Fig. 1, panel 1. This network receives input in a comprehension subnetwork. In Chang (2002), this was modeled as a set of input sentence primes. The hidden state of the comprehension network (activation of nodes at the hidden layer) then constrains the production subnetwork, and influences its subsequent performance. Such a network has been shown to effectively model syntactic priming effects (Chang, 2002).[1]

Each person in an interaction can be represented as a pair of SRNs – receiving input and generating output with production and comprehension subnetworks. Modeling conversation then involves coupling these neural network architectures into a "dyad." We couple these nets by taking the output of "speaker" and use it as the input of the "listener," as shown in Fig. 1, panel 3. On a turn-by-turn basis, we can switch who is doing the producing and comprehending. The networks are trained to predict word sequences in this way, in the context of a coupled "conversation." As shown in Fig. 1, panel 3, there are two levels of coupling in this model. These first-order networks (*agent network*) are coupled in their comprehension and production subnetworks in some way. Interaction is modeled as a coupling between two such networks, as a second-order recurrent network (*dyad network*).

This model can be readily adapted to parameterize constraints on processing. In Fig. 1, panel 2, we show that we can "complete the circuit" in the dyads by connecting production to comprehension in the same way. This simple modification inspired two conditions in a preliminary simulation. First, we studied the ability of dyad nets to predict words in interaction under the original formulation, with only comprehension constraining production. We then tested the contribution of full comprehension-production integration by completing the circuit, and compared its performance to the original formulation.

Like any cognitive model, this framework requires an input space that provides structure to the task. Elman (1990) used simulated sentences generated by a simple grammar, and Chang (2002) used hand-coded semantic and syntactic representations in a simplified grammar. To get input vectors for our model, we used transcripts from an interactive task in which two participants communicate to jointly solve a perceptual task (Fusaroli et al., 2012). Taking the word-by-word

[1]Note in Fig. 1 that Chang's original model only included the constraint on production from prior comprehension.

sequences in these transcripts, we created input activations based on latent semantic analysis representations. This reduces the dimensionality and sparsity of the input space and makes the learning problem more tractable for the network. It also tests the framework with complex naturalistic data.

Input Corpus: LSA Word Vectors

The corpus consists of 16 dyads (32 Danish-speaking individuals) totaling more than 1,600 joint decisions, 25,000 word tokens and 1,075 unique word types.[2] Given the sparsity of the lexical space, we transformed the corpus into a latent semantic analysis (LSA) representation (Landauer & Dumais, 1997). This projects words into a lower dimensional feature (vector) space based on how the words occur in the corpus. We define a word's relative cooccurrence to another word by using a simple 1-step window, so that the cooccurrence of word w_i with word w_j is the total number of times they followed each other, $f(i,j) = NP(w_{i,t}, w_{j,t+1})$, where N is the total number of words in sequence, and P the joint probability that words i and j occurred together at times t and $t+1$, respectively. This count serves as an entry in a 1,075 $\times$ 1,075 matrix M, as the entry at the i^{th} row and j^{th} column. This matrix is, of course, quite sparse, because most words do not cooccur with every other word. LSA was employed as a means to overcome such sparsity, providing a lower-dimensional representation of word similarity based on these distributional patterns: $[U,S,V] = \text{SVD}(M)$.

The left eigenvector matrix (U) provides a more compact representation for individual words. Rather than a complete (but sparse) representation across all 1,075 of its column entries, the SVD solution that LSA uses allows us to take a much smaller number of columns of U instead. These columns represent the most prominent sources of variance in the distributional patterns of the word usage.

When inspecting the singular values (S) of the SVD solution in an LSA model, we find that word usage across all transcripts can be captured by about 7 of these columns of U. A schematic of how we use these feature vectors is shown in Fig. 2, which illustrates a pattern of activity across 7 nodes as the input for these networks. This gives us a 7-dimensional representation of words, where activations can be negative or positive, which requires some modification to the training of our SRN subnetworks.

Training with LSA

Because common backpropagation assumes an activation range of $[0,1]$, we had to modify the input and output activation transformations to suit a $[-1,1]$ range. To do this we changed the standard sigmoid function, used as output activation function, to a tanh function that has the desired properties. In order to propagate error back, we differentiate the tanh function at the output nodes. Because derivative $d\tanh = 1 - \tanh^2$, we obtain

$$\delta_{\mathbf{o}} = \mathbf{o} \circ d\mathbf{o} \circ \mathbf{e} = \mathbf{o}(1 - \mathbf{o} \circ \mathbf{o}) \circ \mathbf{e} \quad (1)$$

Where $\mathbf{o}$ is the output vector of the network, $\mathbf{e}$ is the error associated with each node, and $\circ$ represents elementwise multiplication. $\delta_{\mathbf{o}}$ reflects the error assignment to output nodes. Once this is calculated, we can modify the weights $W_{\mathbf{h}\to\mathbf{o}}$ with

$$\Delta W_{\mathbf{h}\to\mathbf{o}} = \alpha \mathbf{h}^T \delta_{\mathbf{o}} \quad (2)$$

α represents the learning rate parameter, and $\mathbf{h}$ the hidden unit activations of a given subnetwork. We used this approach to modify the weights connecting hidden and output layers. All other layers were treated in the common way with the sigmoid function and its derivative, in accordance with traditional iterated backpropagation.

In order to train the networks using LSA vectors as they interact in dyads, we follow the process illustrated in Fig. 2. In a turn-by-turn fashion, the production subnetwork of one agent net would be trained to predict its "spoken" output, while the comprehension subnetwork of the other agent net would receive this output as input and predict it in a word-by-word fashion.

Simulation Procedure

Training and Testing To assess how well the models capture interactional structure of the empirical data, we trained 16 dyad networks in each of two conditions (comprehension to production only vs. full integration). Each network was trained on one pass on the full transcripts of 15 dyads (almost 25,000 word presentations) and then tested on the remaining target dyad. We set α to .01, and the number of hidden units across all subnetworks to 10.[3] We built a baseline control for each test dyad by shuffling its word order, thus disrupting the sequential structure that the networks were expected to learn. The 'A' or 'B' designation of interlocutors was randomly assigned, but used here for convenience of presentation.

Our performance measure was based on the common measure of cosine between the output and target vectors. Cosine is commonly used with the LSA model, since it captures whether word vectors are pointing in the same direction in "semantic space." Cosine varies from $[-1,1]$, with higher values reflecting better predictions by the network.

Predictions First, we expected that the "content" shared between speaker and listener, projected in LSA space, should allow the networks to learn the statistical structure of interaction. Second, we contrasted three hypotheses about production-comprehension integration. (H1) Fully integrated production-comprehension systems would benefit performance, as the networks are able to receive "more in-

[2] Space limitations prevent us from fully describing the construction of this semantic representation, but we note that we also included a "turn end" marker to ensure that words adjacent across turns were not treated as if they were spoken in the same sequence of words by one person.

[3] Space restricts our parameter search, but we found, in general, that hidden layer size did not greatly impact performance in any conditions in our explorations.

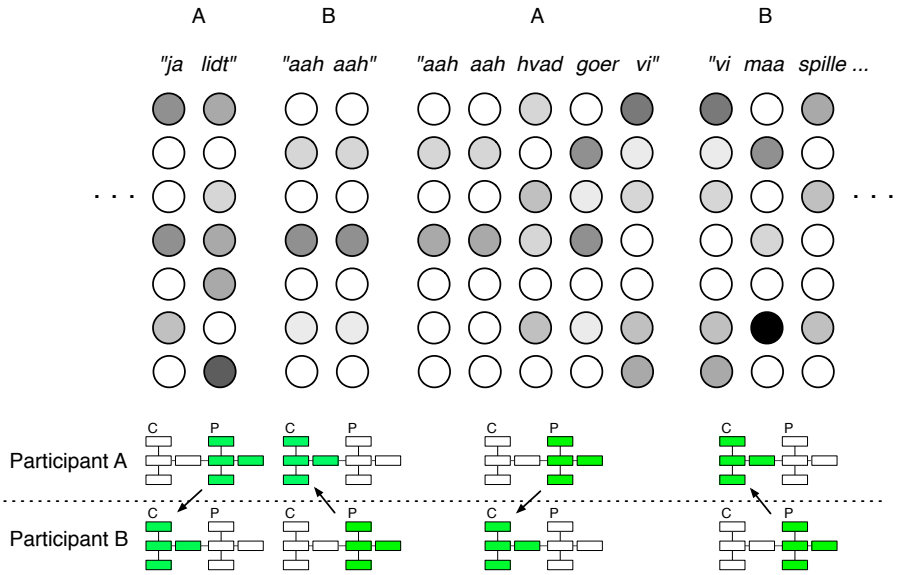

Figure 2: We organized network training by interactive turn. For a given turn, one participant (A or B) is doing the talking. We take the LSA vectors (visualized as a distributed pattern of activation) and have the production network of the speaker on that turn predict its output, and the comprehension network of the other participant predict its input. Within each dyad, the subnetworks of each participant take turns learning to predict the LSA vectors.

formation," in that the comprehension net is now receiving input from production. (H2) Fully integrated production-comprehension systems degrade performance as they introduce noise to the network and an additional set of weights that the network has to learn. (H3) There will be no difference between these networks: Our simplified task has the production and comprehension networks doing very similar things, and so we may not observe any divergence in their performance.

Results

Can Dyad Nets Learn Sequential Structure? When comparing networks in both conditions, it appears that they are very similarly effective at predicting word-by-word LSA vectors in unseen interactions, and that they also show much better performance than the control baseline, in which words are shuffled. This means that networks are processing the order of LSA features, and not simply capturing the activation space in which these LSA features reside. This learning effect is quite large, and is shown in Fig. 3. The appropriate test here is a paired-sample t-test, since each network and its control are trained on matched sets of words with the same network. A t-test across all four subnetworks shows the expected result, for both conditions: t's > 25 and p's $< .000001$.

Does Integration Improve Prediction? The average cosine performance did not differ between the two network conditions, using the same paired-sample t-test across layers, $t(63) = 0.33, p = .74$. This absence of an effect is quite evident in Fig. 3. No reliable difference emerges in direct comparison of any of the subnetworks, either.

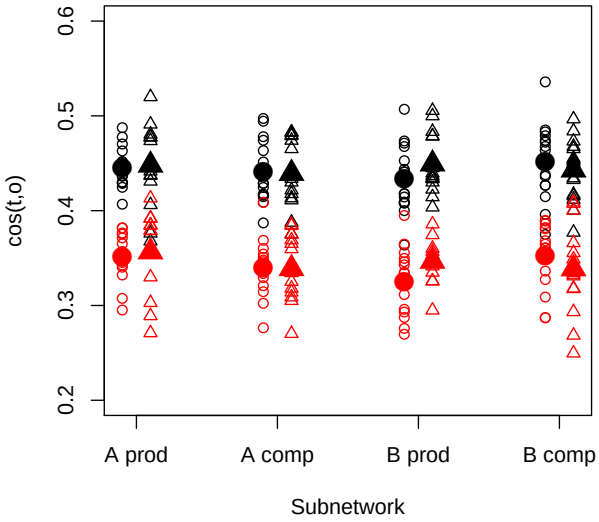

Figure 3: Dyad networks are capable of learning interactional structure. The cosine for agent subnetworks trained on sequential structure show greatly increased scores relative to baseline subnetworks, for which temporal order of the LSA training vectors are shuffled. In general, agent nets with comprehension $\Rightarrow$ production (circle) do not perform differently from agent nets with integration (triangle). They do both show better performance than the control (red). The models are both learning to predict LSA vector sequences. $cos(t,o)$ stands for the cosine of target and observed output vectors.

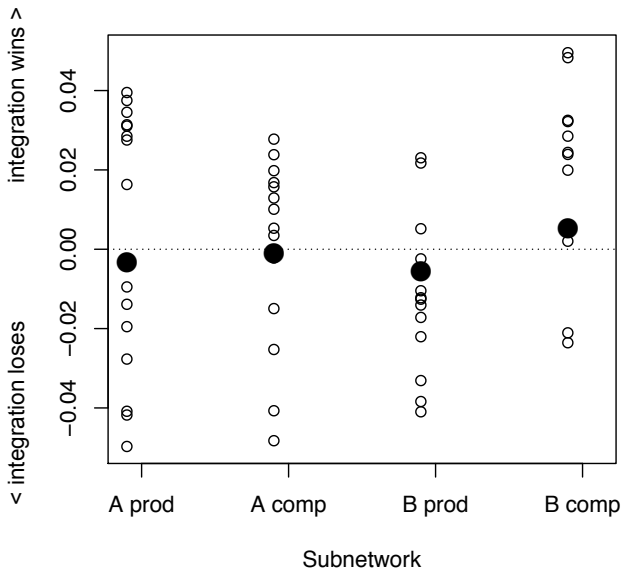

Figure 4: Difference between integrated and unintegrated agent network conditions relative to their respective baselines. It reflects how much more one network can be expected to exceed its baseline relative to the other condition. If integrating production and comprehension improves performance, we expect a positive value on the y-axis.

Does Integration Improve Prediction above Baseline? These results are shown in Fig. 4. In general, as might be expected from the prior analyses, the models are not different from each other in most subnetwork performance. All results are non-significant, with the initial agent net configuration not different from its baseline relative to that of the fully integrated networks, $|t|$'s < 1.

General Discussion

In this paper, we described a flexible computational framework to investigate the cognitive mechanisms underlying linguistic interaction. The first step in this direction is the implementation of coupled neural networks to learn from interaction data. We demonstrated that this adaptation of Chang (2002) is capable of learning the sequential semantic structure in raw, noisy input.

Based on the current debate on interactive alignment, we manipulated their internal cognitive structure to contrast two theoretically motivated models: (i) a model with full comprehension-production integration, and (ii) a model without integration. These alternative coupled networks were then used to model real conversational data in order to investigate hypothesized prediction benefits of full integration. Our results did not reveal an effect of full integration. Put simply, hypothesis (H3) seems to have been supported here: In this computational system, full integration does not bring great gains, if any. Why did we not observe clearer results? To conclude, we outline theoretical and methodological considerations that hint at possible explanations and motivate future implementations of the framework.

First, the results can be interpreted to suggest that 'internal' production-comprehension coupling is in fact not facilitating mutual prediction in this context. This could indicate that recurrent (and thus predictive) structure resides on levels other than the turn-by-turn organization of the conversation. In fact, a recent study (relying on the same corpus) suggests that linguistic patterns critical to performance in the task tend to straddle interlocutors and speech turns making turn-by-turn alignment secondary to recurrent structural patterns at the level of the conversation as a whole (Fusaroli & Tylén, 2016). A future implementation of the model could directly test such ideas (sometimes referred to as the interpersonal synergy model of dialogue: Fusaroli et al., 2014) and compare the performance to other types of conversational interaction that might entail different functional organization.

These results might also be contingent upon a number of methodological limitations that will need to be overcome in future developments. First, the sample size is not impressive and a bigger corpus would possibly enable better training of the networks. Second, in order to deal with the sparse lexical space of real conversations, we reduced the input to LSA vectors. As a consequence both the comprehension and production subnetwork end up dealing with the same kind of data. Integrating comprehension does therefore not add information that is not already contained in the LSA vectors processed in production subnetworks. Thus, the integration is at least partially redundant and cannot be expected to add much to the performance of the model.

There are also more general limitations to overcome. For example, anticipatory dynamics in agent networks should allow overlap at the turn level, as seen in natural interactions. This is a critical feature for modeling the higher-order dynamics of interaction. The PDP approach embraces such computational extensions. For example, networks could be gated, such that off/on states of the production subnetwork will have to be learned by agents. The recurrent property of these networks *should* allow them to predict forthcoming turn switches. The approach offers much in the way of extension, as these networks are, after all, nonlinear function approximators over any arbitrary sets of temporal constraints. For example, we could also develop other input spaces, such as multimodal constraints from nonverbal aspects of interaction, and add them to the verbal components we have explored here.

This flexibility also permits more focused theoretical explorations. The constraints on these networks have theoretical implications that can be readily adapted to further compare and integrate proposed mechanisms, the topic that began this paper. For example, Fig. 1, panel 4 showcases how we might develop the framework to test combinations of other constraints on interaction, such as "common ground." Another example is how internal constraints from one agent network might constrain, and possibly facilitate, the dynamics of the agent to which it is coupled in the dyad network. Theoretically motivated manipulations of this kind would allow more explicit tests of the relationship among these various proposals for the mechanisms of interaction, and compar-

isons to related computational frameworks (e.g., Buschmeier, Bergmann, & Kopp, 2010; Reitter, Keller, & Moore, 2011).

Acknowledgments

Thanks to the Interacting Minds & Center for Semiotics at Aarhus University for its support in bringing the co-authors together for a meeting last year in January, 2015 to discuss this work. Thanks also to Andreas Roepstorff for fun and productive discussions during our visit. The co-authors vibrantly discussed the theoretical status of such a PDP framework, and did not come to a consensus about that status. It did not detract from the fun.

References

Abney, D. H., Dale, R., Yoshimi, J., Kello, C. T., Tylén, K., & Fusaroli, R. (2014). Joint perceptual decision-making: a case study in explanatory pluralism. *Frontiers in Psychology*, *5*.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085.

Bjørndahl, J. S., Fusaroli, R., Østergaard, S., & Tylén, K. (2014). Thinking together with material representations: joint epistemic actions in creative problem solving. *Cognitive Semiotics*, *7*(1), 103–123.

Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two minds, one dialog: Coordinating speaking and understanding. *Psychology of Learning and Motivation*, *53*, 301–344.

Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, *1*(2), 274–291.

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, *16*(5), 893–900.

Buschmeier, H., Bergmann, K., & Kopp, S. (2010). Modelling and evaluation of lexical and syntactic alignment with a priming-based microplanner. In *Empirical methods in natural language generation* (pp. 85–104). Springer.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*(5), 609–651.

Christiansen, M. H., & Chater, N. (in press). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Dale, R., Fusaroli, R., Duran, N., & Richardson, D. C. (2013). The self-organization of human interaction. *Psychology of Learning and Motivation*, *59*, 43–95.

Ferreira, V. S., & Bock, K. (2006). The functions of structural priming. *Language and Cognitive Processes*, *21*(7-8), 1011–1029.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms quantifying the benefits of linguistic coordination. *Psychological Science*, 931–939.

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147–157.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, *40*, 145-171.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11.

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, *16*(2), 114–121.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142.

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in Psychology*, *4*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*(8), 1404–1426.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*.

Oberman, L. M., & Ramachandran, V. S. (2007). The simulating social mind: the role of the mirror neuron system and simulation in the social and communicative deficits of autism spectrum disorders. *Psychological Bulletin*, *133*(2), 310.

Pickering, M. J., & Garrod, S. (2014). Neural integration of language production and comprehension. *Proceedings of the National Academy of Sciences*, *111*(43), 15291–15292.

Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*(4), 587–637.

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, *1*(2), 305–319.

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, *111*(43), E4687–E4696.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(05), 675–691.

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, *12*(6), 957–968.