UNIVERSITY OF CALIFORNIA,
MERCED

Three Scales of Symbol Grounding: From Neural Resonance, to Embodied and Context-Sensitive Language Processing, to Collective Cognitive Alignment

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive and Information Sciences

by

J. Benjamin Falandays

Dissertation Committee:
Professor Michael J. Spivey, Chair
Professor Paul E. Smaldino
Professor Jeffrey Yoshimi
Professor Chris Kello

2022

The dissertation of J. Benjamin Falandays
is approved and is acceptable in quality and form for
publication on microfilm and in digital formats:

Professor Christopher Kello

Professor Jeffrey Yoshimi

Professor Paul E. Smaldino

Professor Michael Spivey, Committee Chair

University of California, Merced
2022

DEDICATION

For my grandmother, Frances, and my nephew, Nico.

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I am incredibly grateful to so many people without whom this would never have been possible: First and foremost, my primary advisor, Michael Spivey, who gave me the freedom to explore every one of my crazy ideas, and from whom I've learned so much about how to be an effective scientist, writer, and mentor. Two of my committee members, Paul Smaldino and Jeff Yoshimi, have acted as additional advisors in everything but an official capacity, and have been no less instrumental in my education. And my fourth committee member, Chris Kello, has been a constant source of eye-opening conversations and insightful feedback. I couldn't have possibly had a better, more supportive, or more patient team behind me. Thank you all.

I have been lucky enough to be taken in by several incredible intellectual communities throughout my PhD, which have been crucial sources of inspiration, feedback, and new ideas. First are the ModSoc and Philosophy labs, which have been second homes for me at UC Merced. Next is the NRT Intelligent Adaptive Systems program, which exposed me to many ideas that now permeate my thinking, and generously offered me several semesters of funding. And finally, the Diverse Intelligences Summer Institute, which absolutely blew my mind and introduced me to so many friends and mentors.

I want to take this opportunity to thank several individuals who have served as important mentors to me before my time at UC Merced: Joe Toscano, my masters advisor, as well as Chip Folk, Irene Kan, Diego Fernandez-Duque, and Matt Mattell at Villanova; Alan Fox, Ray Peters, Helene Intraub, and Fred Adams at the University of Delaware; and Brian Magargal and Neil Kane at Salesianum high school. I would not have reached this point without your support.

I must also thank the people outside of my academic life that have kept me going with their friendship, love, and support. I have made many life-long friends in Merced, including Tim Shea, Chelsea Gordon, Sam Spevack, Jordan Ackerman, Joshua Clingo, LJ Jerome, and Jenni Duran. My bandmates, Matt Turner, Charlie Berrier, Korey Ebinger, and Tevin Williams, gave me a weekly source of catharsis and respite from the life of the mind. I would also like to thank the broader Merced community that has welcomed me in and made this place feel like home—especially the patrons and owners of 17th Street Pub and The Partisan, where I've had countless nights of music, laughter, and great conversation.

Last but not least, I want to thank my wonderful girlfriend, Nicole, who has been the most incredible support system, and without whom I'm not sure I could have ever finished this PhD. Her family, the Cardosos, have taken me in as one of their own and kept me stocked with a constant supply of delicious cookies. Finally, I owe so much to my own family, especially my parents, who have supported me both emotionally and financially in every stage of this process. This is for you guys.

VITA

J. Benjamin Falandays

EDUCATION

Doctor of Philosophy in Cognitive and Information Sciences **2022**
University of California, Merced *Merced, CA*

Master of Science in Experimental Psychology **2017**
Villanova University *Villanova, PA*

Bachelor of Arts in Philosophy, Psychology **2013**
University of Delaware *Newark, DE*

REFEREED JOURNAL PUBLICATIONS

Falandays, J.B., Nguyen, B., & Spivey, M.J. (2021). Is prediction nothing more than multi-scale pattern-completion of the future? *Brain Research.*

Falandays, J. B., Spevack, S., Pärnamets, P., & Spivey, M. J. (2021). Decision-making in the human-machine interface. *Frontiers in Psychology.*

Falandays, J.B., Brown-Schmidt, S., & Toscano, J.C. (2020). Long-lasting gradient activation of referents during spoken language processing. *Journal of Memory and Language.*

Falandays, J. B. & Spivey, M. J. (2020). Theory visualizations for bilingual models of lexical ambiguity resolution. In R. Heredia & A. Cieslicka (Eds.), Bilingual Lexical Ambiguity Resolution. *Cambridge University Press.*

Falandays, J. B., Batzloff, B. J., Spevack, S. C., and Spivey, M. J. (2020). Interactionism in language: From neural networks to bodies to dyads. *Language, Cognition, and Neuroscience, 35*(5), 543-558.

Falandays, J. B. & Spivey, M. J. (2019). Abstract meanings may be more dynamic, due to their sociality: Commentary on "Words as social tools: Language, sociality and inner grounding in abstract concepts" by Anna M.Borghi et al. *Physics of Life Reviews.*

Spevack, S. C., Falandays, J. B., Batzloff, B., & Spivey, M. J. (2018). Interactivity of language. *Language and Linguistics Compass.*

ABSTRACT OF THE DISSERTATION

Three Scales of Symbol Grounding:
From Neural Resonance, to Embodied and Context-Sensitive Language Processing,
to Collective Cognitive Alignment

By

J. Benjamin Falandays

Doctor of Philosophy in Cognitive and Information Sciences

University of California, Merced, 2022

Professor Michael J. Spivey, Chair

This dissertation brings together a collection of four projects that are thematically related through their relevance to the "symbol-grounding problem" in cognitive science: the issue of how the internal activity of a cognitive agent (i.e. neural activity at a biological level, or "representations" at a cognitive level) are *meaningfully* connected to things in the world. This is a general problem for theories of cognition, which must be solved to have adequate theories of *language* more specifically—linguistic meaning is not possible unless we can explain how *meaning* is possible in general. I begin in chapter one with an overview of the symbol-grounding problem, situated in the debate between computational (representational) theories of cognition and non-representational theories such as ecological psychology, enactivism, and embodied cognition. In chapter two, I present a computational model of "neural resonance," which offers an account of the representational role of neural activity that does not require thinking of representations as "encodings" of things in the world, and therefore may not fall prey to the symbol-grounding problem. I show how simple homeostatic mechanisms at the level of neurons may give rise to transient localist representations that can control the action-perception loop of an agent, and also leads to emergent

prediction-like behaviors in language processing. In chapter three, I present two human subjects experiments that investigate the possibility that language comprehension is grounded in sensorimotor simulation. Abstract or metaphorical language is a critical test case for this hypothesis, and I report evidence that such language does not generally depend upon sensorimotor simulation, but that literal language does not *always* depend upon it; rather, we observe a continuum of sensorimotor involvement in language processing. In chapter four, I report the results of another human subjects study on the context-flexibility of phonetic representations in Spanish-English bilinguals. This experiment provides support for the notion that perceptual categories crucial for language processing can be flexibly adapted to fit the current context. In chapter five,I present an agent-based model of "collective cognitive alignment" which addresses a crucial step in the emergence of language: the coordination of shared perceptual categories. Finally, in the concluding chapter I reflect on how we may construct a theory of cognitive science that takes *meaning* seriously, and allows us to preserve an unbroken "grounding wire" as our theories move from the cognitive activity of the simplest lifeforms up to the most complex forms of cognition as exemplified in human language and culture.

Chapter 1

Introduction

This dissertation brings together a collection of projects that are thematically related through a focus on the "symbol-grounding problem", the issue of how a cognitive agent can connect its internal activity, which we may call "symbols" or "representations," to referents in the world (Harnad, 1990; Searle, 1980). While there has been much debate about the precise requirements for grounding, Harnad (1990) has suggested that two properties may be key: (1) the capacity for internal activity to pick out referents in the world, and (2) conscious experience. This problem may be framed as being about the phenomenon of *meaning*: when a cognitive agent is successfully able to ground its representations, we may say that the agent has access to the meaning(s) of those representations. Meaning, in turn, would seem to be closely bound up with our concept of what is for an agent to have conscious experience, or perhaps even to have a *mind* at all. That is, if we observe a system that looks and acts like an intelligent agent, and yet on closer inspection the system seems incapable of grasping the *meaning* of the information that it has access to, then we will likely say that the system does not actually have a mind. For all of the progress that cognitive science has made towards developing advanced artificial intelligence, we are no closer today to anything resembling a mechanistic answer to the symbol-grounding problem, and indeed, as I will argue throughout this manuscript, dominant theories of cognition associated with the computer metaphor of mind may be fundamentally unable to ever offer such an answer. Thus, cognitive science—the interdisciplinary study of mind—would render the world entirely mindless.

Some cognitive scientists may be perfectly okay with a field that leaves behind the issue of meaning. Perhaps meaning is irrelevant, so long as we can predict behavior with sufficient accuracy, and create artificial systems that do what we would like them to do. Or perhaps meaning and conscious experience are simply illusions, "epiphenomena" of no real consequence at all. Some may think that meaning matters, but is simply too hard of a problem to deal with, and would prefer to focus on more pragmatic concerns. Others go so far as to suggest that meaning is not a properly

scientific problem, or that the questions have been poorly framed and will dissolve entirely after some semantic clarifications. To all of these charges I would disagree strongly, though I will not spend time here trying to refute this position directly. Instead, I will try to let the work that follows speak for itself in showing that a cognitive science of meaning is both possible and valuable. For now, I will simply suggest that *if* living organisms indeed experience the world as full of meaning, *if* such experience is at all important for explaining how and why living organisms do what they do, and *if* our theories of cognition are fundamentally unable to account for such experience, then perhaps we have missed something big.

The symbol-grounding problem is relevant both in theories of cognition quite broadly (e.g. why a pattern of neural firing in your brain, or a mental representation, means something to you), and more specifically in the study of language (e.g. why a word means what it means). These two domains are related in important ways. Language and other symbolic systems represent the most *complex* manifestations of meaningful cognition of which we are aware, but they must be built atop fundamental systems of meaning at the level of individual cognition. That is, one cannot get to the meaning of a word unless one can get to the meaning of the pattern of neural firing that occurs when one hears or reads that word. Thus, if our dominant theories of cognition cannot deal with meaning generally, then they will also fail to explain language. And while one could certainly undertake a study of meaning at the individual level without bringing language into the picture, there may be several benefits to keeping in mind the most complex forms that meaning can take on. For one, this can engender a greater awareness of where our theories are likely to fail—to be "unscalable"—and thereby grow more open to reconsidering some entrenched assumptions that have led us astray. Furthermore, if language is built on top of more basic systems of meaning, then perhaps we can feed two birds with one seed, learning about how humans have meaning *at all* by studying how they get the meaning of a word or sentence. Towards these ends, some of the work in this dissertation will focus on the more complex, difficult cases even *within* the domain of language, such as the comprehension of abstract or metaphorical sentences (chapter 3), and the coordination of systems of meaning among individuals with no prior shared knowledge structures (chapter 5). These phenomena may provide a conservative test of how likely our existing theories are to succeed in the long run.

Ultimately, however, scientists interested in meaning would like to end up with a theory that is not anthropocentric, and will not force us to suggest that the phenomenon of meaning magically popped into existence at the birth of the first spoken word or even the first *H. Sapiens*. We may also wish to avoid the opposite extreme that could be called "pan-psychism," the suggestion that meaning has *always* been around and is in everything to some extent—a fundamental force of the universe. A middle-way would be desirable: a scientific theory that explains what meaning is and how it works, how it arises where there was once none, which systems "have" it and which don't, and how it may evolve into more complex forms. That is, we should like to have a continuous tether from the world of physics to the world of human

language—an unbroken "grounding wire." In service of this goal, I have organized this dissertation to proceed from from the smallest scale of analysis on which I have approached the problem of meaning to the largest (though we could just as well have gone in the reverse order).

Chapter 2 will address the phenomenon of meaning at the level of an individual organism interacting with its environment, focusing in particular on the role that the nervous system plays in the phenomenon of meaning. Using a toy model—a reservoir computer—I will attempt to offer a how-possible account of the way that a bundle of thoughtless cells, with no sense of a world outside the skull, can give rise to an organism that "represents" aspects of its environment that are meaningful for action. This model demonstrates how simple homeostatic[1] mechanisms at the level of individual cells, when brought together in the right way, can exhibit a kind of collective intelligence that seems to encode features of the environment and even predict what will happen in the near future. However, this model neither encodes nor predicts, but instead "resonates" with patterns of stimuli. Models of this kind may offer a theoretical bridge between cognitivist theories of mind, based heavily on the notion that the brain's job is to "infer" the world outside of the mind, and alternative theories that tend to eschew the concept of representation entirely, but may then struggle to explain what it is that brains have to do with cognition. I will try to show how this concept of "neural resonance" applies equally well to the case of a single organism's action-perception loops, where we take the environment to be the physical surroundings, as to cases in which we treat the linguistic behavior of others as the "environment" being navigated. This chapter will serve as a conceptual foundation as we work our way upwards towards more complex manifestations of meaning.

Chapter 3 presents two behavioral experiments that address the extent to which comprehension of language is grounded in systems of action and perception. Theories of "grounded cognition," such as Barsalou's (1999) Perceptual Symbol System approach, have suggested that humans may reach the meaning of a word of sentence by partially reactivating regions of the brain associated with perception and action, or mentally "simulating" the physical meaning of an utterance. While many important results have come out of this work, cases of abstract, metaphorical, and idiomatic language remain difficult to explain within such a framework. For example, how is one to understand the meaning of "infinity" through a pattern of sensorimotor activity? Research on the involvement of sensorimotor systems in the processing of such abstract language has produced some conflicting results, which I attempt to shed some light on. I consider the possibility that a distinction between language that is grounded in sensorimotor activity versus ungrounded may be overly simplistic; Instead, there may be a continuum of sensorimotor involvement, determined both by immediate context as well as by more global statistical relationships among words in a language. This is demonstrated, in Experiment 1, through a test of the "action-sentence compatibility"

[1] technically "allostatic," though I will prefer the former term as it may be slightly more familiar to some readers

(ACE) effect, which reveals that the degree to which responses are influenced by the mis/match between the action implied by a sentence (e.g. hand-related action in literal/metaphorical context: "Edgar caught your ball/attention") and the action required for a response (i.e. responding with the hand or foot) varies as a function both of current context, and the overall contexts in which a word is usually encountered. Experiment 2 deployed an eye-tracking "Visual-World" paradigm design in order to more closely examine how such effects play out over the course of sentence processing. This study sheds light how language processing may nonetheless involve the body *without* involving the reactivation of sensorimotor systems, with action actually *facilitating* abstract understandings by interfering with literal ones.

Chapter 4 applies a similar focus to questions of the contextual-flexibility of language processing, this time considering how immediate context or recent experience shapes low-level perceptual processes. In a mouse-cursor tracking study, Spanish-English bilinguals and monolingual English speakers categorized artificial phoneme stimuli as either a /b/ or /p/ sound. In one condition, Spanish-English bilinguals saw a brief page of instructions in Spanish, while in another the instructions were presented in English. Participants were required to click on pictures that corresponded to /b/ or /p/ words in either language (e.g. *beso/peso* or *bear/pear*). These quite subtle manipulations (relative to others that have been used in the past) were sufficient to trigger a shift in the category boundaries, such that bilingual participants divided up the phonetic space more like a monolingual Spanish speaker in one case, and more like a monolingual English speaker in another. Both chapters 3 and 4 point to ways that perceptual processes and comprehension of signals are "tuned" by the context and what the body is doing. These are not the highly intentional processes of a "rational" processor, where one first identifies a context that is then used to infer the meaning of some word or phrase. Instead these are very subtle, rapidly adjustable, and automatic processes by which action, the body, and the immediate physical and social environment alter your understanding and behavior.

The fuzziness and context-sensitivity of language processing revealed in Chapters 3 and 4 might seem like a recipe for disaster, from an engineering perspective. How are we supposed to exchange meanings across individuals when our processing and response to the very same signal can change, outside of our conscious awareness, on even a moment-to-moment basis (let alone on the scale of days, weeks, months, and years)? This is the issue I will begin to take up in Chapter 5. I argue that in many ways the fuzzy, noisy, and embodied nature of human cognition is not (merely) a challenge for language processing, but in fact the very thing that allows it to exist at all. Using an agent-based model that adapts a model of individual category learning to a cultural setting, I show that constraints on communication, including transmission noise, limitations on lifespan and learning, and aspects of demographic structure all contribute in crucial ways to the possibility that a shared space of signals can form and stabilize where once there was none. Conversely, we see that "high-fidelity" transmission may paradoxically undermine the possibility of a stable symbolic system.

Taken as a whole, this body of work may begin to shed some light on how we can build an unbroken "grounding-wire" of cognitive theory, from the processes that allow an individual organism to have meaning up to those that give rise to complex systems of symbolic meaning, such as human language. Because this collection of work does not *quite* bring our grounding wire all the way down to the earth—instead starting with a model of human cognition that is already somewhat complex—it will first serve to fill in some details about what "meaning" is at its most fundamental level. In what follows, I present a brief overview on theoretical stances within cognitive science that are relevant to meaning and symbol grounding, which will place this work in historical and philosophical context, and set the stage for many of the specific topics covered later on.

1.1 On Information and Meaning-Making

In a historical treatment of the concept of information, Hoffmeyer and Emmeche (2014) invite the reader to consider the latin root of the word—*informare*: "to bring something into form." This etymology suggests that the original meaning of the term referred to a *process* of imparting knowledge by altering the shape of a system. In contrast, they suggest that the modern, technical concept of information "reflects the atomization of knowledge which has been a scientific ideal through the last hundred years," with information now being viewed primarily as a substance that can be transmitted. This historical shift is exemplified in the concept of information developed from Claude Shannon's seminal work on Information Theory.

With the goal of developing communication technology that could efficiently communicate in the face of noise, Shannon operationalized information as the reduction of uncertainty or surprise upon receiving a signal. Formalizing the reduction of uncertainty requires that we also specify an *expectation*: a set of possible messages and the associated frequency distribution with which those messages are observed. If only one signal is possible, uncertainty will be 0; A receiver will always know which message is going to be received. However, if there is more than one possible signal, then we have some uncertainty as to which signal will be observed at any given time. Receiving a signal, then, can be understood as reducing uncertainty, with the magnitude of the reduction dependent upon the frequency distribution of possible signals—the less probable the signal observed, the more substantially our uncertainty has been reduced. Shannon redefined information as precisely the magnitude of this reduction, which can be measured in "bits." Thus, the less probable a signal, the higher the informational content.

Notice, however, that this definition of information requires there to be a prespecified set of possible signals with a stable frequency distribution. Neither condition necessarily obtains in natural systems. Moreover, Shannon himself admitted that his theory considered the aspects of information that were stripped of their semantic con-

tent. Once a receiver has confirmed the identity of a signal, deciphering its meaning requires *decoding* the signal by virtue of a code that maps signals onto meanings. For example, in the case of Morse code, sequencees of electrical pulses are decoded into alphabetical and numeric characters.

However, even this decoding process will not straightforwardly get us to the meaning of a signal. Even once a receiver has identified and decoded a message, such as decoding electrical pulses into strings of text, how are they to know what the resultant *text* means? If we assume that a receiver now needs to use a *mental* code to map strings of text onto meanings, we simply begin an infinite regress of decoding, which at no point grounds out in true meaning.

Nonetheless, many cognitive scientists, psychologists, and neuroscientists take this to be exactly what brains are doing: encoding and decoding signals (Brette, 2019). This is particularly true for theories of language processing, which has resulted in a marked discontinuity between treatments of linguistic meaning and biological meaning. The fundamental limitations of this view of cognition have been spelled out over many years by the philosopher Mark Bickhard, who explains that "encodingist" views have no way to explain the normative content of mental representations (Bickhard & Terveen, 1996), because encodingism is characterized by the idea that mental representation is fundamentally a causal correspondence relation between internal (brain) states and external states of the world. Bickhard writes 2009a:

> "If the causal relationship exists, then the representation exists, and it is correct; if the causal relationship does not exist, then the representation does not exist at all. These are the only two possibilities; they leave no way to account for the case in which 'a representation exists but is false about what it is representing."'

Thus, encodingist models leave mental representations entirely arbitrary, devoid of the normative content that gives them meaning for the organism.

Along similar lines, Brette (2019) has also recently argued that the coding metaphor is more harmful than helpful in the cognitive sciences, because assigning meaning to neural codes requires knowledge of experimental context, which is only available from the perspective of an outside observer, but not from the subjective perspective of the agent in whom the "code" is observed. For example, consider an experiment in which a researcher records neural activity from a participant who views stimuli that vary only in color (wavelength). While the researcher can potentially recover the wavelength of a stimulus from the observed neural activity—and thus may take neural activity to be encoding wavelength—neurons that are responsive to wavelength may also be responsive to light intensity. Experimenters interested in studying the neural encoding of wavelength will of course hold light intensity constant, but the brain of the participant has no way to know that this is the case. Thus, from the perspective of the brain, it would be impossible to tell whether its own activity corresponded to

changes in wavelength, light intensity, or both. Based on several examples of this kind, Brette (2019) concludes that "neural codes have much less representational power than generally claimed or implied." The potential for neural codes to serve as the foundation for representation has also been shaken by several recent demonstrations of the phenomenon of "representational drift," wherein the correspondences between neural activity and stimulus properties may change dramatically over the period of days or weeks Deitch, Rubin, and Ziv (2020); Rule, O'Leary, and Harvey (2019); Schoonover, Ohashi, Axel, and Fink (2020).

In summary, neuronal "codes" (1) are unable to carry normative content, (2) do not actually have the power to encode stimulus properties in a context-free way and without supposing an idealized observer, and (3) are highly unstable. As such, models of cognition built on the computer metaphor and the syntactic notion of information derived from Information Theory are fundamentally unable to account for the phenomenon of meaning. This "symbol grounding" issue has long been recognized in cognitive science, and all attempts to deal with it within an encodingist framework have inevitably reached a logical circularity. We cannot, for example, suppose that internal representations simply borrow their normative content from other internal representations. This point is made clear by Searle's famous "Chinese Room" thought experiment, and is also supported by Hume's argument that it is impossible to derive norms from facts Cohon (2004). We also cannot suppose that representations obtain their normative content by virtue of some isomorphism with the world (i.e. mental representations are a "copy" of the world), as this would force us to conclude that organisms already in some sense know the things that they are to represent. One might try to circumvent this problem by supposing that evolution has granted us normative representational content innately, as Fodor argued, or that it is obtained through learning. However, both evolution and learning themselves depend upon the normativity we would wish them to provide—if representations cannot be wrong, there can be no error, and thus no selection, learning, or adaptation.

Major attempts to deal with the symbol-grounding problem, all of which in some way try to narrow the theoretical fissure between perception and action, can be arranged along a continuum from fully internalist to fully externalist approaches to cognition, relating to whether or not cognition is taken to be a "skull bound" phenomenon. On the internalist pole we can place theories of grounded cognition, exemplified in the work of Larry Barsalou 1999, as well as approaches based on the free-energy principle Friston (2010). The former framework suggests that symbols in the brain are reactivations of neural patterns associated with perception and action. For example, proponents of grounded cognition might suggest that reading the word "car" results in a partial reactivation of neuronal populations that have been active in previous instances in which one saw a car, grasped a steering wheel, smelled exhaust fumes, etc. This view, however, is still fully "encodingist"—taking mental representations to be fundamentally correspondence relations between brain activity and things in the world—and thus fails to avoid the issues raised above.

Approaches based on the FEP, on the other hand, hold that mental representation is a kind of "controlled hallucination" wherein the brain itself generates the perceptual patterns that are expected to be observed, and refines this generative model based on mismatches between predicted and observed patterns of sensory stimulation. On this and related views associated with the "Bayesian brain" hypothesis, cognition requires inferring the hidden causes of sensation, which are not available directly to an agent due to an informational boundary—a so-called "Markov blanket"—that separates organism and environment. Approaches based on the FEP go some of the way to dissolving the artificial bisection of action and perception in classical "cognitivist" frameworks, by emphasizing that action and perception continuously constrain one another, but they still suffer from an inability to account for the normative nature of mental representations Bickhard (2016a). What the brain has access to, in these models, is a discrepancy between a predicted sensory stimulation and an actual sensory stimulation, but there is no way for those discrepancies to become *about* something for the agent—instead, they can be understood as prediction errors only, again, from the perspective of an external observer who has access to both the state of the world and the activity of the brain.

On the opposite end of the continuum are fully externalist models of cognition, exemplified by ecological and radically embodied frameworks. Based on the concept of "direct perception" originating in the work of J.J. Gibson, these approaches hold that information exists in the environment in the form of structured energetic flows, which are sufficient to specify to an organism the available affordances for interaction M. Anderson and Chemero (2019). As such, proponents of this view suggest that organisms do not need to represent their environment at all, but merely to "attune" the dynamics of the brain and body to the relevant energetic arrays. For example, the movement of an organism generates an "optic flow"—changes in the pattern of light stimulating the eyes—which, by virtue of "(more or less) lawful relationships between the surfaces and the structure of the light" is a reliable cue to the distance between the organism and nearby object, among other things (Balasubramaniam, Riley, & Turvey, 2000; Tsao & Tsao, 2021). For proponents of the ecological view, such as M. Anderson and Chemero (2019), this kind of "ecological information"—the information available in the changes of sensory stimulation as an organism moves around an environment—is "inherently semantic" because it supports action, and thus has normative value for the organism. Ecological and radical embodied frameworks can be considered "realist" views, in that they take information to be something that exists in the world, awaiting for an organism to utilize it.

In the middle of the internalist-externalist continuum sit frameworks that can be described as "mutualist," in that they give roughly equal importance to organism-internal and external processes in grounding meaning. For example, from the perspective of the enactive framework initiated by Varela, Thompson, and Rosch (2017), the ecological approach to meaning "attempts to build an ecological theory of perception entirely from the side of the environment [...and...] neglects [...] the codetermination of animal and environment." The enactive approach emphasizes that what counts as

a "signal" from the environment is determined by the unique sensorimotor coupling between organism and environment—the organism is "attuned" to the environment in a way that carves up the world into signals and noise. For this reason, the enactive approach has been described as "constructivist" in that it frames information not as something that exists independently in the environment, but instead as something that is created in the dynamic relationship *between* an organism and environment.

The enactive approach holds that living organisms are "autopoeitic" systems, which means they act so as to continually maintain and regenerate a boundary that separates the organism and environment, such as a cell wall. Here the enactive approach builds on the ecological approach to meaning, suggesting that the environment is revealed as meaningful to the organism insofar as it presents affordances for action *that contribute to the maintenance of a boundary between organism and environment.* Importantly, by self-producing a boundary, an autopoietic system simultaneously constructs an environment—that which is outside of the boundary. Any organism-environment boundary must necessarily be permeable in some ways and/or at some times, because the organism needs to acquire metabolic resources from and expel waste to the environment in order to maintain the boundary. But while the internal processes that maintain the boundary must be coupled to external processes in some way that suffices to keep the whole boundary-constructing process going, the space of possible couplings is large and can be freely explored. Thus, the physical structure, needs, and perceptual repertoire of an organism may change, thereby altering precisely what is the organism and what is the environment, while a boundary of some kind is nonetheless maintained without interruption. Therefore, on the enactive view, affordances and information are *not* something that exists out there in the world, but instead are dynamically constructed as the boundary between organism and environment is continuously negotiated. However, several authors have recently argued that the ecological approach need not be understood as ignoring the codetermination of organism and environment, and thus that ecological and enactive approaches may be integrated towards a unified, post-cognitivist approach Baggs and Chemero (2018, 2019); Feiten (2020); Heras-Escribano (2019).

Another mutualist approach that shares much in common with the enactive view is the "interactivist" framework initiated by Bickhard (2009a). While the enactive position holds that normativity is grounded in autopoeitic systems that create and maintain their own boundary with the world, the interactivist framework emphasizes that this sort of boundary-constructing process only occurs in systems that are far from thermodynamic equilibrium (Bickhard, 2016a). Thus, Bickhard argues that a missing key to the puzzle of grounding meaning lies in understanding the thermodynamic conditions that allow for self-maintaining systems to emerge Bickhard (2016b). While the second law of thermodynamics states that the entropy (or "disorder") of an isolated system only increases, it is now known that order can increase locally when it serves to dissipate an energy gradient more efficiently. In other words, thermodynamic systems appear to be attracted towards the organization that maximizes the rate of entropy production, which sometimes requires local, temporary decreases of

entropy. This phenomenon in which ordered systems arise from disordered systems in the presence of an energy gradient is known as "self-organization."

Importantly, because order emerges in service of efficiently dissipating an energy gradient, self-organized systems tend to quickly run themselves towards thermodynamic equilibrium. For example, Bickhard (2009b) writes:

> A candle flame maintains above combustion threshold temperature, induces convection, which brings in fresh oxygen and gets rid of waste, vaporizes wax in the wick for combustion, and melts wax in the candle so that it can percolate up the wick [...] A candle flame, however, can only do one thing—burn. It has no options and cannot select among options. If it runs out of wax, for example, there are no alternatives that it has the capacity to select, that might correct this threat to its continued existence.

In contrast, living systems are *recursively* self-maintenant—they "maintain self-maintenance" (Bickhard, 2009b) and act so as to mitigate the tendency to approach equilibrium. This insight is complementary, rather than contradictory, to the enactive approach, and more recent formulations of the enactive approach have explicitly recognized how the self-individuating property of autopoetic systems depends upon far-from-equilibrium dynamics (E. Thompson, 2010). A selectively-permeable boundary, for example, may contribute to maintaining the condition of being far from thermodynamic equilibrium, but such a boundary can only form when a system is already far from equilibrium.

According to Bickhard, the drive of a living system to keep itself far from thermodynamic equilibrium is key to understanding the emergence of normative representations—internal patterns of activity that are *about* things in the world in a way that has inherent value for the system. In the phenomenon of chemotaxis, for example, *E. Coli* can select among several possible actions—swimming straight when moving up a sugar gradient, or tumbling randomly when no sugar gradient is detected—in order to locate food sources. Because these choices may succeed or fail to contribute to self-maintenance, recursively-self-maintenant systems functionally anticipate the conditions of their own survival—they take actions, based on cues, that contribute to survival *if* those cues are reliable indicators of resources, threats, obstacles, etc (see also: Rosen, 2012). As such, the signals that an organism registers from the environment via its sensory apparatus become *about* the conditions for survival in a way that gives them truth value from the perspective of the organism itself. Bickhard (2009b) writes:

> [Given some sensory perturbation] there is an implicit predication that "this" is one of those environments in which the initiated interaction will proceed as anticipated. That predication, therefore, might itself be true or

> false: the environment might or might not be among the supportive kinds [...] Initiating the activity, therefore, presupposes that those supportive conditions hold.

Thus, Bickhard argues that the implicit anticipatory behavior of recursively-self-maintenant, far-from-thermodynamic systems, constitutes a minimal form of *representation*.

If we take meaning and representation to be grounded in anticipatory action-perception loops that contribute to self-maintenance, we can begin to see more clearly why it is misleading to think of an organism's internal activity as constituting an "encoding" or representation of states of the world. The ecological, enactive, and interactivist frameworks emphasize that internal activity does not need to *stand for* things in the world, but rather to regulate an organism's sensorimotor coupling with the world, by virtue of which it maintains itself. The goal for an individual neuron, and indeed the brain as a whole, is not to "represent" but simply to survive, to maintain the conditions of their own self-maintenance. This requires spreading activity around in the brain in such a way as to keep the entire network alive, and it requires that the activity is connected to sensorimotor systems in such a way as to keep the organism moving, obtaining life-sustaining resources, and avoiding threats. In this way, we can see how the emergent normativity of recursively self-maintenant systems offers a foundation for the meaning of perceptual signals from the perspective of the organism. The meaning is not *inside* the neuronal response to a stimulus, but instead is in the way that neuronal activity, bodily processes, behavior, and environmental processes are coordinated in such a way as to keep the whole system going as long as possible.

Chapter 2

Neural Resonance: A Potential Bridge Between Representational and Non-Representational Models of Cognition

2.1 Introduction

As I described briefly in the introductory chapter, an ongoing debate in the cognitive sciences concerns the appropriateness of the computer metaphor for the mind and/or the brain. Thinking of the mind as analogous to a computer was a key inspiration for many thinkers important to the founding of cognitive science as a field some 40 years ago, and remains a popular notion today. For these *Cognitivist* thinkers, cognition is a process of performing logical operations over internal "representations" or "symbols" that stand for entities and ideas. Yet this basic concept actually predates modern computing technology by about a century, going back at least to the psychophysics work of Hermann von Helmholz in the 1850's, who first popularized the notion of perception as *inference*. This view of the mind has always had its detractors, notably in the Ecological school of thought associated with J.J. Gibson, which grew into the more recent movements of Embodied Cognition and Dynamical Systems Theory. Gibson emphasized that organisms have no need to *represent* the world outside, and instead can "resonate" to structured flows of energy—an idea he called "direct perception." Notably, in eschewing the notion of representation, this latter school of thought has tended to focus on what goes on at the level of the organism and environment, leaving open the issue of how neural activity figures into the story. Recently, there have been increasing calls to finally reintroduce neural dynamics into the picture (Raja, 2018, 2021).

While the debate over the validity of the computer metaphor has continued to rage, some have claimed that it is a matter of mere semantics. For example, Richards and Lillicrap (2021) argue that if one defines "computer" simply as "some physical machinery that can in theory compute any computable function," then the brain is clearly a computer in a very literal sense (but so are many other things that we often do not think of as computers). But if one's definition of "computer" includes all of the baggage commonly associated with how most computing technology works today—especially the idea that internal activity corresponds to an *encoding* of things in the world—then a computer clearly becomes a very *weak* metaphor for the brain. It is the ubiquity of this "encodingist" position, as Bickhard has called it (1996), that leads the symbol-grounding problem to rear its ugly head time and again, and which makes this debate much more than a matter of semantics. The underlying issue at the heart of the computer-metaphor debate is therefore not *whether* we can usefully think of the mind as a computer in some respects, but instead how, *if* the mind is like a computer in the sense of operating over internal representations, are we to explain that organisms seem to experience actual things in the world, and not just meaningless symbols in the mind? And if, on the other hand, we were to abandon the computer metaphor on the basis of this symbol-grounding problem, then what on earth is going on inside the head?

It was necessary to begin this section by clarifying the *particular* commitment of the computer metaphor position with which non-representationalists take issue—the encodingist commitment—else it would seem contradictory to turn around and use *another* type of computer to offer a potential path forward in this dilemma. In this chapter, I will describe a simple model of cognition as a *reservoir* computer, which does not perform logical operations over stored representations or encodings. Instead, this computer consists of a set of selfish nodes, analogous to individual neurons, that act only locally in order to keep their own activity near a viable target level. Through individual nodes adjusting connection weights with neighbors and internal parameters in order to maintain homeostasis, the network as a whole comes to *resonate* to structured patterns of energy out in the world. In doing so, the network also produces "transient localist representations", as Rodny, Shea, and Kello (2017) have described them: semi-stable patterns of activity that drift over time, and which could potentially be *read* as encodings by an outside observer, though we know that the computer itself has no access to meanings by virtue of that activity. At the same time, we can see that local, homeostatic regulation leads the network to become functionally anticipatory, which Bickhard (1996) has argued is a key foundation for normativity (read: *meaning*). That is, the network appears to take action on the basis of what it "predicts" will occur next, implying that its transient localist representations have *meaning*—adaptive value that is available to the network itself, and not just decodable for an outside observer. This simple model may offer a bridge between the cognitivist and non-representational schools of thought, by showing that internal "encodings" may be an emergent product of a brain "resonating" to structured flows of energy in its environment.

To build this argument, I will first offer a brief background on reservoir computing models, emphasizing the features that make them a better model of cognition than standard localist, sequential processors. Then, I will describe the model in detail and present two case studies. First, I will show how this model can control an action-perception loop in a simple agent, organically producing object-tracking behavior. Second, I will show how these same dynamics apply to language processing, and all that is required is to think about the linguistic behavior of conspecifics as the changing "environment" that an individual must navigate. In the latter case, I will show how the reservoir network produces behaviors that closely resemble the signatures of "predictive coding" theories, despite the fact that the reservoir makes no predictions. These case studies may help to demonstrate how we need not throw away the notion of representation entirely in order to have a theory of cognition that does not fall prey to the symbol-grounding problem, although we may need to make substantial changes to the way that we understand the role of brain activity in many cognitive processes.

2.2 Background: Reservoir Computer Models

The model presented in this chapter is a simplified form of a reservoir computing model (Kello, 2013; Lukosevicius, Jaeger, & Schrauwen, 2012; Szary, Kerster, & Kello, 2011). The canonical reservoir computer consists of three layers: an input layer, a reservoir layer, and an output layer. The reservoir layer typically contains a large number of nodes that are sparsely interconnected via non-updating random weights, with each node possessing a nonlinear activation function. We can understand the reservoir as a variant on Elman's (1990) simple recurrent network, where the context layer and the hidden layer are now one highly interconnected set of nodes, and no learning is performed on their weights. The activity of the reservoir layer can be described as a projection of a relatively low-dimensional input pattern into a much higher-dimensional space. This high-dimensional mapping has the potential to carry a lot of information that can subserve complex mappings between inputs and outputs. However, the representation in the reservoir layer is complex, noisy, and distributed across many nodes, making it difficult to interpret directly. Therefore, canonical reservoir computing models must learn a mapping from the reservoir layer to a set of output nodes by virtue of a teaching signal. When this training signal is the next input pattern, we could fairly say that such a network is doing predictive coding (i.e. explicit use of predictions). More often, though, reservoir computing models are used to perform complex pattern classification functions, which are not thought of as predictive.

Kello (2010) introduced a reservoir computer model with spiking nodes that use a "self-tuning" algorithm to turn on and off local connections in pursuit of a "critical branching ratio"—that is, a situation in which each spike tends to produce one additional downstream spike, on average. When a reservoir network achieves a critical

branching ratio, spikes will propagate through the network without resulting in the network "freezing up" (i.e. all nodes becoming fully active or fully inactive). The critical branching ratio has been shown to maximize the informational capacity of the network. The resulting fluctuations of the network allow it to contain a kind of implicit, contextual memory, because spike patterns from the previous step will influence processing of the current inputs. Information about the present input as well as about past inputs is preserved in the intermixed, complex, instantaneous spike pattern of the reservoir. And when input is suddenly cut off from the reservoir layer, the endogenous activity of the network may persist for a few timesteps, producing a behavior known as "fading memory." Critically, this kind of "memory" is not best thought of as a representation of the past, but instead as a kind of momentum from the past into the future. Under the right contexts, we suggest that such pattern completion may be indistinguishable from an ostensible "prediction."

Kello and colleagues have shown the applicability of their reservoir computer model to a wide range of phenomena in cognitive science (Dale & Kello, 2018; Kello, 2013; Kello et al., 2010; Kello, Kerster, & Johnson, 2011; Rodny et al., 2017; Szary et al., 2011). First, this model provides enhanced biological plausibility over prior connectionist models of cognition, in that it produces phenomena that have been observed in human brains, such as neural avalanches and power-law scaling of fluctuations in spike patterns. Szary et al. (2011) showed that this model has the capacity to represent visual motion, in that it is able to integrate activity over many timesteps. Rodny et al. (2017) emphasize the ability of the network to produce *transient* localist representations: semi-stable patterns of activity in response to particular inputs, that changes over time. In this way, the reservoir computer model naturally produces the phenomenon of "representational drift" that has recently confounded some neuroscientists (Deitch et al., 2020; Rule et al., 2019; Schoonover et al., 2020). Rodny et al. (2017) suggest that these transient representations may be crucial for the context-sensitivity of cognition, allowing internal activity to change subtly or dramatically in a matter of moments to adjust to a present situation. Dale and Kello (2018) suggest that the reservoir computer offers a viable model of the composite nature of linguistic meaning, by virtue of the properties of dynamic memory, timescale integration, and multimodal integration, which allows the network to be sensitive to patterns of information across many different sources and scales.

The application of the reservoir computer as a model of cognition is generally consistent with the view of mental representation as trajectories of a system through a state space of activity, such as described by Yoshimi (2012) and Onnis, Farmer, Baroni, Christiansen, and Spivey (2008). This view may allow us to retain the concept of "representations" where it is useful, while remaining grounded (i.e. *without* thinking of representations as encodings). Nonetheless, this position still requires us to radically reconsider what it is that representations do. They are not, from this perspective, akin to atoms of cognition that enter into logical operations, as some early proponents of the Cognitivist school of thought suggested. Nor do these types of representations "carry" any meaningful content. Rather, representations understood

in this fashion correspond to the forward momentum of a system that has entrained to patterns in the environment. When such a process of entrainment results in semi-stable, recurring patterns of activity in the brain, these may appear to an outside observer as "encodings" of aspects of the environment, but these patterns never need to be "decoded" for the cognitive system itself.

2.3 Model Description

2.3.1 Conceptual Overview

When we zoom in on the brain, we see living cells, competing for resources and trying to maintain homeostasis in a rapidly changing environment, rather than tiny prediction or inference engines. Neurons are often studied as homeostatic or allostatic systems (MacLean, 2003; O'Leary & Wyllie, 2011; Turrigiano & Nelson, 2004): all neurons need to remain within some viable range of activity or else atrophy and die, but the "preferred" range of activity is highly variable across individual neurons. Neurons also adapt to perturbations so as to maintain viability. That is, they have self-preserving mechanisms (e.g., modifying synaptic receptors, changing membrane potentials, etc.) that allow them to retain viability under variable conditions. Adaptation of this kind can be thought of as a form of implicit memory, because it means that changes in response to past inputs are carried into the future and will influence future outcomes. Successfully adapted individuals may appear to us to have been prepared for future conditions as if they predicted those conditions, but most readers will probably agree that individual neurons do not make nor represent predictions. In this section, we introduce a computational model designed to illustrate how functionally-predictive behavior and apparent encodings or representations may emerge in a distributed network of simple, homeostatic systems, like neurons. However, this model also serves to illustrate that the functional-level description may only hold under certain conditions, and requires an outside observer to interpret it as a prediction. In other words, while we can read the model's behavior as prediction, there are no predictions represented inside the system.

Conceptually, one can think of the nodes in our network as individual agents that must obtain a minimum amount of energetic resources in order to survive. Energetic resources flow into the network via external perturbations. However, external perturbations do not reach every node, and even nodes that do receive external perturbations will not receive one on every time step. Furthermore, nodes will continually leak energy at a fixed rate. To deal with this variability in energetic influx, nodes that do not receive the minimum amount of energy from external perturbations will try to seek out input from neighboring nodes. This is represented in the model as an increase in positive connection weights with neighboring nodes.

However, too great an influx of energy to an individual is also undesirable. Nodes that receive too much energy have three strategies available simultaneously. First, they may try to decrease their positive input from neighbors, or even receive inhibitory input. Second, they may try to increase their "metabolism," dissipating more energy per unit of time. The latter strategy is implemented in the model as an increase in the "target" input value at each time step, and represents the idea that nodes can adapt to handle increased amounts of input. However, the target value has a floor, such that all nodes need some input to survive. Nodes will try to stabilize their actual input near their target value through some combination of input received from external perturbations, and input received from neighbors. Third, nodes that receive a rapid influx of energy far above their target energy may "spike," dissipating most of their input at once, and simultaneously emitting a signal that may become informative to neighbors.

Critically, in order for nodes to stabilize their activity, they must receive the right amount of input at the right time. For example, a given node may be stimulated by an external perturbation that comes at t_1 and t_3, but receive no external perturbation at t_2 and t_4. Thus, at $t_1$and t_3, the node may need inhibitory input from neighbors to remain near its target activity level, while at t_2 and t_4 it may need excitatory input from neighbors. This means that nodes must learn to implicitly anticipate the temporal structure of perturbations, as well as the temporal structure of spiking activity in neighboring nodes, in order to stabilize their own activity.

2.3.1.1 Reservoir Network and Node Properties

Our network contains a set of N processing nodes in the reservoir, where $N = 250$ in the first model below, and $N = 100$ in the second. Each node was randomly connected to other nodes in the reservoir with a link probability of $p_{link} = .1$, such that each node had approximately 25 neighbors in the first model, and 10 neighbors in the second. The weight of each link was randomly initialized by drawing from a normal distribution with mean 0 and s.d. of 1.

Each node n is characterized by 4 variables: a current activation level x_n, initialized at 0; (2) a fixed leak rate lr of .25 (e.g. if the activation level of a node is 1 at time t, the activation level will be .75 at time t_{+1}, in the absence of further input); (3) a variable target activation level, initialized at $T_n = 1$; (4) and a variable spiking threshold T'_n, which was always equal to $2T_n$ (e.g. when the target T_n was 1, the spiking threshold T'_n was 2). The value of the target T_n was given a lower bound of 1 (the value at initialization), ensuring that all nodes needed at least some continuous, positive input in order to remain near their target value.

When the activation level of a node is greater than or equal to the threshold value T'_n, it spikes, broadcasting a signal value of 1 to its neighbors. If the node does not achieve its threshold value, it fails to spike, broadcasting a signal of 0 to its neighbors. Thus,

the network-endogenous input (i.e. excluding external perturbations to the network, discussed below) of each node n at time t was simply the sum of n's weights with neighbors that spiked at time $t-1$ (see Eq. 2.1 below). Activation cannot go below 0.

When a node spikes, it "dissipates" some activation by subtracting the threshold value. For example, if a node spikes with an activation level $x_n = 10$, a target value $T_n = 5$, and a threshold value $T'_n = 10$, its activation will immediately drop to 0. If the same node spikes with an activation level of $x_n = 15$, its activation will instead drop to 5. As we will describe further below, the effect of this rule is that, if a node is to remain close to the target activation level, its activity must either remain under the threshold value, so as to minimize the dissipation of activity, or go substantially over the threshold value, so as to cancel out the effect of dissipating activity upon spiking. Thus, neurons that spike with only a small margin will recruit more input from active neighbors, resulting in a pseudo-Hebbian effect whereby nodes increase weights with neighbors that fired on the previous time step.

2.3.1.2 Activation Dynamics

On each iteration, the input from external perturbation as well as from within the network is summed for each node. The activation vector x of the reservoir at time step t given input vector i can be written as follows:

$$x_t = x_{t-1} \bullet lr + i \bullet W_i + s_{t-1} \bullet W_x \tag{2.1}$$

In Eq. 2.1, lr represents the leak rate (.25), W_i represents the input weight matrix, s_{t-1} represents the vector of spikes at the last time step (1 for a node that spiked, 0 for a node that did not spike), and W_x represents the recurrent weight matrix inside the reservoir. Importantly, W_x contains no self-connections, such that inputs come exclusively from network-external perturbations or from the spikes of neighbors. The activation vector x is then compared with the threshold vector T' to compute the next spike vector s. Nodes that are above the threshold activation value result in a spike value of 1, and nodes that are below threshold result in a spike value of 0:

$$s_n = \begin{cases} 1, & x_n \geq T'_n \\ 0, & x_n < T'_n \end{cases} \tag{2.2}$$

For any neuron that spikes, the activation level will be decreased by the respective threshold level, prior to computing errors:

$$x(s) = x(s) – T'(s) \tag{2.3}$$

In Eq. 2.3, $x(s)$ is the activation vector of neurons that spiked on the current iteration, and $T'(s)$ is the corresponding threshold vector.

2.3.1.3 Learning

In the early time steps of our model, some reservoir nodes may receive one or more external perturbations on a given loop through our grammar, but many of the nodes in the reservoir will receive no external input. As such, input-connected nodes will initially reach their threshold activation level and therefore spike, but most nodes will undershoot both their target value and threshold value, and therefore will not spike. Based on the initially random internal weight matrix, activation will then spread throughout the network, continually intermixing with the activation from external perturbations. Since threshold values are initially relatively small ($T' = 2$ for all nodes at initialization) relative to the magnitude of activation from input nodes (+5), nodes that spike may remain over their threshold value for subsequent iterations, despite dissipating some activation through spiking and some through leakage. This means that some nodes may spike repeatedly as a result of a single input, in the absence of recurrent negative feedback.

On each iteration, every processing node in the reservoir has an opportunity to adjust its weights with other processing nodes as well as its target activation value, with the homeostatic goal of reducing the discrepancy between actual and target values. (Input weights are never altered.) Using terms from control theory, what we are calling homeostasis amounts to a form of "automatic gain control," whereby a processing unit attempts to maintain a stable output level despite variations in input.

The update for weights in the reservoir matrix W_x is determined by three things: the total error E (the difference between the actual and target activation; Eq. 2.3), the number of weights available for updating N_n, and a learning rate L_{W_x}. Critically, nodes can only update connections with neighbors that spiked on the previous iteration. This means that if no neighbors spiked on the previous iteration, the number of weights N_n that can be updated is 0. Otherwise, N_n is equal to the number of neighbors that spiked on the previous iteration. The update for target value T (and, by extension, threshold values $T' = 2T$) is determined only by the total error E and a learning rate L_T. Based on qualitative examination of piloting results, we set $L_{W_x} = .1$ and $L_T = .01$, as these values allowed the model to stabilize around small error levels (mean E .1) relatively quickly (within 1000 sentences, or 4000 iterations) while avoiding the problem of overfitting (i.e. if learning rates are too high, weights/targets may change drastically at every time step, making the model unable to reach an equilibrium).

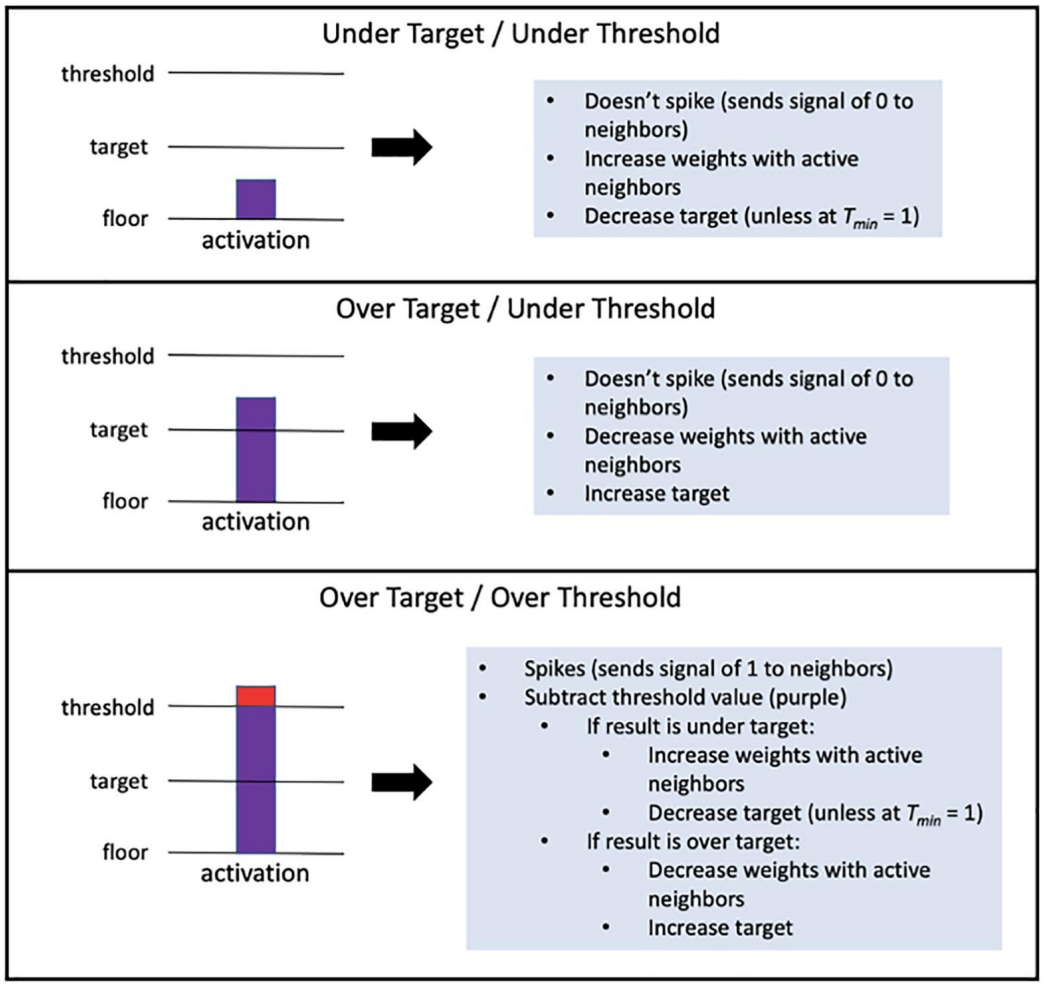


Figure 2.1: A schematic showing the rules for weight/target updating, depending upon a node's current activation level relative to the node's current target level and spiking threshold value (which was always twice the current target level)

$$E_{n,t} = x_{n,t} - T_{n,t} \tag{2.4}$$

Incoming weights to a node n from an active neighbor n' $W_{n',n}$ are adjusted so as to approach the target activation level according to the following learning rule:

$$W[n,n']_{t+1} = W[n,n']_t - \frac{E_{n,t}}{N_{n,t}} L_{W_x} \tag{2.5}$$

Looking at the final term of Eq. 2.4, we can see that the total error E_n is distributed evenly across all N_n weights that are available for updating on a given iteration. See

Figure 2.1. When the error E_n is positive (a node overshoots its target), weights with active neighbors are reduced, and vice versa when error is negative. The inverse operation is applied to target, such that the target is increased when error is positive, and decreased when error is negative, according to the following equation:

$$T_{n,\ t+1} \ = \ T_{n,\ t+1} + \ E_{n,\ t} * L_T \tag{2.6}$$

Note again that E_n is computed after any node that has spiked on the current iteration has "dissipated" its threshold activation level T'_n. For example, if a node spikes with an activation level $x_n = 1.5T'$, its activation after subtracting the threshold value will be exactly equal to its target value ($1.5T' - T' = .5T' = T$), and thus its error on the current iteration will be 0 ($E_n = x_n - T_n = 0$), resulting in no change of the node's parameters. We reiterate this point to emphasize that spiking does not necessarily correspond to a failure to maintain homeostasis for our nodes nor a "prediction error" (although some have proposed such an understanding of spiking activity, e.g. Fiorillo, Kim, and Hong (2014)), but can also be a way for nodes to maintain homeostasis in the face of implicitly "expected" over-stimulation.

2.4 Model 1: Neural Resonance and Action-Perception Loops

In this subsection, I show how the allostatic reservoir network described above may be used to control the action-perception loop of a simple agent embedded in an environment. A still shot of the agent-environment system and relevant model components is shown in the top-left panel of Figure 2.2. Similar to a model from Hotton and Yoshimi (2010), the agent is represented as a circle fixed at the origin of a plane, and the stimulus is represented as a point that moves in a circle around the origin. The agent is imbued with 2 arrays of sensors, analogous to two eyes, positioned at +30 degrees (left sensor, red point) and -30 (right sensor, blue point) degrees relative to the heading angle of the agent. Each eye consists of an array of 120 input nodes, analogous to retinal cells, that are spaced evenly ±60 degrees from the center of each sensor, giving each eye a 60-degree field-of-view in either direction. Each input node in a sensor array spikes in the presence of a stimulus at its current angle. Given that the left and right eyes are positioned 60 degrees apart, and each eye contains sensors extending 60 degrees in each direction, the field-of-view for each eye overlaps in the entire space between them. In other words, when a stimulus is present at an angle that falls *between* the two eyes, both eyes are able to "see" the stimulus simultaneously.

The array of input nodes constitutes a distinct layer from the reservoir network.

Each of the 240 total input nodes was randomly connected to a node in the reservoir network with a probability of $P_{link} = .1$. The activation level of input nodes was reset at each timestep and input nodes did not utilize the allostatic mechanism, but rather were set to spike always and only when the stimulus was present at the position corresponding to each sensor. All weights from the input to the reservoir layer were set to +1, and there were no connections from the reservoir to the input layer.

The stimulus is represented as a point that moves in a circle around the agent at a speed of 1 degree per timestep (green point in the top-left panel of Figure 2.2), therefore taking 360 timesteps for a full rotation. The stimulus began by moving counter-clockwise, and was set to suddenly switch directions every 720 time steps, or two full rotations.

In addition to having two arrays of input sensors, the agent was also given an output layer of two nodes corresponding to "effectors" for turning left (top-right panel of Figure 2.2, red bar) and right (blue bar). Each node in the reservoir was randomly connected to each effector node again with a probability of $P_{link} = .1$. All connection weights from the reservoir to the output layer were set to +1, and there were again no connections in the opposite direction. Like input nodes, effector nodes did not use the homeostatic mechanism and their activity was reset at each timestep. Unlike input nodes, effector nodes did not spike. Instead, the total input to each effector node at each time point was averaged over the number of incoming connections, producing a value between 0 and 1 for each effeector. The difference between the activation level of the left and right effector was computed as $D_{effector} = x_{left} - x_{right}$. When this difference was positive (i.e. the left effector had more input), the agent turned left $10 * D_{effector}$ degrees, and vice versa when the difference was positive.

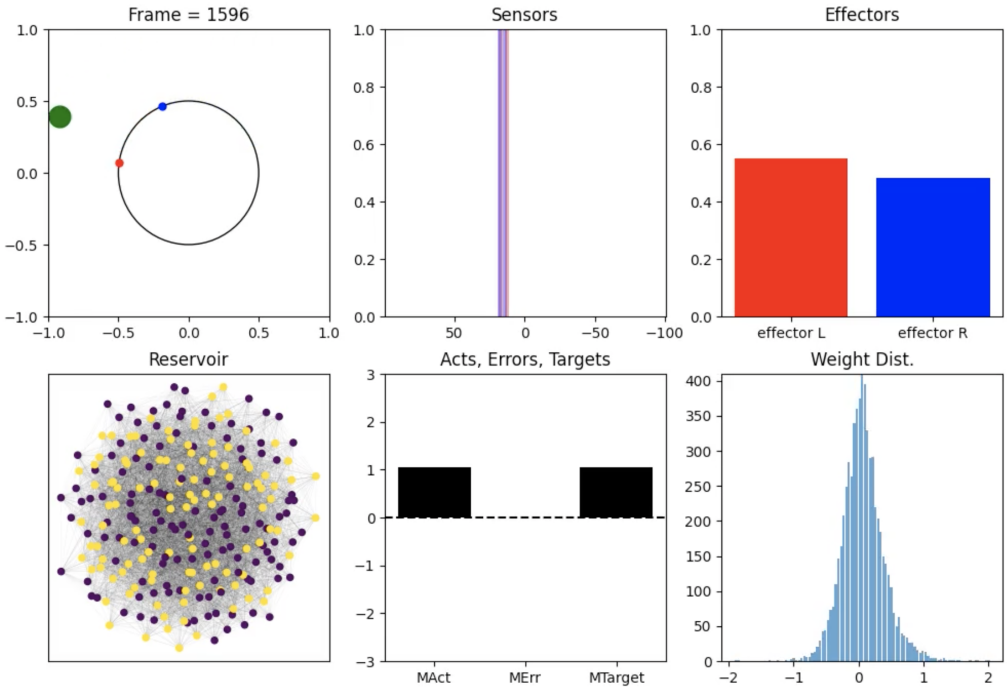


Figure 2.2: A still of the model as it controls the action-perception loop of a simple agent that can turn left or right. The top-left panel shows the agent (large unfilled circle) with two sensors (red and blue points) and the stimulus (green point). The top-middle panel shows the activation level across the array of red and blue sensors. The top-right panel shows the current activation level of the effectors for turning left (red) and right (blue). The bottom-left panel shows the reservoir, with spiking nodes shown in yellow. The bottom-middle panel shows the current mean activation across the reservoir nodes, the mean error (discrepancy between target and activation, and mean target. The bottom-right panel shows the distribution of learned weights within the reservoir.

The simulation begins with the stimulus positioned at 0 degrees and moving counter-clockwise, with the agent facing 90 degrees. When the agent's sensors first detect the presence of the stimulus, activation begins to spread through the network. This activity spreads also to the effector nodes, which initially begin moving the agent erratically left and right. But after the local allostatic mechanism proceeds for a few hundred time steps, we observe a sudden shift of behavior: the agent locks on to the stimulus and begins rotating in the same direction, at a similar speed. This can be seen in Figure 2.3, which shows the heading angle of the agent (black) and the stimulus (green) over 7200 timesteps (20 rotations of the stimulus) in a representative run. At approximately 1250 timesteps, we can see the black points corresponding to the agent begin to track the green ones corresponding to the stimulus. When the stimulus changes directions, the agent turns, with a short delay, to follow it. Changes of direction for the stimulus appear as green "V" or inverted "V" shapes in Figure 2.3. One such change occurs between 2000-2500 timesteps in the displayed run of

the simulation, and here we can see that the black line corresponding to the agent switches direction shortly after the green line corresponding to the stimulus. In some cases, we can see that the agent temporarily loses track of the stimulus, as between 4000-4500 time steps in this run.

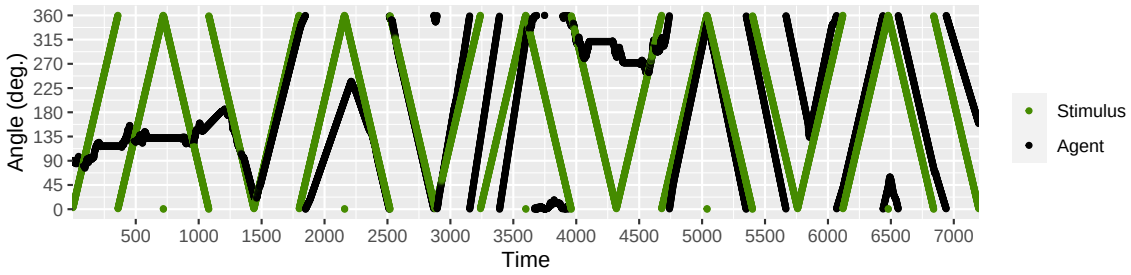


Figure 2.3: The angle of the stimulus (green lines) and the agent (black lines) over time.

Why does this apparent object-tracking behavior emerge in a network that has no explicit directive to track the stimulus? This behavior can be explained by virtue of the fact that it allows the network to stabilize its own activity. When the stimulus initially passes over the sensors, the spikes in the network are initially chaotic. If, when this activity spreads to the effector layer, the agent turns in the *opposite* direction from the stimulus, activity will stop entering the network entirely, and the reservoir will eventually stop spiking until the stimulus comes back around (or the agent comes back around to the stimulus). Because this movement undermines the flow of input into the network, it provides little basis for updating connection weights. That is, nodes can only update connections with neighbors that are spiking, so if the activity of the entire network dies out quickly, no updating will occur for a period of time. But if, on the other hand, the activity that spreads to the effectors leads the agent to turn in the *same* direction as the stimulus, the network will continue to spike for a longer period of time, providing more opportunity for the network to learn. In sum, behaviors that maintain a consistent flow of input to the network are implicitly rewarded, while behaviors that undermine the input to the network are not. In this way, the network spontaneously learns to track the stimulus, "attuning" its own movements to changes in the position of the stimulus.

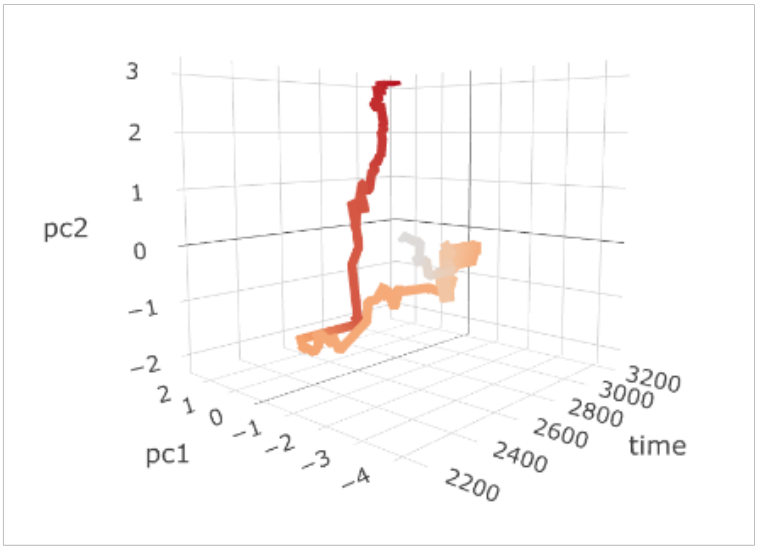


Figure 2.4: The first two principal components of the reservoir network's activity from 2200-2300 timesteps. Earlier timepoints are shown in lighter colors, with later timepoints in dark red.

To better understand what is happening inside the network as this occurs, we can zoom in on the activity of the network during a time period in which the agent is successfully "tracking" the stimulus. In order the visualize the activity across the 250 reservoir nodes, we can begin by reducing the dimensionality to just two dimensions using principal components analysis of the spike patterns over time. Figure 2.4 shows the first two principal components on the x- and y-axes, with time plotted on the z-axis. This figure shows how the activity of the network is changing from 2200-2300 timesteps of the run plotted above. Here we can see that, despite maintaining stable tracking behavior, the network is gradually drifting through its state-space of activity. The network begins in the negative region of PC1, with PC2 at around 0, then moves into positive values of PC1 and negative values of PC2, and finally ends up in the top, back corner of the state-space, corresponding to positive values of both PCs. This illustrates that the stable *coordination* of the agent with the stimulus is associated with gradual drift in the internal activity of the network

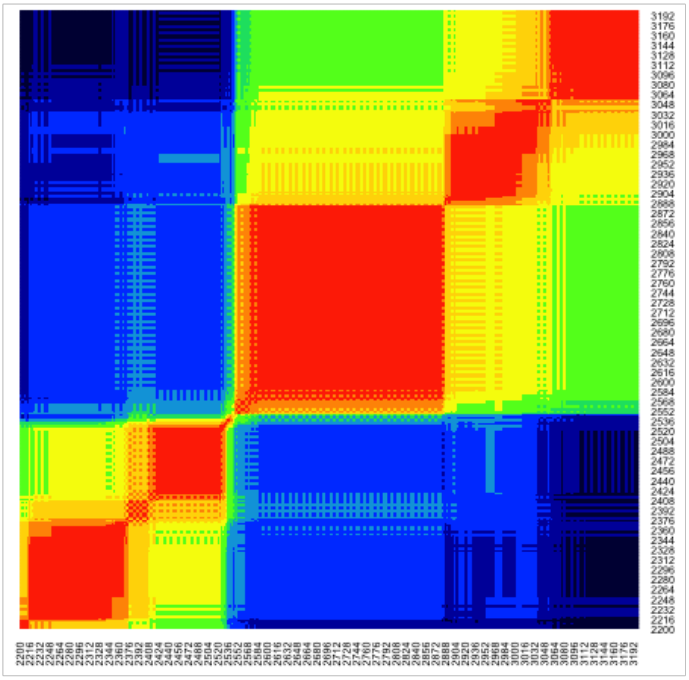


Figure 2.5: The auto-correlation of the reservoir network's spike patterns from 2200-2300 timesteps. Strong correlations appear in bright red, with weaker correlations proceeding through orange, yellow, and green, until the weakest correlations shown in blue.

We may also consider the degree to which this network "reuses" spike patterns over time. This can be visualized by examining the auto-correlation of the network's spike patterns across time, plotted in Figure 2.5. Note that during the plotted time period from 2200-3200 timesteps, the stimulus and agent make two changes of direction: first near the beginning of this time window (around 2250 timesteps) and again near the end (around 2750 time steps). These changes of direction appear in Figure 2.5 as distinct regions in which the activity of the network is strongly auto-correlated. For example, the bottom-left region of Figure 2.5 contains two smaller red regions of high auto-correlation near the diagonal, with only moderate correlations (yellow blocks) between activity across these two time windows. This corresponds to the first change of direction of the agent. Thus, each time the agent switches direction, a semi-stable pattern of activity is found that persists for some time, but this pattern of activity may look quite different at different moments when the agent is apparently doing the same thing. Note, for example, that in the very first and very last regions of Figure 2.5, the agent is successfully tracking the stimulus and moving in the same direction at roughly the same speed, but the network's activity is very weakly correlated across these two regions.

This simple model hopefully may begin to illustrate the utility of thinking as neural activity as "resonating" to flows of energy in the environment. This phenomenon is

inherently multi-scale, and can only be understood by considering brains, bodies, and environments in conjunction. The allostatic nodes in this network seek to maintain viability by getting a consistent level of input over time. If the environment is changing in a structured way, and the agent remains still, allostasis may be achieved simply by adjusting connection weights such that the changing input produces a consistent flow of activity through the network. However, when the agent is allowed to move, driven by its own internal activity, this threatens to undermine stability of input flowing through the network. That is, the change in input is not just driven by a changing environment, but also by the movement of the agent *through* the environment. In such cases, the network needs to find reliable *trajectories* through its state space of activity that keep the input relatively stable for periods of time. These adaptive trajectories can be thought of as transient localist representations that mediate between inputs and outputs, and thus have adaptive value for the network itself. In spite of this value, the network never needs to "decode" these representations into meanings such as "stimulus coming from the left, turn right." Instead, the activity of the network may *functionally* represent such meanings, but only temporarily, and only in the context of the current state of the agent and the environment.

2.5 Model 2: Neural Resonance and Language Processing

Having considered how the concept of "neural resonance" can be used to understand the function of internal representations in a simple action-perception loop, I will next consider how this same concept may apply to language processing. In this case, we may loosen our notion of the "environment" to include not just physical objects, but also the linguistic behaviors of conspecifics. Just as a stimulus moving across the field-of-view generates a changing pattern of activity over the sensors in the first model, we may think of an unfolding spoken sentence as generating a changing pattern of activity over auditory sensors. In order to maintain viability in this case, the network needs to learn something about the temporal patterns present in a language.

Model 2 has a few differences from Model 1. First, this model contains no output or "effector" layer. Instead, this model represents an agent that is passively listening to a series of sentences. Second, since this problem is a bit simpler (as I have modeled it), Model 2 uses fewer nodes in the reservoir—100, rather than 250. Third, instead of having an array of sensors analogous to retinal cells, Model 2 instead has an array of just 5 input nodes, each analogous to a pattern of activity generated for a distinct word. This simple model of linguistic processing uses just 4 different "words," along with a representation of a pause between words, and hence there are 5 different input patterns. In all other ways, the mechanisms of Model 2 are exactly the same as Model 1 above.

2.5.0.0.1 Relevance to "Predictive-Coding" Theories of Language Processing. This model was first introduced by J. B. Falandays, Nguyen, and Spivey (2021), and was framed in the context of "predictive-coding" models of language processing. Before delving into the model's behavior, it is worth briefly reviewing the relevance of the model in that domain. Predictive coding models are the subclass of predictive *processing* models in which prediction errors are the signals between processing units. That is, predictive coding models make *explicit* use of predictions in order to learn, whereas predictive *processing* represents a more general category of models that produce prediction-like behaviors, but do not necessarily represent prediction errors. Predictive coding models, such as those associated with the Free-Energy Principle and the Bayesian Brain Hypothesis, are arguably the modern torch-bearer of the computer-metaphor of mind, and very much take the activity of the brain to correspond to encodings of things in the world (or more technically, encodings of *discrepancies* between expected and encountered stimuli in the world). As such, predictive coding models are subject to the symbol-grounding problem. If it can be shown that many of the behaviors that have been taken as evidence for predictive coding in the brain can *also* be produced by a model such as ours, then this would suggest that we need not fall back on predictive coding to explain many cognitive phenomena.

A key pattern that has been taken to support predictive coding in the brain is a reduction in signal when stimuli are predictable, such as the reduced N400 for semantically regular sentences (Kutas and Federmeier, 2011) or reduced reading times for predictable words (Smith and Levy, 2013). This pattern is hypothesized if, as the prediction-error-minimization framework holds, feed-forward connections in the brain primarily carry error signals, which should be smaller in magnitude when expectations are met (Clark, 2013, Rao and Ballard, 1999). Nonetheless, Luthra et al. (2021a) recently reported analogs of this prediction-error signal inside TRACE, which does not make explicit predictions. Luthra et al. found that both the total amount of lateral inhibition at the word layer and the total amount of feedback from the word layer to the phoneme layer, two indices of competition and activity in TRACE, were reduced when the incoming phonetic features were predictable. This occurs because predictable inputs are, by definition, inputs for which the information contained in later segments is partially redundant with respect to early segments. In other words, the guess that a rational agent would make based on partial information is likely to be correct, and full information merely corroborates this guess. So when TRACE gets predictable incoming phonetic featural input, this means that the true lexical target is likely to be among the set of lexical nodes already most strongly activated based on early word segments, and later segments will only confirm a winner, if necessary. But when inputs are less predictable, this means that bottom-up input in later segments is inconsistent with the set of lexical candidates that first became active, leading to more lexical nodes becoming active and greater competition occurring among them. Thus, a reduction in signal for predictable inputs (i.e., a signature of predictive coding) can also emerge from a pattern-completion system such as TRACE, without recourse to prediction error signals.

Luthra et al. (2021a) compared the dynamics seen in TRACE with an SRN, which does make explicit predictions. However, Luthra et al. point out that the error signal of an SRN is essentially external to the network—it is not a function of the activity inside the network, and does not even need to be present for a trained model to function (in fact, error signals are typically absent during testing). The nodes of the SRN's hidden layer, which are more analogous to population codes in the brain, do not show the signal reduction that is the hallmark of predictive coding—if anything, quite the opposite. Thus, it appears that the neural signatures of predictive coding in the brain may actually be more consistent with the pattern-completing, non-predictive behavior of TRACE than with the explicitly predictive behavior of an SRN.

In what follows, we will show that these same hallmarks of predictive coding *also* appear in our reservoir network. This serves to demonstrate how simple allostatic mechanisms that look only into the recent past can generate patterns of behavior that appear to extend adaptively into the future—in other words, to serve as predictions of future input. If, as I have suggested in the introduction to this chapter, that anticipatory behavior is foundational for normativity or *meaning* that is available to a cognitive agent directly, and *not* just decodable for an outside observer, then we may understand this model as offering a "how-possible" explanation for the grounding of language.

2.5.1 Input

The input to this network was a sequence of "sentences" of the form [subject, verb, object, space]. There were two possible noun inputs ("man", "dog") that could serve as both subject and object, and two possible verb inputs ("walks", "bites"). With the addition of an input encoding a "space," this created 5 total inputs. The input sequences were generated by moving probabilistically through a transition matrix, shown in Figure 2.6. This created 8 possible sequences that appeared with the probabilities shown in Table 2.5.1. These inputs were represented as 1-hot encodings across five input nodes (e.g. input 'man' is represented as [1, 0, 0, 0, 0]). Input nodes were connected to the main network through an input connectivity matrix, which was generated stochastically. Each input node was randomly connected to nodes in the reservoir with a link probability of Plink = .1. In addition to the randomly initialized internal weight matrix, these were the only sources of stochasticity in the model, which otherwise operates entirely deterministically subsequent to initialization.

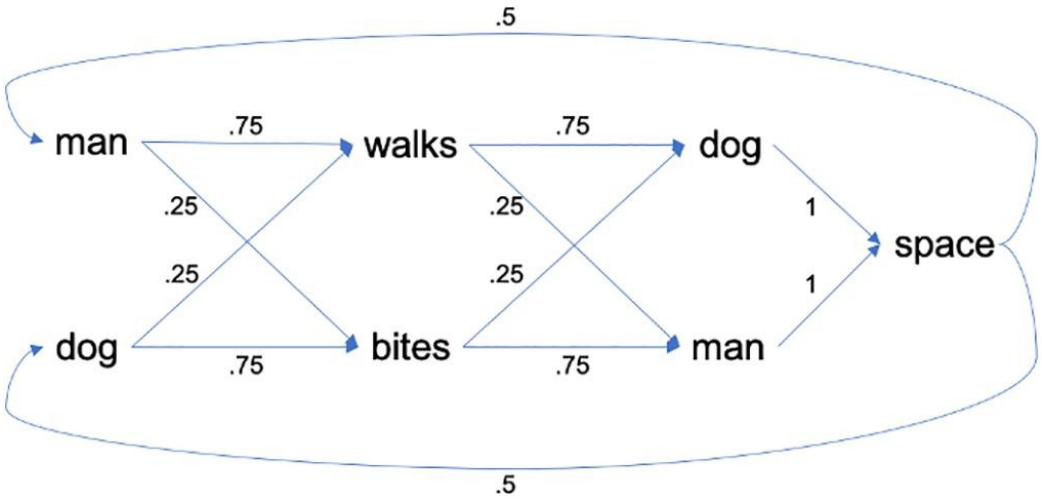


Figure 2.6: Transition matrix used to probabilistically generate input sequences to the network, of the form [subject, verb, object, space]. Numbers adjacent to arrows indicate the probability of that particular transition.

Table 2.1: The set of 8 possible sentences that could be generated by the transition matrix and the probability of occurrence. Note that each sequence was always followed by a "space" input.

sequence	probability
man walks dog	.28125
man walks man	.09375
man bites dog	.03125
man bites man	.09375
dog walks dog	.09375
dog walks man	.03125
dog bites dog	.09275
dog bites man	.28125

Active input weights are set to a value of +5, and inactive weights are set to 0. This means that nodes received an external perturbation of magnitude +5 if they were connected to the current input pattern, and no external perturbation if they were not connected to the input pattern. Thus, each input node, when active (only one was active at a time), produced a perturbation of +5 to approximately 10 randomly selected reservoir nodes, and a perturbation of 0 to all other reservoir nodes.

2.5.2 Outcomes

Because of the stochastic assignment of input patterns, separate runs of our model can produce distinct activity patterns. For present purposes, we are interested primarily

in the "fading memory" property of this model, which can be seen extended for 3 or more iterations on the majority of runs. Interested readers may run our code themselves in order to observe other possible outcomes.

However, to clearly illustrate the properties of this model, it will be helpful to first analyze the outcome of a single, representative run of the model in detail. Figure 2.7 shows in red/yellow columns the spike pattern on the last 3 of 1000 training sentences of four inputs each ([subject, verb, object, space]) on one run (i.e. there were 4000 total training iterations). Then, the model was run for an additional 4 timesteps of testing (blue/yellow columns of Fig. 2.7), in which external input was provided for only the first timestep (which was a subject input, in this case 'man'), and the model propagated its own activity for the final 3 timesteps with no further input (columns labeled 'NA' in Fig. 2.7).

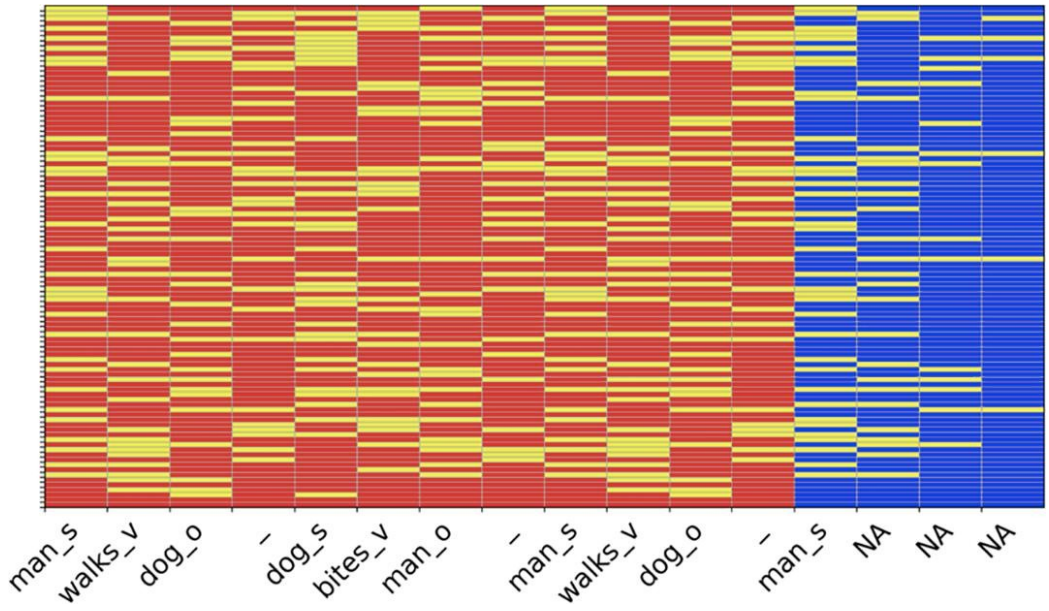


Figure 2.7: The spike pattern for the last 3 sentences (12 iterations) of training (red/yellow columns) on a representative run of the model, plus four test iterations of testing during which external input was cut off after the subject noun (blue/yellow columns). Nodes are arranged on the y-axis, and time is on the x-axis. Active nodes are shown in yellow, with inactive nodes in red/blue. For the last three iterations, 'NA' is shown in the column labels to indicate that no input was provided.

If we visually inspect any two columns of Figure 2.7 in which the same input was given, we can notice similar spike patterns recurring over time, which were only weakly present in the early iterations of the model, if at all. This can be seen even more clearly in the autocorrelation matrix from the same spike pattern over time (Figure 2.8). In Figure 2.8, we can see that the highly correlated spike patterns occur in response to different instances of the same input in the same position of a sentence, which could be read as emergent population codes for each input. We found that

these population codes were quite consistent over at least the final 100 timesteps of the model (autocorrelations were between .65- .77).

The crucial behavior, for our purposes, occurs when the external input is turned off in the final three iterations, seen in the blue-and-yellow columns of Figure 2.7, and in the 'NA' rows of Figure 2.8 (demarcated by dotted lines at the bottom). Despite having received no input, the network itself generates a spike pattern that resembles a slightly-degraded version of the population codes corresponding to the inputs that could have appeared in those positions. For example, the first iteration without input on this run, based on the transition matrix, would most likely have been "walks" with P = .75, or "bites" with P = .25. The pattern that emerges in the absence of input (the second blue column in Figure 2.7 and the first NA row in Figure 2.8) is highly correlated with the previous instance on which "walks" was actually presented (r = .60-.66), and the next most highly correlated pattern coincided with previous instances where the verb was "bites" (r = .34). This effect continues, though fading slightly, for the following 2 test iterations, which would have most likely been "dog" followed by a "space" input. In the absence of input, the network generates a pattern that highly resembles the most-likely completion of the terminated input sequence, and slightly resembles other less-likely sequences.

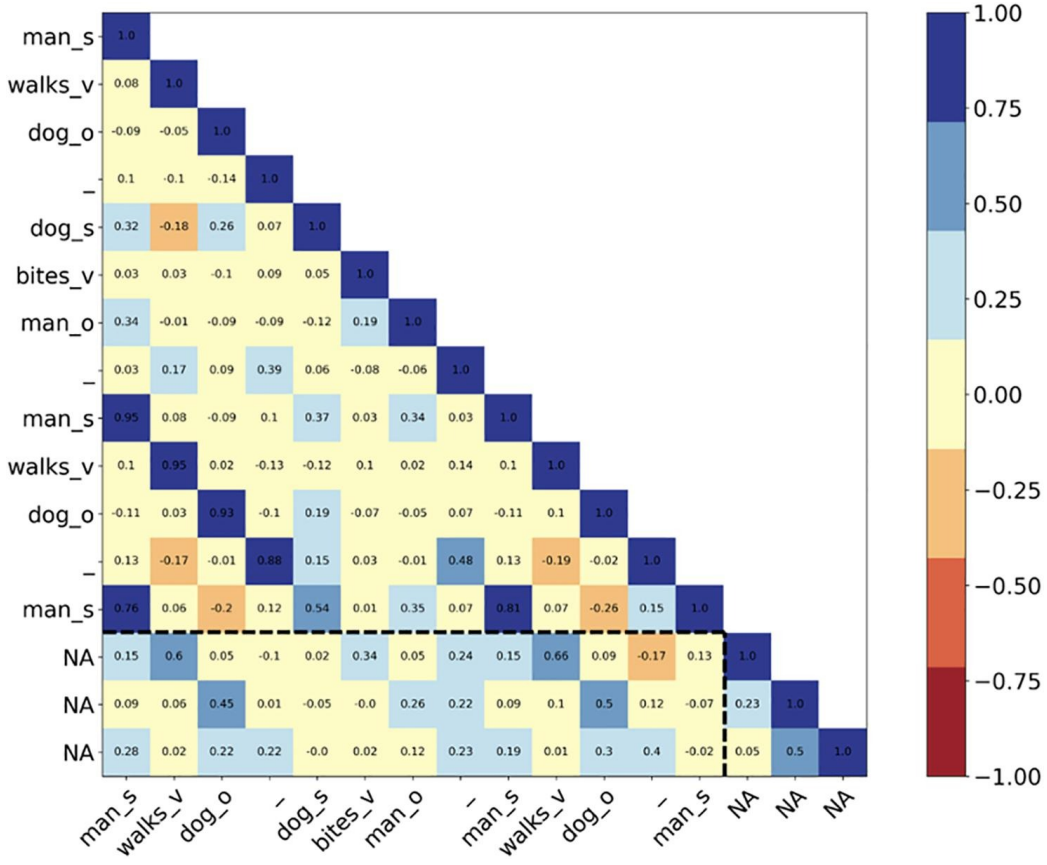


Figure 2.8: The autocorrelation matrix for the last 12 iterations of training, plus four test iterations during which only a subject noun was provided as input. Before the input turns off (above/left of the dotted lines), we can see that the reservoir has developed population codes—highly correlated spike patterns for separate instances of the same input. When the input is turned off (below/right of the dotted lines), we can see that the fading memory produces spike patterns that remain highly correlated with the likely next inputs, as if the network generates the pattern it "expects."

2.5.2.1 Fading Memory "Predictions"

To evaluate the consistency of this effect, we conducted 500 independent runs of the model. At the end of each run, we saved the final state of the model, then conducted 6 tests during which one or two subsequent iterations of input were provided (e.g. ['man', no input] or ['man', 'walks', no input]). We then computed the average correlation between the spike pattern that appeared after input was cut off, with the previous spike patterns that coincided with each input/position combination. If population codes emerge in this homeostatic reservoir network, and if the network does pattern-completion through time, we would expect the spike patterns on the final timesteps where no input was given to nonetheless resemble previous spike patterns corresponding to the inputs that could have appeared in those positions. For this analysis, we considered only the last 100 training sentences (400 iterations), to account for the fact that, while there appear to be emergent population codes that are relatively stable for a period, these may continue to change over longer time spans. The results presented in Table 2.2 represent the grand average of these correlations over all 500 runs.

As Table 2.2 shows, the "fading memory" output of our model tends to resemble the population codes for the input that would have most likely appeared in the next position, if the input had not been cut off. For example, given input of 'man' followed by no input, the activity of the network is most highly correlated with previous instances where 'walks' appeared (which followed 'man' 75% of the time). The next most similar population code was 'bites', which appeared 25% of the time following 'man' during training.

Interestingly, we can see that the model's fading memory output tends to be specific to the position at which the input is received. For example, given ['man', 'walks'], the output is most highly correlated with instances where 'dog' appeared in the object position (and, slightly less so, with instances where 'man' appeared in the object position), but has a much lower correlation with the activity observed for instances where 'dog' appeared in the subject position. Importantly, the external perturbation was the same in these two cases. This again points to the fact that our model became sensitive to the temporal structure in inputs, rather than simply the input patterns themselves.

Furthermore, we can notice that the fading memory output appears sensitive to the overall frequency with which different sentences could occur. For example, given ['man, walks'] or ['dog, walks'], an input of 'dog' appeared in the next position 50% of the time. However, the fading memory patterns are slightly different in these two cases: the output is more highly correlated with previous instances of 'dog' when the network was probed with ['man, walks'] (r = .452) than when it was probed with ['dog, walks'] (r = .402). In this way, the network appears sensitive to the fact that the overall probability of observing the sequence ['man', 'walks', 'dog'] (P = .28) is higher than the probability of observing ['dog', 'walks', 'dog'] (P = .09).

Table 2.2: Fading memory "predictions": the average correlation coefficient b/w each output with prior instances of each input-at-a-position (end of training, 500 runs). When input is cut off after one or two iterations, the model's fading memory produces spike patterns that are most highly correlated (bold cells) with the population code corresponding to the most-likely next input (e.g. input of 'man' followed by no input results in a pattern resembling previous instances in which 'walks' was actually presented). The next most-highly-correlated input pattern corresponds to the next-most-likely pattern that would have appeared (e.g., the fading memory after input of 'man' is also correlated with 'bites').

Possible Inputs at Each position [subject, verb, object, space]							
Test Input Presented	'man'_{subj}	'dog'_{subj}	'$walks'_{verb}$	'$bites'_{verb}$	'dog'_{obj}	'man'_{obj}	space
$[man]$	0.099	0.016	0.467	0.319	-0.009	0.065	0.173
$[man, walks]$	0.06	0.061	0.17	.0004	0.452	0.213	0.036
$[man, bites]$	0.049	0.052	-0.008	0.199	0.237	0.402	0.038
$[dog]$	0.02	0.107	0.3173	0.469	0.069	-0.01	0.176
$[dog, walks]$	0.057	0.0546	0.205	-0.003	0.402	0.237	0.04
$[dog, bites]$	0.06	0.06	0.002	0.171	0.213	0.452	0.039

This pattern completion behavior emerges because the model comes to depend upon its own self-generated activity to keep nodes poised near their target activation value. In fact, the external perturbations ultimately account for very little of the total activation pattern at each time step, most of which is driven endogenously. As such, when the external input shuts off, many of the nodes that constitute the population code for what would have been the next external input will derive their incoming activation mostly or entirely from the previous spike pattern. Because there is repeated structure in time, learning to stabilize present states based on the past is functionally equivalent to predicting the future. In other words, by adapting to past inputs, our nodes are learning what to do next time that pattern occurs in order to remain close to their target activation value. Needless to say, adaptation would not look like a

form of prediction if the past conditions that spurred adaptation did not recur in the future. Thus, our model is not so much predicting external input as it is entraining to the temporal structure of external input.

2.5.2.2 Hallmarks of Predictive Coding

In addition to generating prediction-like effects in its fading memory, our model also exhibits a signature of predictive coding that we also saw in TRACE: decreased activation for predictable inputs relative to unlikely inputs. We take the average activation of nodes in our network as a reasonable analog to an EEG signal from a real brain. We tested for a this signature of predictive coding in two ways. First, Figure 2.9 shows the mean activation of the network, obtained from the final 400 timesteps (100 sentences) of training for 500 distinct runs of the model, as a function of the transitional probability of the sequence (x-axis) and the item type (subject, verb, object, or space; colors in Fig. 2.9). This plot shows that mean activations were higher when low-probability transitions were observed. We can also see that activations were higher for verb inputs, relative to object inputs. This may be due to the fact that the Shannon entropy of subject nouns was 1 bit (each subject noun could appear with P = .5), while the entropy of verbs was .81 bits (verbs appeared with P = .25 or .75). This means that verbs were preceded by more surprising inputs, in an information theoretical sense, than were object nouns. If activation is higher for surprising inputs, and this activation may accumulate over subsequent timesteps, then it is reasonable to expect activations to be lower, on average, at the object position of a sentence relative to the verb position.

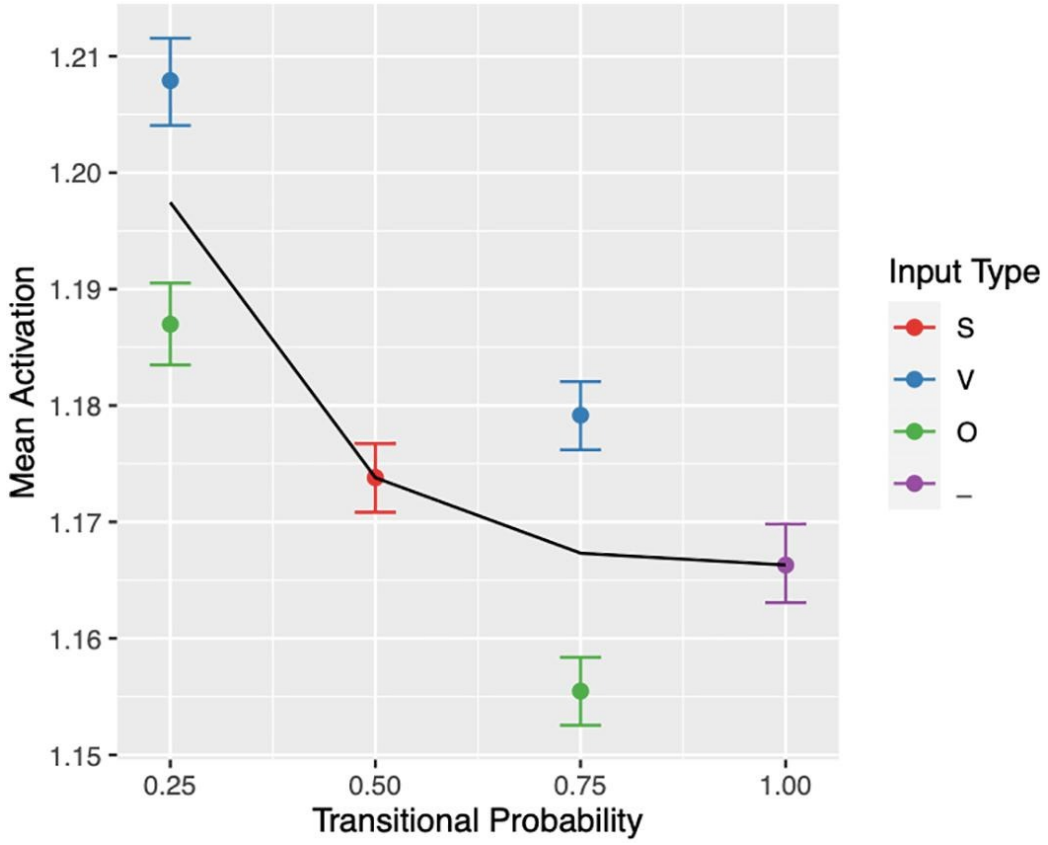


Figure 2.9: The average activation value of nodes as a function of the transitional probability of the observed input pattern. Colors correspond to possible sentence positions (subject, verb, object, or space). Each subject input appeared with .5 probability, and a space (\") input always appeared in the fourth position of a sequence. Object and verb inputs had associated transitional probabilities of either .25 or .75. The black line shows the mean value, averaging across input types. Error bars represent a 95% confidence interval around means obtained from the final 400 timesteps of training across 500 distinct runs of the model.

We next considered what may happen when the network is presented with an "ungrammatical" sequence—a sequence of items that was never encountered in training. During training, the network always observed sequences of the form [subject, verb, object, space]. We conducted a test using these same elements in a new order, [subject, object, verb, space]. In this new order, the first and fourth items are consistent with the sequences observed in training, while the second and third inputs were never observed in those positions during training. In Figure 2.10, we have plotted the average activation of the network during the final 100 timesteps of training (25 sentences of four elements each; left of the dashed line), plus the average activation of the network given each of 8 possible ungrammatical sequences. These values were averaged over 500 distinct runs of the model. The mean activation of the network for the final 100 timesteps of training was 1.17 (SD = .009). As Figure 2.10 reveals, activation increases substantially in response to the first ungrammatical input, increases further

with the second ungrammatical input, and drops at the conclusion of the sequence with a space input, which was consistent with the patterns observed in training. The mean activation values in response to the ungrammatical inputs are greater than 1.2, more than 3 standard deviations above the average activation in the final timesteps of training. This shows that the network's response to input sequences with a cloze probability of 0 are similar to the ones observed with low cloze probability (e.g. when transitional probabilities were .25, Fig. 11).

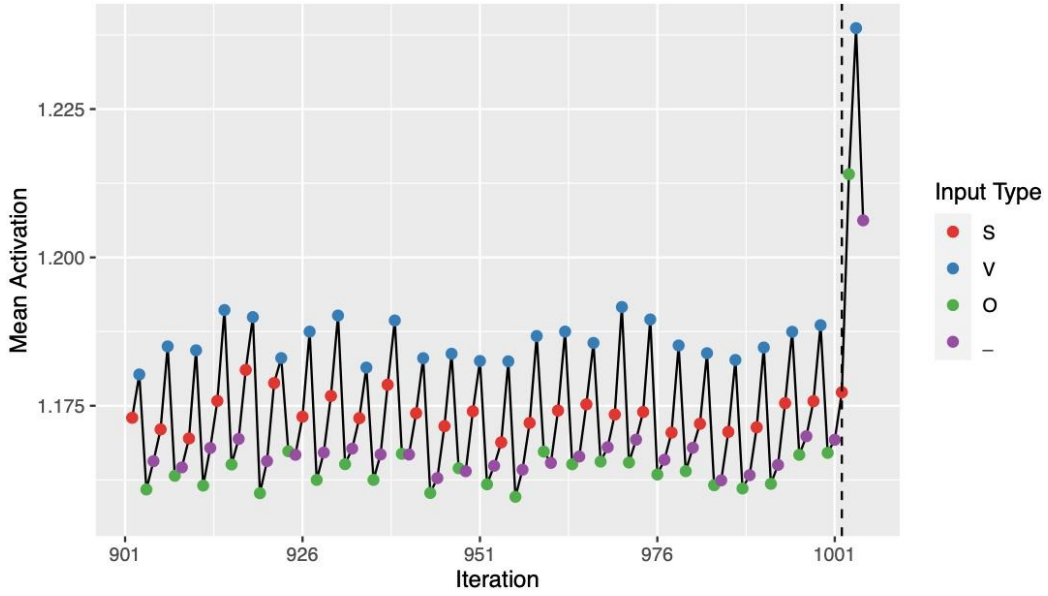


Figure 2.10: The average activation value of nodes as a function of iteration, for the final 100 iterations of training (left of the dashed line), plus a final sequence (four iterations) of testing with a sequence order that never appeared in training (subject, object, verb, space). Colors correspond to the item type. Points to the left of the dashed line represent averages over 500 distinct runs of the model. Points to the right of the dashed line represent averages over 8 possible ungrammatical sequences presented to 500 distinct runs.

2.5.3 Discussion

The simplified reservoir computing model presented here has no externally-imposed target activation pattern (in the form of a teaching signal), and yet distributed "population codes" naturally emerge. Through a learning algorithm by which individual nodes pursue a constant level of activity near an arbitrary target, the population code present at any given time point becomes critical for generating the full population code at the next time step, only a small subset of which may be partially attributable to external perturbation. As such, the "fading memory" of our reservoir tends to continue the established sequential pattern for a few time steps in the absence of

any input. We also observe one of the ostensible signatures of predictive coding—an decrease in signal when input sequences are more predictable—despite the fact that our network does not generate predictions.

While predictive coding models of neural computation take neural activity (largely, spiking behavior) to encode prediction errors (e.g. Fiorillo, Kim, & Hong, 2014), our learning algorithm treats spiking as a viable way to maintain homeostasis, so spikes in our model do not necessarily encode errors. Nonetheless, we find the ostensible signature of predictive coding in our model, with greater activity in the network in response to lower-probability or 0-probability (ungrammatical) sequences. Given that this pattern can emerge in systems that do not make use of predictions nor prediction errors, such as TRACE and our reservoir model, and given that it may sometimes not emerge in systems that are explicitly trained to predict, such as an SRN (Luthra et al., 2021), caution may be warranted before interpreting similar patterns in human data as support for a "realist" stance on predictive coding in the brain (see Colombo & Seriès, 2012).

If we didn't know any better, it would be tempting to say that this network has learned to "predict" upcoming inputs. However, while we can describe this form of adaptation as functionally predictive, it is important to note that what the system is doing is simply pattern completion, where the network learns to generate patterns that contribute to local homeostasis in the midst of variable external perturbations. As such, the activity of the network is not literally an encoding of predictions or prediction errors. Nor is the network simply encoding inputs, since much of the activity is endogenously driven. Rather than exhibiting "weak anticipation" by generating predictions of future input that are labeled as such, this homeostatic reservoir network does "strong anticipation" (Dubois, 2003; Stepp & Turvey, 2010) by coupling with its environment instead of representationally modeling it. This illustrates the fact that systems can be implicitly anticipatory, without forming explicit predictions.

2.6 Conclusion

In this chapter, I have presented a reservoir computer model with simple, allostatic nodes, that is capable of spontaneously generating object-tracking behavior and seemingly predictive behaviors to linguistic inputs. This model may serve as a potential bridge between representational "Cognitivist" theories on cognition, and non-representational "Ecological" theories, by providing a role for internal brain activity that nonetheless does not need to be understood as an encoding. Bickhard, a prominent critic of the Cognitivist position, emphasizes that encodings certainly do exist and admits a role for them in cognition, but simply suggests that they cannot be the end of the story, lest we end up with a world devoid of *meanings*—that is, activity that is grounded, or has normative value for an agent itself and not just an external observer. The present model may begin to illustrate precisely what the role

of encodings may be, and how they may be grounded. At the same time, reservoir computer models like ours and those of Kello and colleagues also demand a major reconceptualization of what representations are and how they function. Far from the stable, grandmother-cell story of old, the representations produced by these networks are transient, noisy, fuzzy, and context sensitive.

Chapter 3

A Continuum of Sensorimotor Grounding in the Comprehension of Literal and Metaphorical Sentences

3.1 Introduction

Some readers may be familiar with the classic children's stories of Amelia Bedelia, the excessively literal maid. One day, when her employer Mr. Rogers tells Amelia that it's time for them to "hit the road," Amelia picks up a stick from the yard and proceeds to wallop the street. How can we understand Amelia's unusual behavior from a language processing standpoint? As Amelia incrementally processes the speech input of Mr. Rogers, she first hears the verb "hit." Let us suppose this triggers a pattern of activity in Amelia's motor cortex—particularly the areas dedicated to the hands and/or arms—which is highly similar to the pattern of activity one would find if Amelia was actually performing the action of hitting something. Amelia has understood the symbol "hit"! But, upon hearing "...the road," Amelia fails to do what an individual without her dysfunction would do: suppress the first pattern of activity, and instead initiate a pattern that corresponds to the action of leaving, most likely in a vehicle. Amelia has, what we might call, a symbol un-grounding problem (or, perhaps a symbol re-grounding problem).

Since Searle first posed the "Chinese room" thought experiment in 1980, one of the paramount issues in cognitive science has been the so-called "symbol grounding problem" (Searle, 1980; see also: Harnad, 1990). Searle's thought experiment purported to demonstrate that an algorithmic instantiation of language processing, in which input symbols are transformed into output symbols based on a set of rules, has no

access to the meaning of those symbols, since the symbols have never been linked (i.e. grounded) to their referents. Over the past two decades, theories of "grounded cognition" (Barsalou et al., 1999) have attempted to overcome this problem through variations on the argument that the brain regions involved in processing the meaning of an utterance are the very same ones that are involved in sensorimotor experience with the referents of the utterance. For example, Barsalou's Perceptual Symbol Systems hypothesis (Barsalou, 2008; Barsalou et al., 1999) proposes that understanding the meaning of the word "cat" occurs through a partial re-activation of the brain areas that previously encoded the sight of a cat's pointy ears, the feeling of soft fur, the sound of a meow, and perhaps the smell of a litterbox. This is often described as the brain running an embodied "perceptual simulation" of a cat.

Abstract or metaphorical language represents a critical test case for the sensorimotor simulation hypothesis, as it is difficult to imagine how one might comprehend *abstract* language—with no clear sensorimotor component to the meaning—by virtue of reactivation of sensorimotor representations. While a large body of work has demonstrated evidence for sensorimotor simulation in response to concrete language, which may in some cases have a functional role in language processing, the evidence is much more mixed regarding abstract language. Here, we present two experiments designed to investigate whether abstract language processing results in behavioral patterns associated with sensorimotor activation, and whether this activation serves a functional role in comprehension, should it exist.

Experiment 1 consisted of a concrete/abstract judgment in which participants read abstract or literal sentences with hand- or foot-related action verbs (e.g. "punch" or "kick" respectively), and responded with either the hand or foot. If either type of sentence elicits the reactivation of associated sensorimotor regions, we expected to observe a congruency effect, with responses made with the hands being faster for sentences containing hand-related verbs, and vice versa for responses with the foot. A second goal of this study was to examine the extent to which such congruency effects are moderated by several common psycholingusitic variables, including the frequency, familiarity, and figurativeness of sentences, in additional to the contextual diversity of the critical words.

Experiment 2 deployed a "Visual-World" paradigm eye-tracking design and a manual interference task to investigate the extent to which sensorimotor activity played a functional role in language comprehension. Participants listened to abstract or literal sentences with hand-related verbs, and made an abstract/literal judgment using their foot. For half of the trials, participants performed a concurrent simple motion with the hands, intended to generate interference with activity in hand-regions of motor cortex. If activity in these regions is functionally relevant for the comprehension of language, we predicted that the hand motion would interfere with the processing of these sentences. This could be observed as a change in the fixations to the task-relevant 'Abstract' and 'Literal' response boxes. If the comprehension of both abstract and literal language rely on sensorimotor regions, we predicted no change in the

relative proportion of fixations to each response option while performing the hand motion. If, on the other hand, literal but *not* abstract language processing relies on sensorimotor regions, we predicted that the hand motion would selectively boost the proportion of fixations to the 'Abstract' response box. That is, participants may momentarily become more inclined towards an abstract interpretation of a sentence containing a hand-related action verb while they are moving their hands.

3.1.1 Review of the Literature

By now, a staggering body of research has been dedicated to the topic of sensorimotor grounding in language. We will merely review a representative sample of it here (but for detailed reviews, see: Fischer & Zwaan, 2008; Hauk & Tschentscher, 2013; Kiefer & Pulvermüller, 2012; Meteyard, Cuadrado, Bahrami, & Vigliocco, 2012). Some of the first clues hinting at sensorimotor grounding in language comprehension came from neuroimaging studies revealing somatotopic activation of the motor cortex in response to action words in either a visual (Hauk, Johnsrude, & Pulvermüller, 2004) or auditory (Buccino et al., 2005) format. Similarly, Tschentscher, Hauk, Fischer, and Pulvermüller (2012) found somatotopic activation of hand regions during a counting task when no overt hand movements were taking place, suggesting that participants were mentally simulating counting on their fingers. Willems, Hagoort, and Casasanto (2010) extended these findings to show body-specificity in somatotopic activation: Right-handers showed activation in left premotor cortex (since motor control is organized contralaterally in the brain) during lexical decisions to manual-action verbs, while left-handers showed activation in right premotor cortex. These results have been interpreted as evidence that language processing can result in detailed mental simulations of performing actions consistent with the ones described in verbal stimuli.

The simulation hypothesis has been supported by a number of behavioral studies as well. Early results that were consistent with this view came from Tucker and Ellis (1998) showing that when people judged an image of an object with a handle as being upright or upside-down, reaction times showed an interaction between the location of its handle interacted with the hand that they were using to respond. For instance, when the handle happened to be on the left side of the object, and the response they were making used the left hand, reaction times for judging the orientation of the object were faster. Tucker and Ellis interpreted this finding as evidence for activation of a motor affordance potentiating action of the hand that was closest to the handle, even when grasping the object was impossible. Similar results have been found for verbal descriptions of those objects and their handles (D. C. Richardson, Spivey, & Cheung, 2001, , experiment 2). When Stanfield and Zwaan (2001) presented sentences like "John hammered the nail into the wall," they found that participants were faster to identify an image of a horizontally oriented nail. With "John hammered the nail into the floor," they were faster to identify an image of a vertically oriented nail. D. Richardson, Spivey, Barsalou, and McRae (2003) found that processing action

verbs whose image schemas have a horizontal or vertical axis interferes with a visual discrimination task in the congruent region of visual space. Andres, Finocchiaro, Buiatti, and Piazza (2015) reported an action-sentence compatibility (ACE)[1] effect for responses to hand-/foot-related verbs when responding with the hands or the feet, while Ahlberg, Dudschig, and Kaup (2013) found a corresponding effect for responses to hand-/foot-related nouns (but, interestingly, found no effect for verbs). This effect has also been found in full-sentence processing, rather than just in single words. For example, Scorolli and Borghi (2007) reported an ACE effect for verbal or manual responses to sentences containing mouth- or hand-related verbs.

While the above results, and many others, point to the existence of mental sensorimotor simulations, a key question concerns the possible functional role of such simulations. Mahon and colleagues (Mahon, 2015; Mahon & Caramazza, 2008) have raised the possibility that the proposed "simulations" are attributable to mere spreading activation and/or mental imagery formed after comprehension is complete, and therefore are not functionally relevant to language comprehension per se. This concern has been addressed in a number of studies. The results of a recent study by Strozyk, Dudschig, and Kaup (2019) offer some support for Mahon's position: those authors found a consistent ACE effect, such that hand responses were consistently faster for hand words than foot words (and vice versa), but also found that a simultaneous hand or foot tapping task did not inhibit responses to hand and foot words, respectively. The authors took this as evidence that participants mentally simulated the meanings of presented words, which primed congruent responses, but that this occurred in a post-processing stage that was not functionally involved in the comprehension of those words. However, a number of other studies offer evidence to the contrary. Pulvermüller (2005) found that brief TMS to hand or leg regions selectively sped participants' responses to hand or leg verbs. Shebani and Pulvermüller (2013) found that a simultaneous hand or foot tapping task interfered with memory for hand/arm or foot/leg verbs, respectively, indicating that motor cortex participates in word encoding. Yee, Chrysikou, Hoffman, and Thompson-Schill (2013) expanded upon these findings by showing that the degree of manual interference found when processing hand-related nouns is modulated by the degree of experience manipulating the referents of those nouns. Willems, Labruna, D'Esposito, Ivry, and Casasanto (2011) showed that TMS stimulation to left motor cortex sped participants' right-hand responses to manual (but not non-manual) action verbs. Similarly, Gijssels, Ivry, and Casasanto (2018) reported complementary effects of inhibitory or excitatory stimulation to premotor cortex on reaction times to manual verbs. This sampling of results clearly demonstrates that the sensorimotor representations can, at least in

[1]Note the ACE was originally reported by Glenberg and Kaschak (2002) in the context of compatibility between the direction of transfer implied by a sentence and the direction of responding. However, a recent multi-lab replication attempt has overturned these earlier results, finding that it did not replicate in any of 18 labs (Morey et al., 2021). On the other hand, a recent meta-analyses from A. Winter, Dudschig, and Kaup (2022) reported that the ACE was reliable, though small. Work involving compatibility effects between response effectors (e.g. responding with the hand or foot) and bodily-related verbs is not necessarily undermined by these results.

some contexts, have a functional role in language comprehension or memory.

It is important to emphasize that we should not now conclude that all of language processing or conceptual access recruits sensorimotor information in the same way (or at all), though grounded cognition views have at times been caricatured as arguing exactly this. Quite to the contrary, grounded cognition views have always emphasized the situatedness of cognition (Barsalou, 2016), meaning that cognition is a dynamic product of brain, body, and broader context (linguistic context included). It is precisely this context that drives the recruitment of sensorimotor information on an as-needed basis. It is now widely agreed that conceptual access occurs within a highly distributed brain network (Pulvermüller, 2005), in which there are hubs (i.e. "convergence zones"; damasio1989a) that integrate information across modalities and that mediate responses to stimuli in complex ways (Barsalou, 2015). It is perhaps immaterial whether these integrative hubs in the network are referred to as "amodal" (Mahon, 2015) or "multimodal" (Barsalou, 2016). The important question now is not *whether* language comprehension involves sensorimotor representations, but instead how context determines when sensorimotor representations will or will not be recruited for linguistic processing, and precisely how such effects play out over the time course of language processing.

3.1.1.1 Sensorimotor Grounding in Abstract/Metaphorical Language Processing

Abstract and metaphorical language present a particularly useful test case for examining this issue. Intuitively, it seems difficult to explain how one could understand an abstract concept such as "truth" or "justice" by recourse to sensorimotor simulation. By definition, the meanings of abstract concepts are spread across contexts that are highly diverse in their experiential content. Similarly, metaphorical language such as "hit the road" actually has relatively little to do with any literal actions implied by the verb "hit," and therefore a sensorimotor simulation of the verb in this context might seem to be counterproductive.

Nonetheless, there have been some attempts at explaining abstract and metaphorical language in terms of sensorimotor grounding. For example, Barsalou et al. (1999) offered an account of understanding the words "true" and "false" whereby individuals compare sensorimotor simulations to the perceived world and assess the match. However, this explanation is clearly limited to situations in which sensory evidence is available for comparisons. In a political discussion about the value of justice, for example, there may be no such information available. To overcome this issue, Lakoff and Johnson (2008) and others (e.g. gibbs2002a) have argued for the ubiquity of sensorimotor metaphor usage to describe or understand abstract ideas, which has been referred to as Conceptual Metaphor Theory. On this view, abstract concepts such as "love" are better understood as umbrellas for a number of different sensorimotor metaphors (e.g. "love is a journey" or "love is a battlefield"). Which conceptual

metaphor is activated, they argue, depends upon context, but they maintain that such abstract ideas are nonetheless frequently reduced to ideas grounded in sensorimotor experience.

However, the neuroimaging and behavioral data regarding the conceptual metaphor view (Lakoff & Johnson, 2008) are quite mixed (for a review, see: Dove, 2016). Some neuroimaging studies have found evidence of a role for the motor system in abstract and metaphorical language. For example, an MEG and an fMRI study by Boulenger, Hauk, and Pulvermüller (2009) and Boulenger, Shtyrov, and Pulvermüller (2012), respectively, reported somatotopic activation in motor cortex in response to literal as well as idiomatic/metaphorical sentences involving action verbs (e.g. "grasp the ball" vs. "grasp the idea"), as well as greater modulation of fronto-temporal regions for metaphorical sentences relative to literal ones. Consistent results have been obtained by Lacey et al. (2017) and Lauro, Mattavelli, Papagno, and Tettamanti (2013), in addition to comparable patterns for sentences containing tactile (Lacey, Stilla, & Sathian, 2012) and gustatory metaphors (Citron & Goldberg, 2014). Similarly, some behavioral studies have shown effects consistent with sensorimotor grounding of abstract and metaphorical language. For example, D. Richardson et al. (2003) found ACE-like response time effects for abstract verbs as well as literal ones. Santana and Vega (2011) found the same patterns for metaphorical uses of action verbs, in addition to literal uses of action verbs and abstract verbs. With fictive motion sentences, where an action verb is used figuratively, as in "The road runs through the valley," Matlock (2004) showed that readers were slower when a context sentence made the path seem more difficult to travel—even though the sentence does not depict any literal travel. Similar results were found in eye-movement patterns while participants heard those sentences and viewed a corresponding scene (D. Richardson & Matlock, 2007).

On the other hand, several studies have reported finding such somatotopic activation for literal sentences, but not for their metaphorical counterparts. For example, an fMRI study by Quadflieg et al. (2011) found that a classifier trained on neural data from a visual height discrimination task could also discriminate between neural responses to verbal items literally describing height, but not between responses to items about power or valence, which have been previously shown to activate height-related spatial metaphors (Schubert, 2005). Aziz-Zadeh, Wilson, Rizzolatti, and Iacoboni (2006) found that the same brain areas were active when watching videos of actions as when reading literal, but not metaphorical, sentences involving the same action verbs. Similar patterns of results have been reported by Raposo, Moss, Stamatakis, and Tyler (2009), Desai, Conant, Binder, Park, and Seidenberg (2013), and Rüschemeyer, Brass, and Friederici (2007). In a behavioral study similar to that of D. Richardson et al. (2003), Bergen, Lindsay, Matlock, and Narayanan (2007) failed to replicate the findings with abstract verbs, reporting that abstract verbs did not elicit spatially grounded representations that interfered with visual processing.

In a critical review of the literature on embodied metaphor, Casasanto and Gijssels

(2015) have argued that, while conceptual metaphors may be ubiquitous, there is actually little evidence to indicate that such metaphors result in functionally-relevant input from sensorimotor systems. ACE-like semantic congruency effects, they argue, have been established in cognitive science for decades and explained satisfactorily without recourse to embodied simulations. As such, congruency effects observed for abstract or metaphorical language may be of little use in distinguishing between grounded and amodal/symbolic hypothesis. Furthermore, the authors argue that many of the neuroimaging studies that have found evidence consistent with embodied grounding of abstract and/or metaphorical language suffer from statistical and methodological issues that call into question their validity. So, does all of this mean that we are safe concluding that literal language is grounded in the body, but abstract and metaphorical language is not?

On the contrary, we suggest that the inconsistent results found above may be the result of the highly context-dependent, interactive and dynamic nature of language and cognition (for reviews, see S. Anderson & Spivey, 2009; J. Falandays, Batzloff, Spevack, & Spivey, 2018; S. Spevack, Falandays, Batzloff, & Spivey, 2018; Willems et al., 2011). Ambiguity in language is ubiquitous at multiple levels of representation (J. Falandays, Brown-Schmidt, & Toscano, 2020), and in the broader field of language processing, it is now well established that even literal language is highly sensitive to context (e.g. tabossi1988a). As such, sensorimotor grounding in any type of language may be better thought of as a matter of degree, dependent upon the linguistic and environmental context, rather than a binary choice (Chatterjee, 2010; Zwaan, 2014). For example, Bergen and Wheeler (2010) found that when the sentence uses the imperfective aspect, as in the present progressive sentences "Richard is beating the drum" and "Richard is beating his chest," the action-sentence compatibility effect is especially robust. However, a simple past tense sentence like "Richard beat the drum" is not as effective at inducing the ACE. They suggest that the imperfective aspect in "Richard is beating the drum" emphasizes an ongoingness of the event described by the sentence and thus engages the motor system more so. By contrast, the simple past tense emphasizes the completion of the event, and so the motor cortex may be less involved (see also: Huette & Anderson, 2012). Aravena et al. (2014) measured the grip force of participants as they listened to literal sentences containing hand-related action verbs. These authors found that participants increased their grip force shortly after processing the hand verb when the verb was the focus of the sentence (e.g. "John signs the contract"), but not when the focus was shifted to the agent's mental state (e.g. "John wants to sign the contract"). Furthermore, when the sentence context led to a strong prediction of an upcoming hand-related action verb, grip force increased even when the action verb was replaced by a pseudo-word. Similarly, grounded effects seem to disappear in contexts where semantic access is unnecessary. For example, a study by Tomasino, Fink, Sparing, Dafotakis, and Weiss (2008) found a faciliatory effect of TMS to motor cortex when participants responded to hand verbs in a motor imagery task, but not in a silent reading or frequency judgment task, while Mirabella, Iaconelli, Spadacenta, Federico, and Gallese (2012) found an ACE effect in a semantic judgment task but not in a Stroop color judgment task. Results like these suggest

that, in general, context may be just as important in shaping the recruitment of sensorimotor information as the actual words used.

In line with this view, much of the work on sensorimotor grounding of abstract and metaphorical language has revealed an important role for context. For example, an fMRI study by Desai, Binder, Conant, Mano, and Seidenberg (2011) found that the neural response to abstract and metaphorical sentences in motor cortex was inversely correlated with their familiarity. These authors concluded that abstract/metaphoric language goes through a "gradual abstraction process" whereby sensorimotor systems are recruited to a lesser degree as familiarity with the phrases increases. Lauro et al. (2013) and a meta-analysis by Yang and Shu (2016) reported motor activity related to literal, abstract, and metaphorical language, but not idiomatic language. Because idiomatic language is highly familiar and conventionalized, these results fit with Desai et al.'s (2011) proposal. Further supporting a distinction between novel and conventionalized metaphorical language, Pobric, Mashal, Faust, and Lavidor (2008) found that TMS to right hemisphere selectively impaired processing of novel metaphors, while TMS to left hemisphere selectively impaired processing of literal phrases and conventionalized metaphors.

However, even highly conventionalized, idiomatic language can appear to be grounded in the motor system, under the right conditions. Gibbs (1993), Lakoff (1987), and Müller (2009) have all argued that many seemingly "dead" metaphors (e.g. those for which the original metaphorical relationship is now opaque), in fact retain strong semantic connections to their source domains. Supporting this view, an ERP study by Goldstein, Arzouan, and Faust (2012) presented participants with novel or conventionalized metaphors, for which they were required to explain the meanings. They found that the waveforms in response to novel metaphors, after being explained, became more similar to unexplained conventionalized metaphors. Meanwhile, explained conventionalized metaphors became more similar to unexplained novel metaphors. The authors took these results to show that individuals can rapidly modulate their processing of metaphorical phrases, such that novel metaphors can become conventionalized, thereby decreasing the recruitment of featural information, while "dead" metaphors can be revived, increasing their recruitment of featural information.

An alternative hypothesis is that abstract and metaphorical language does not become disembodied as familiarity increases, but rather differently embodied. The words-as-social-tools hypothesis from Borghi et al. (2019) proposes that abstract language acquires meaning through social interaction, thereby becoming grounded in interoceptive, metacognitive, and speech-related brain areas. While those authors present a convincing argument that abstract language comprehension has a different developmental trajectory than concrete language, we have pointed out that this conclusion also warrants consideration that abstract language may be more dynamic and context-dependent than concrete language over shorter timescales as well (J. Falandays & Spivey, 2019). Furthermore, Gibbs (1993) argued that idiomatic meanings of phrases may be activated in parallel to literal meanings, suggesting that simply

asking whether or where concepts are grounded is likely to be the wrong question entirely. Instead, language comprehension may recruit several different modal and amodal/multimodal regions simultaneously, with the balance of activation shifting according to context.

Several theoretical perspectives have now been put forth that frame sensorimotor grounding effects within a highly parallel, interactive, and dynamic framework (J. Falandays et al., 2018). For example, the language-as-situated-simulation hypothesis (Barsalou, 2008) holds that language comprehension involves the interaction of multiple subsystems that integrate sensorimotor experience and the distributional statistics of language (see also, andrews2009a, louwerse2012a. Similarly, Kemmerer (2015) proposed that conceptual knowledge is stored in a dynamically flexible, hierarchical network whereby recruitment of sensorimotor systems depends upon the task and linguistic/environmental context. We are in agreement with these perspectives, and in this paper, take up the task of exploring the contextual variability and temporal dynamics of sensorimotor grounding in more detail.

3.2 Overview of the Present Study

The purpose of the present study is to clarify the origins of inconsistent patterns of results related to the sensorimotor grounding of literal and metaphorical language by exploring the degree to which linguistic and task context modulates these effects. We constructed a set of literal phrases containing hand-/arm- or foot-/leg- related action verbs, along with matched metaphorical sentences. These phrases varied on several psycholinguistic variables, including the frequency and contextual diversity of the verb and noun and the frequency and predictability of the full phrase. The verbs and object nouns of these sentences were rated for their relatedness to hands/arms and legs/feet. In experiment 1, we investigate the ways in which each of these factors contribute to the action-sentence compatibility effect in a literal/metaphorical judgment task.

In experiment 2, we investigate the time course of sensorimotor grounding effects in an eye tracking study with a simultaneous manual-interference task. The time course of grounding effects is a underexplored issue. This is partly due to the fact that most neuroimaging studies lack sufficient temporal resolution to draw conclusions about when activity in sensorimotor regions is seen (but cf. Boulenger et al., 2012)). Meanwhile, reaction time experiments typically get a measurement only once at the end of comprehension, making them equally unsuited for this purpose. However, several studies have shown that timing is of critical importance in the ACE effect and in the interpretation of its significance (Borreggine & Kaschak, 2006; Diefenbach, Rieger, Massen, & Prinz, 2013; D. C. Richardson et al., 2001; Vega, Moreno, & Castillo, 2013). For example, Borregine and Kaschak (2009) manipulated the time during sentence processing at which a motor response was cued, finding that the ACE effect emerges early, diminishes or disappears completely, and then re-emerges again.

The authors took this to show that the ACE effect is present when semantic features are activated, but suppressed while features are being bound into coherent wholes (see also papeo2009a). Thus, the time course of effects provides crucial clues as to the role of sensorimotor activity in comprehension (Hauk, 2016).

It is important to note that our design is unable to distinguish between the possibility that metaphorical language is disembodied, or whether it is differently embodied (*a la* borghi2019a). For example, failing to find an ACE effect or an effect of manual interference when processing hand-related metaphorical verbs may mean that little to no sensorimotor information is recruited, or that different sensorimotor information is recruited, such as mouth regions. However, finding substantial contextual variability in sensorimotor grounding would be sufficient to show that searching for a specific, individual locus of grounding is not a viable path forward. Rather, we would be forced to conclude that embodied grounding of language is a highly dynamic and context-dependent phenomenon.

3.3 Experiment 1

Experiment 1 leveraged the action-sentence compatibility effect to examine the presence of sensorimotor representations in processing literal and metaphorical phrases. We sought to determine which psycholinguistic variables contribute to grounding, such as the frequency of the phrase, the verb, and the noun. Participants listened to sentences containing an action verb related to either the hands/arms or the feet/legs and were tasked with indicating whether the sentence was literal or non-literal by responding with either a hand or a foot press. If sensorimotor systems are somatotopically activated while processing both literal and metaphorical statements, we predict an interaction effect between the body-association of the verb and the effector used for responding.

3.3.1 Method

3.3.1.0.1 Design Experiment one used a 2 (verb-body association: hands/arms or feet/legs) X 2 (response effector: hand or foot) X 2 (sentence type: literal or metaphorical) repeated-measures design. In one block of trials, the hand-press was assigned to a "literal" response and the foot-press was assigned to an "abstract" response. In the other block of trials, these response contingencies were flipped. Participants were pseudo-randomly assigned such that half the participants saw the response contingencies in one order, while the other half of participants saw them in the other order. Within each block, trial order was fully randomized for each participant.

3.3.1.0.2 Materials Our stimuli were inspired by those used by Boulenger et al. (2012), who reported two key bits of evidence: First, both literal and idiomatic phrases elicited activity in motor cortex. Second, idiomatic phrases elicited stronger activity in fronto-temporal regions. Their stimuli consisted of matched literal and idiomatic sentences containing hand/arm or foot/leg verbs, such as "Pablo kicked the statue" (literal) or "Pablo kicked the habit" (idiomatic). Boulenger et al. (2012) found that both motor activity and fronto-temporal differences emerged 150-250ms following the end of the sentence. This time course is well suited to a reaction-time design: if their neurological observations correspond to functionally-relevant activity, effects should be present at just the time participants are launching their responses.

We first constructed a list of hand-/arm- or foot-/leg-related verbs. All verbs were used in the past tense. These verbs were rated by 21 independent participants from the same population as the study. Participants rated the degree to which each verb was related to each of the two categories on slider bars ranging from 0 ("Not at all") to 100 ("Very much so"). We selected 15 maximally distinct verbs from each category. Linear regression analyses confirmed that hand/arm verbs were rated as more hand-/arm-related than foot/leg verbs (b = 72.838, SE = 1.516, t = 48.04, $p < .001$) and less foot-/leg-related than foot/leg verbs (b = -64.467, SE = 1.944, t = -33.16, $p < .001$). Since the same verbs are used in both literal and metaphorical sentences, it is not possible for those groups to differ in hand/arm or foot/leg associations.

The verb groups were compared on lexical frequency, contextual diversity, and length based on measurements from the SUBTLEX-US database (Brysbaert & New, 2009). The measurements in this database have been shown to conform more closely to contemporary human behavior than older, more classic databases such as Francis, Kucera, Kučera, and Mackie (1982) and CELEX (1996). We compared the log-transformed word-frequency and contextual-diversity score of the groups using linear regression, counting only each unique verb. This analysis showed that hand/arm verbs had significantly higher frequency scores than foot/leg verbs (b = .623, SE = .304, t = 2.05, $p < .05$), but the two groups did not differ significantly on length or on contextual diversity, the latter of which is arguably the more important psycholinguistic measurement for present purposes (Adelman, Brown, & Quesada, 2006; Plummer, Perea, & Rayner, 2014).

For each of these verbs, three matched pairs of literal and metaphorical sentences were constructed, which differed only in their final word. This produced four groups of sentences with 45 stimuli in each: literal hand/arm, metaphorical hand/arm, literal foot/leg, and metaphorical foot/leg. An analysis of the frequency, contextual diversity, and length of the final words revealed no significant differences between groups. 24 independent participants from the same population as the study rated the hand/arm and foot/leg association of each noun, using the same scale as for the verbs. Linear regression analyses showed that nouns used with hand/arm verbs were significantly less associated with the foot/leg than were nouns used with foot/leg verbs (b = -3.386, SE = 1.487, t = -2.276, $p < .05$), but the two groups did not differ signifi-

cantly in hand/arm associations. Both groups showed a main effect of usage category, such that nouns used in literal sentences were significantly less associated with the hand/arm than were nouns used in metaphorical sentences (b = -3.565, SE = 1.597, t = -2.232, p <.05), and significantly more associated with the foot/leg (b = 16.184, SE = 1.495, t = 10.823, p < .001). There were also interactions between usage category and verb-body association, such that the previous main effects were stronger for literal sentences than for metaphorical sentences (for hand-/arm-association ratings: b = 19.99, SE = 2.219, t = 9.007, p <.001; for foot-/leg-association ratings, b = -15.32, SE = 2.078, t = - 7.372, p < .001). These differences will be accounted for in the statistical analysis of the experimental data through the use of linear mixed-effects modeling.

23 independent participants from the same population as the study rated each phrase on its familiarity, interpretability, naturalness, figurativeness, and imageability using a 5-point Likert scale. Linear regression analyses showed no differences between groups on familiarity, interpretability, or naturalness. As would be expected, literal phrases were rated as significantly less figurative than abstract phrases (b = -1.154, SE = .073, t = -15.831, p < .001) and as significantly more imageable than abstract phrases (b = .257, SE = .086, t = 2.973, p < .01). The variables of familiarity, interpretability, naturalness, and imageability were highly correlated with each other (r > .9). Therefore, to deal with a potential multi-collinearity problem in the statistical analysis stage, we performed principal components analyses on these four variables, and obtained the first principal component, which accounted for 93.38% of the variance. This component was used in the analyses in place of the four separate measures, and for convenience will be referred to as the familiarity PC.

To estimate the frequency of each phrase, which is expected to relate to the degree of conventionalization of a phrase, we collected the number of Google results returned when searching the full phrase, without the proper name, in quotation marks. We refer to this as the phrasal frequency. Because the distribution of phrasal frequencies was exponential, this variable was log transformed. Phrasal frequency was significantly positively correlated with the human ratings for familiarity (r = .499, p < .05), interpretability (r =.428, p < .05), naturalness (r = .413, p < .05) and imageability (r = .358, p < .05). Linear regression analysis of phrasal frequency showed significant effects of usage (literal vs metaphorical; b = -.775, SE = .381, t = -2.034, p < .05) such that literal sentences were less frequent than metaphorical ones, and body association (hand/arm vs foot/leg; b = 1.14, SE = 3.81, t = 2.999, p < .01) such that hand/arm sentences were more frequent than foot/leg ones.

Next, we estimated the cloze probability of the final word in each sentence using the output from Open AI's recurrent neural network, GPT-2 (Radford et al., 2019), which is pretrained on a large corpus. To compute this measure, each sentence, beginning with the verb, was padded on the left with a start-of-text marker. The sentences were then fed to the network word by word, producing a prediction of the upcoming word based only on the full preceding context of each sentence. We then obtained the

logit values—the network's predictions for each word in its vocabulary—at the point of disambiguation. These values were fed through a softmax function to obtain the probability of each word in the vocabulary at that point in the sentence. We then obtained the probability corresponding to the actual disambiguating word. Because this distribution of values was also exponentially distributed, it was log transformed. Linear regression analysis showed no significant differences between groups on the cloze probability of the disambiguating word.

Finally, for the subject of each sentence, 18 traditionally male names and 18 traditionally female names were selected. This list of 36 names was spread across the list such that each of the 6 instances of a verb was paired with a different name. Each verb was also paired with only male or only female names. Each name was used 5 times across the stimuli.

3.3.1.0.3 Participants 103 participants from the University of Merced community were recruited using the Sona research participation system. Participants provided informed consent in accordance with IRB policies and were compensated for their time with course credits. Participation was restricted to those with normal or corrected-to-normal vision, who were fluent in English, right-handed, and reported not having dyslexia, other reading disabilities, or any physical problems that would prevent simple movements with the hands.

3.3.1.0.4 Apparatus The response-collection apparatus consisted of a Makey Makey circuit board. This device allows for easily routing touches to any electrically-conductive object onto keyboard buttons on a computer. The left keyboard button was connected to a small metal pad used for hand responses, while the right keyboard button was connected to a larger metal pad for foot response. Participants wore a metal ring on their left hand which was also connected to the Makey Makey. This forms an incomplete circuit between the participant, the Makey Makey, and the metal pads. When participants touched either metal pad, the circuit is completed and a response is registered. Participants were required to be barefoot in order to use the foot pad. Before beginning the experiment, the researcher explained how the response collection system worked and verified that the participant knew how to use it.

3.3.1.0.5 Procedure Participants completed the experiment individually in the lab in a single session lasting approximately 30 minutes. Before beginning the experimental task, participants first completed a Vividness of Visual Imagery (VVIQ) questionnaire (Marks, 1973). After completing the survey, participants were seated in from of an Apple Macintosh desktop computer, on which the software OpenSesame was used to present the experiment.

Before each block of the experiment, participants read instructions specifying the

current response contingencies (whether the hand-press corresponded with a "literal" or "abstract" response). Although the latter type of stimulus is more accurately understood as metaphorical language rather than "abstract" language (since some metaphors may contain only concrete words and refer to concrete situations, e.g. "hit the road"), piloting showed that our participants had an easier time understanding and responding to the distinction when it was posed as literal vs. abstract rather than literal vs. metaphorical. We also chose not to pose the distinction as literal vs. non-literal in order to make response options visually distinct.

The main phase of the experiment was divided into two blocks of 90 trials, with a self-paced break in between. On each trial, a sentence was presented on a computer screen (at a viewing distance of 60cm) in white, size 50, Mono font on a black background. Sentences were presented incrementally, with each of the four segments of a sentence (Subject, verb, intervening word/phrase, disambiguating noun) presented in the center of the screen for 250ms, followed by a 50ms inter-stimulus interval (total stimulus-onset asynchrony = 300ms). With four segments, this resulted in 1200ms total between the onset of the first word and the appearance of the response prompt. After the full sentence was presented, participants saw a red fixation dot in the center of the screen, cuing them to respond. Participants had 5 seconds maximum to respond with either the hand or foot to indicate whether the sentence was literal or abstract. After responding, the text "Next trial" was displayed for 500ms, followed by 500ms of a white central fixation dot, which indicated that the next trial was beginning.

While stimuli differed in the length of the intervening word/phrase (from zero to three one-syllable words), all words of the intervening phrase were presented simultaneously (or a blank screen was presented, if the intervening phrase was size zero) for 250ms, followed by a 50ms ISI, such that all sentences still took the same amount of time to be displayed. Our pilot study showed that participants had no difficulty reading up to three one-syllable words (e.g. "...out on the...") during the 250ms presentation. In previous rapid serial visual presentation (RSVP) tasks, words have been recognizable at a rate of 63ms per word (Forster, 1970). Moerover, eye-tracking data of reading has shown that short function words in the parafovea can be processed without being fixated (Rayner, 1998; Rayner & McConkie, 1976).

Before each experimental block, participants performed 20 "warm-up" trials to familiarize themselves with the current response contingencies. On warm-up trials, participants saw a single word, "Literal" or "Abstract," below a simple graphic of the corresponding body effector (i.e. a silhouette image of a hand or foot) and executed the corresponding response. This was intended to train an association between the response effector and its meaning, or to reverse the association from the previous block. Participants then completed 20 "practice" trials consisting of literal and metaphorical sentences in the same format as the experimental stimuli, but for which the verbs were expected to be unrelated to either the hands or the feet. Participants could not move on from either of these sections until they achieved at least 80% accuracy.

3.3.1.0.6 Preprocessing: Exclusions and Transformations An items-level analysis of accuracy called our attention to one item with only 6% mean accuracy: the phrase "walked in the park." Manual inspection of the stimulus set showed that this phrase was mistakenly coded as metaphorical, most likely due to its similarity to the expression "a walk in the park," when in fact the present usage arguably only fits with a literal interpretation. This item was therefore removed from all analyses.

In pre-processing the data, initial exploration revealed that participants were less accurate in classifying metaphorical statements than literal statements. Therefore, rather than exclude participants based on an accuracy cutoff that averages over both statement types, we used signal detection theory to determine each participant's discriminability index for the literal/metaphorical distinction. We excluded any participant with a discriminability index more than one standard deviation below the mean. This cutoff excluded 14 participants, resulting in a total of 86 participants in the final dataset, used for all subsequent analyses. Beginning with 15394 trials (179 trials X 86) participants, we next excluded outlier trials as a function of reaction time. Raw reaction (RT) time data was highly right-skewed, as is typical, and therefore was log transformed to satisfy statistical assumptions of normality. We then computed the mean and standard deviations of the log RTs independently for each participant and each response effector (hand or foot). Next, we eliminated any trial for which the log RT was more than two standard deviations above the group mean (433 trials) and any trial for which it was more than two standard deviations below the group mean (227 trials). This left 14724 trials, or 95.7% of the original dataset.

The action-sentence compatibility effect would manifest as an interaction between the response effector used and the degree to which the sentence associates with the response effector. To make this effect more interpretable in light of the moderating variables under investigation, we first recoded trials as being congruent—where the correct response (a hand or a foot press) was consistent with the body-association of the sentence—or incongruent. This recoding aggregates across hand and foot responses, and therefore transforms the ACE interaction effect into a main effect of congruency. The degree to which the noun and verb are associated with the hand/arm versus foot/leg was also transformed into a "congruency bias." For example, a verb that is strongly associated with the hand/arm would have a high, positive congruency bias when the correct response uses the hand, while it has the inverse congruency bias when the correct response uses the foot.

3.3.2 Results

3.3.2.1 Accuracy

The model tables for all analyses of experiment 1 are presented in appendix A. The percentage of correct responses as a function of congruency (congruent vs incongru-

ent) and usage category (literal vs metaphorical) is plotted in Figure 3.1. These data were analyzed using a logistic mixed-effects model, implemented using the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015). The maximal model included fixed effects for congruency (congruent vs incongruent) and usage category (literal vs abstract), as well as their interaction. The random effects structure included random intercepts for participants and items. Step-wise additive model comparison (Table A.1) revealed a significant effect of sentence type (b =.72, SE = .225, χ^2 = 7.3981, p < .01), such that accuracy was higher for literal sentences than for abstract ones. Because the potential differences between literal and abstract sentences in the ACE effect is a main focus of this study, we conducted planned comparisons on each sentence type independently, despite the lack of a significant interaction. These analyses (Table A.2) detected an effect of congruency in the literal condition (b = -.25, SE = .16, χ^2 = 8.1005, p < .01), but not in the metaphorical condition (Table A.3).

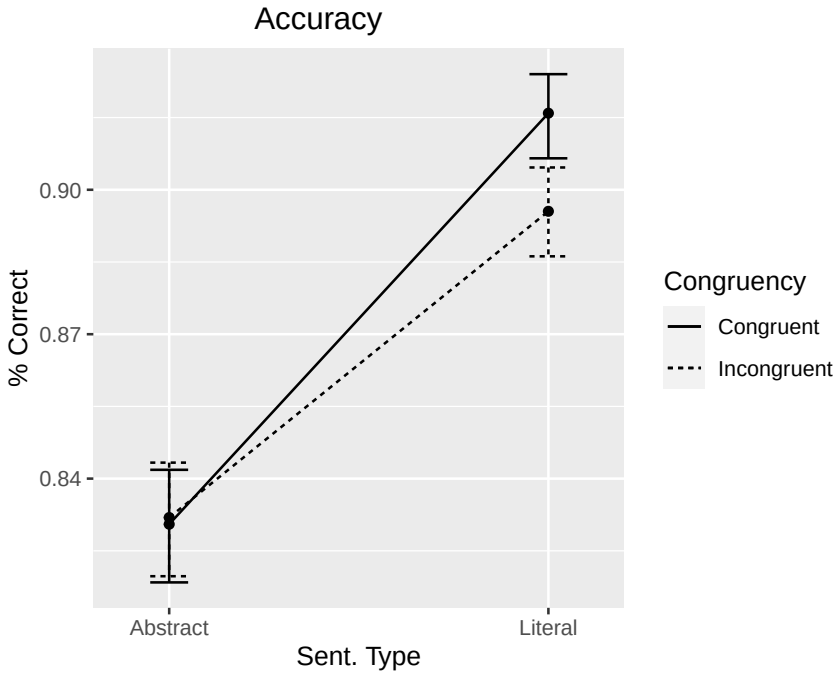


Figure 3.1: Accuracy as a function of sentence type and sentence-response congruency. Error bars represent 95% C.I. obtained from nonparametric bootstrap sampling.

3.3.2.2 Reaction Time

In analyzing reaction times, we first excluded all trials where participants responded "incorrectly" (1862 trials, 12% of the data). In many cases, metaphorical phrases can be reasonably interpreted as literal (as Amelia Bedelia hitting the road shows), so we make room for the possibility that stimuli were interpreted differently than they

were coded. However, because our research questions relate to the access of literal or metaphorical meaning, excluding trials where participants' responses did not match the coding of the phrase helps limit our analysis to cases in which participants actually interpreted the phrases as such (while acknowledging that there are surely some "false alarms" among the data). This final exclusion left us with 12862 trials.

We next collapsed the 2 (response-effector) X 2 (sentence-body-association) into a single effect of congruency (Figure 3.2). The log-transformed reaction time data were analyzed using linear mixed effects modeling, implemented using the lme4 package in R. Fixed effects included congruency (congruent or incongruent) and usage (metaphorical or literal), which were treated as factors, as well as their interaction. The random-effects structure included random intercepts for participants and items. As shown in Table A.4, there was a main effect of usage category (b = -.074, SE = .026, χ^2 = 5.58, p < .05) such that literal sentences elicited faster responses than metaphorical ones, as well as a main effect of congruency (b = .009, SE = .01, χ^2 = 11.9162, p < .001) such that congruent responses were faster than incongruent responses. There was also a usage X congruency interaction (b = .03, SE = .01, χ^2 = 4.2032, p < .05) such that the congruency effect appeared in the literal condition (b = .04, SE = .01, χ^2 = 16.03, p < .001), but not in the metaphorical condition.

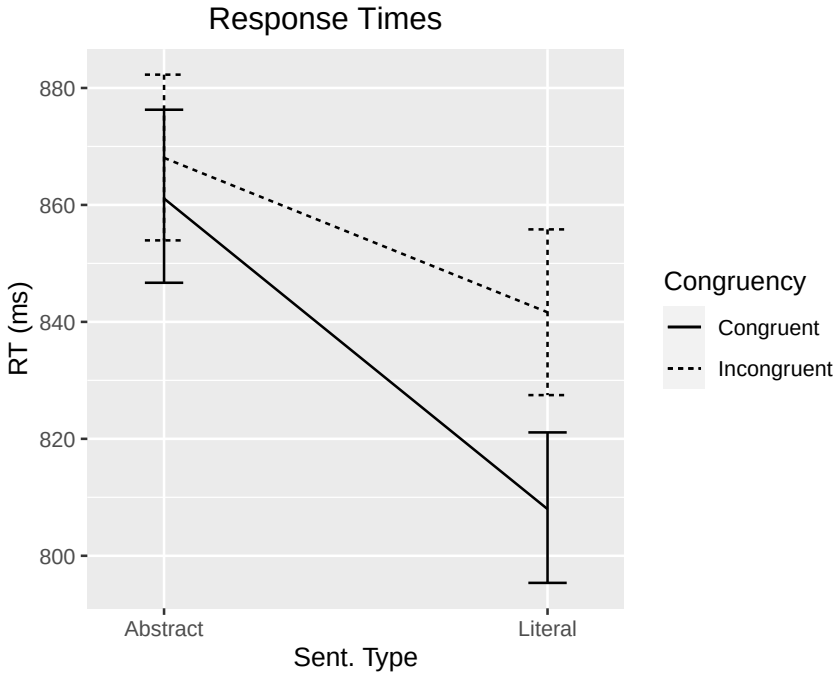


Figure 3.2: Mean response time as a function of sentence type and sentence-response congruency. Error bars represent 95% C.I. obtained from nonparametric bootstrap sampling.

3.3.2.3 Variability of ACE Across Items

We next examined the degree to which the congruency effect on reaction times was moderated by relevant psycholinguistic variables, including the frequency, familiarity, and figurativeness of the phrase, and the contextual diversity of the noun and verb. We also considered the possibility that the degree to which the verb and noun were associated with the congruent response would moderate the ACE, over and above whether a sentence was grouped into the hand/arm or leg/foot stimulus categories. Finally, we considered whether participant scores on the Vividness of Visual Imagery Questionaire would predict the magnitude of the ACE.

We first began by entering all above-mentioned psycholinguistic variables into an omnibus model, along with effects of sentence type, congruency, and their interaction. Random effects include intercepts for participants and items. This analysis (Table **??**) detected significant three-way interactions of sentence type and congruency with phrasal frequency, figurativeness, and the contextual diversity of the verb. Each of these effects are plotted below and discussed individually.

3.3.2.3.1 Phrasal Frequency.

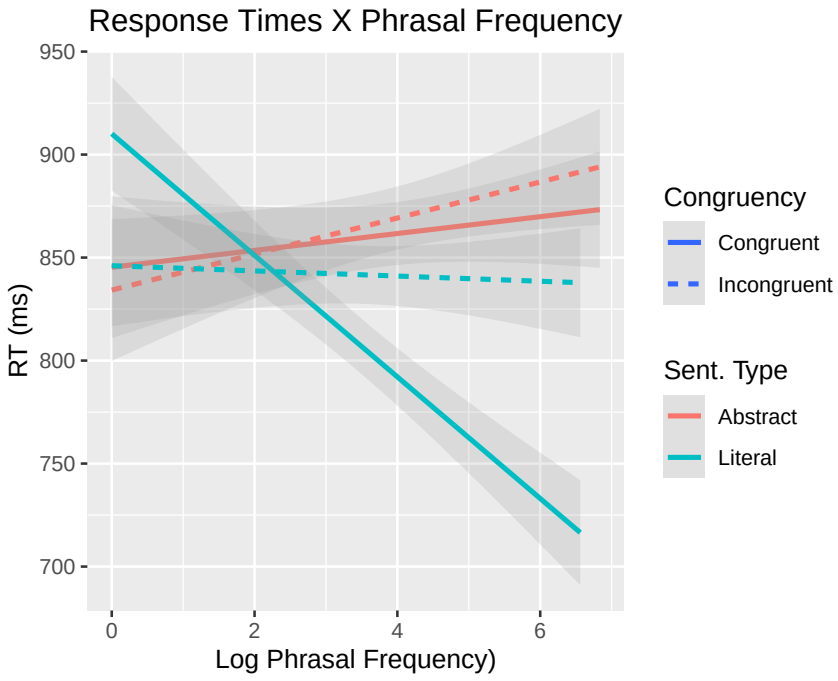


Figure 3.3: The effect of the frequency of the phrase on reaction times.

Our analysis revealed a significant interaction between congruency and phrasal frequency for literal sentences, but not for abstract ones. As Figure 3.3 shows, the

congruency effect was much larger for the most frequent literal sentences relative to less frequent literal sentences.

3.3.2.3.2 Figurativeness.

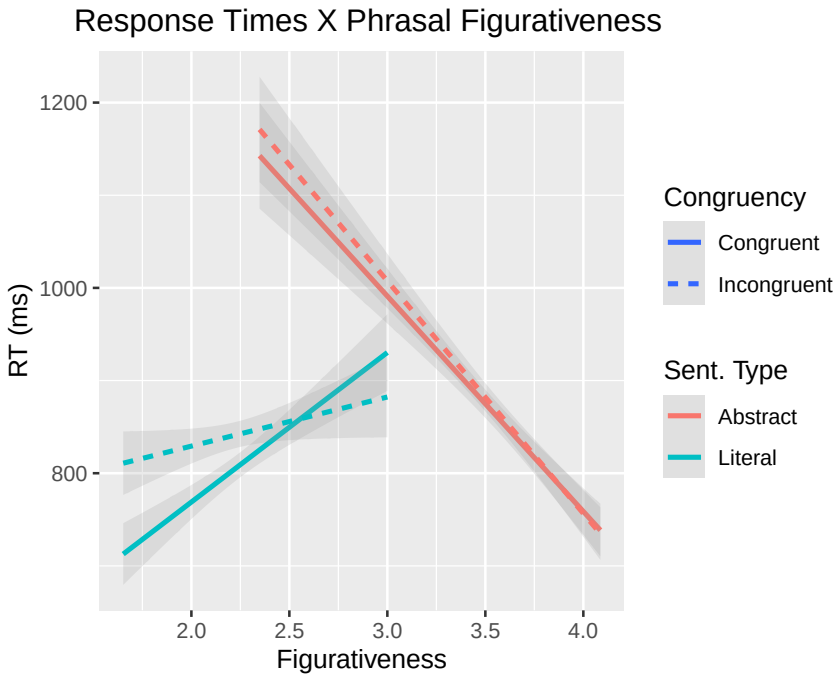


Figure 3.4: The effect of the figurativeness rating of the phrase on reaction times.

We also found a significant interaction between congruency and figurativess for literal sentences, but not for abstract ones. Figure 3.4 shows that literal sentences that were rated as more figurative elicited a smaller congruency effect.

3.3.2.3.3 Contextual Diversity of the Verb.

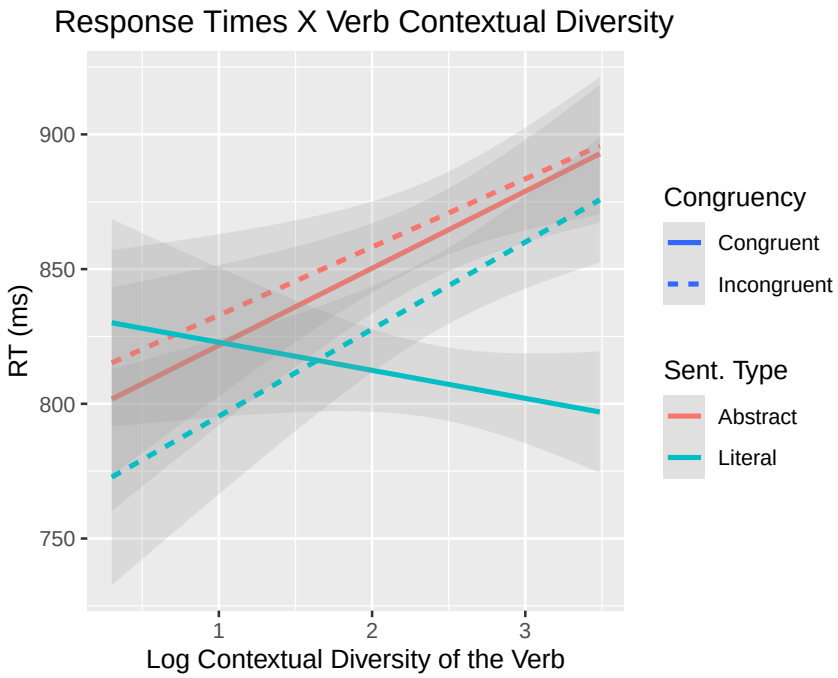


Figure 3.5: The effect of the contextual diversity of the verb on reaction times.

Finally, we found a significant interaction between congruency and the contextual diversity of the verb, again present only for literal sentences. As Figure 3.5 shows, this interaction was such that more contextually-diverse verbs elicited a larger congruency effect in literal contexts.

3.3.3 Discussion

The results of Experiment 1 demonstrate a clear congruency (ACE) effect for literal sentences: participants were both more accurate and faster when responding to congruent literal sentences relative to incongruent literal sentences. The faciliatory ACE effect on reaction times also showed substantial contextual moderation. The ACE effect on reaction times for literal sentences strengthened with increases in phrasal frequency and contextual diversity of the verb, but diminished with increases in figurativeness. These results are consistent with the notion that recruitment of motor information during literal sentence comprehension is highly sensitive to linguistic context. On the other hand, metaphorical sentences showed no evidence of a main effect of congruency in either the accuracy or reaction time measures.

Our results are potentially consistent with the hypothesis that congruency effects are the result of sensorimotor simulations or mental imagery that occur after the processing of a sentence is complete. However, it is worth noting that we detected no effect of participant scores on the Vividness of Visual Imagery questionnaire. Our results do

not fit with the more specific hypothesis that such simulations are specifically of the *self* performing an action, as argued for by Papeo, Corradi-Dell'Acqua, and Rumiati (2011). While those authors found increased motor activity only for first-person uses of action verbs, all of our sentences were in the third person, past tense, and yet effects are still found.

3.4 Experiment 2

The results of Experiment 1 were consistent with the notion that action-sentence congruency effects are the result of sensorimotor simulation or mental imagery that occurs *after* processing of a sentence. As such, those results are unable to tell us whether sensorimotor information is functionally relevant for language processing itself. Experiment 2 examined the possibility of a functional role for sensorimotor representations during literal and metaphorical sentence comprehension by employing a simultaneous manual-interference. This manipulation was inspired by previous work by Yee et al. (2013), who had participants engage in a concurrent "patty-cake" motion with the hands while making a concrete/abstract judgment in response to single words. This manipulation was intended to be analogous to transcranial magnetic stimulation of hand regions in motor cortex, which has previously been shown to selectively interfere with the processing of manipulable objects. While a simple hand motion likely exerts a weaker and more spatially diffuse effect, it has the advantage of simplicity, avoids potential criticisms of TMS methods, and requires no assumptions about the organization of brain regions involved with particular tasks. Yee et al. (2013) found that this manual task increased error rates and reaction times more for items that were commonly interacted with using the hands (e.g. *pencil*) relative to similarly-concrete items that were not commonly interacted with using the hands (e.g. *tiger*), providing support for the notion that the processing of hand-related words relies on the same brain regions involved with action using the hands. Similar results were obtained in a study from Shebani and Pulvermüller (2013), who found that moving the hands or feet selectively impaired memory for hand- or foot-related action verbs, respectively.

We hypothesized that, if metaphorical meaning is grounded in sensorimotor processes, a simultaneous hand-motion during sentence processing should interfere with activation of both a literal and metaphorical meaning. Conversely, if sensorimotor and "amodal" representations operate in parallel, a simultaneous hand-motion should selectively interfere with activation of literal meanings. Based on the results of Experiment 1, we predicted that we would find the latter pattern.

A second goal of this experiment was to shed some light on the time course of any effects, should they be present. Previous work has used a variety of different methods, and found some conflicting results, which has made some patterns difficult to interpret. Some studies have considered the processing of commonly manipulable nouns,

while others have looked at the processing of action verbs. Most have presented single words in isolation, while some have examined the processing of sentences. However, studies that have used sentences as stimuli often still take as their dependent measure a single response that occurs at the end of a trial, such as a reaction time or an accuracy measure. This makes it difficult to ascertain which aspect of a sentence specifically impacted processing. Eye-tracking methods may help to clarify the situation in that they may allow us to consider the processing of both verbs and nouns *while* they are being processed in real time.

We made use of the "Visual-World" paradigm, in which participants view a screen containing task-relevant items in bounded regions of interest, and eye fixations are recorded while participants hear a sentence. Our participants saw a screen simply containing the words "Literal" and "Abstract" on opposite sides. Since the use of a concurrent manual task made it impossible to elicit responses with the hands, participants instead made responses using only the feet. On each trial, participants heard a literal or abstract sentence containing a hand-related verb. At the end of each sentence, arrows appeared below each response option, indicating the direction in which to respond with the foot (UP or DOWN for a toe- or a heel-press, respectively). These response contingencies changed across trials, which made fixations to at least one response option necessary for accurate responding on each trial. Since eye movements are quite low-cost, rapid, and automatic, and it is well-established that humans make anticipatory fixations to objects that are expected to be task-relevant, we hypothesized that fixations to the 'Literal' or 'Abstract' response boxes would be indicative of participants' relative preference for a literal versus abstract interpretation of each sentence as it unfolded.

3.4.1 Method

3.4.1.0.1 Design. Experiment 2 used a 2 (usage type: literal or metaphorical) X 2 (manual interference: present or not) repeated measures design. Participants were randomly assigned to do a simultaneous manual-interference task in either the first or second block of the experiment. Trial order was fully randomized for each participant.

3.4.1.0.2 Materials. The stimuli for experiment 2 consisted of just the 90 hand/arm sentences used in Experiment 1. Leg-related items were not included, so as to avoid mixing the potential effect of the hand motion manipulation with possible ACE effects when responding with the foot. Thus, all items in Experiment 2 correspond to "incongruent" combinations from Experiment 1.

The sentences were recorded by a male talker in a quiet room using an Audio-Technica AT2020 microphone and a Focusrite 2i2 audio interface, then saved to the computer

as WAV files for editing. Recordings were digitized at a sampling rate of 44,100 Hz with a bit depth of 24. All audio editing was performed using Praat (Boersma & Weenink, 2013). The sound levels for all sentences were normalized to the same mean intensity. The acoustic onsets of the verb and the disambiguating noun, as well as the offset of the noun, were recorded manually in Praat.

3.4.1.0.3 Participants. Sixty-one participants from the University of Merced community were recruited using the Sona research participation system. Participants provided informed consent in accordance with IRB policies and were compensated for their time with course credits. Participation was restricted to those with normal or corrected-to-normal vision, normal hearing, who were fluent in English, right-handed, and reported not having dyslexia, other reading disabilities, or any physical problems that would prevent simple movements with the hands

3.4.1.0.4 Apparatus. Responses were again collected using the Makey Makey (described in the "Apparatus" section of experiment one). However, experiment 2 used two foot-pads to collect responses rather than one hand- and one foot-pad. Participants rested their right foot on a bar such that their heel was suspended above one pad (the DOWN response pad) and their toes were suspended above another pad (the UP response pad).

Participants' eye gaze was recorded using an Eyelink II head-mounted eye tracker. Prior to the beginning of the experiment, the eye tracker was calibrated using a standard nine-point grid, and the subject was shown how to perform a drift correction, which took place once every 15 trials. Eye movement data was collected via the Eyelink control software and custom MATLAB scripts. Data from the right eye were collected using both pupil shape and corneal reflection at a sampling rate of 250Hz.

3.4.1.0.5 Procedure. Participants completed the experiment individually in the laboratory in a single session lasting approximately an hour. Before beginning the experimental task, participants first completed a Vividness of Visual Imagery (VVIQ) questionnaire (Marks, 1973). After completing the survey, participants were seated in front of a computer at a viewing distance of 60cm, taught how to use the response apparatus, and the eye tracker was calibrated. Participants then put on high-quality headphones (Sennheiser HD-558 open back headphones), and the volume was set to the participant's most comfortable level.

The main phase of the experiment was divided into two blocks of 45 trials with a self-paced break in between. On each trial, participants saw two 300 x 300 pixel, white-outline boxes on a black background. The boxes were vertically centered on a 1920 x 1200 pixel screen and placed on the left and right sides of the screen, respectively, with 130 pixels between the edge of each box and the closest edge of the

screen. At the top of each box was the label "Abstract" or "Literal" presented in white size 70, Segoe UI font. The locations of the labels were randomized for each participant, but remained the same throughout a single session.

Before each trial, participants saw the trial number appear in the center of the screen for 500ms, indicating that the next trial was beginning. Each trial began with a white fixation dot presented at the center of the screen while a spoken sentence was played in the participant's headphones. At the end of each sentence, the fixation dot disappeared and a small, dark-grey UP or DOWN arrow appeared in each box. The arrows indicated to participants in which direction they should respond with the foot for each of the two possible responses (literal vs abstract) on that trial. The arrows were designed to be difficult for participants to accurately see with peripheral vision, such that they needed to direct their focus to their box of choice in order to determine the correct response. For example, if the participants heard a literal sentence, they would fixate the box labeled "Literal," observe the current direction of the arrow, and respond with the foot in that direction. The response contingencies for the arrows were pseudo-randomly assigned such that UP corresponded to abstract half the time and literal half the time, and vice versa for the DOWN response.

Before beginning the main phase of the experiment, participants completed 20 practice trials using the same filler stimuli as in the practice trials of experiment one, which were unrelated to either the hands or the feet. Participants were required to get a score of at least 80% in the practice phase in order to proceed and had two attempts to do so.

3.4.2 Results

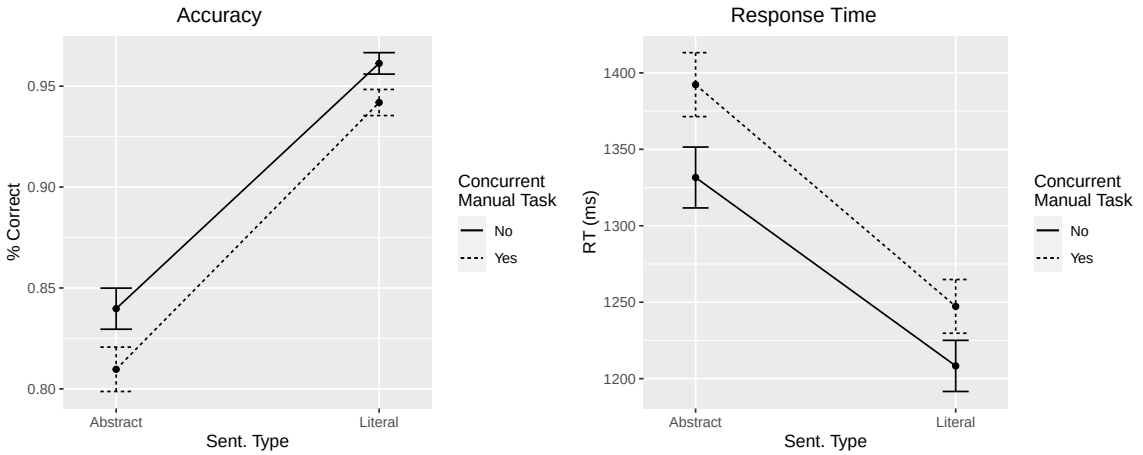
The model tables for all analyses of experiment 2 are presented in appendix B. Results below are broken down into two sections. First, we plot and analyze behavioral measures that were obtained once at the end of each trial: accuracy (whether the sentence was correctly categorized as abstract or literal) and reaction time.

3.4.2.1 Accuracy and Reaction Time

Mean accuracy and reaction time as a function of sentence type (literal vs abstract) and concurrent hand motion are plotted in Figure 3.6a. These data were analyzed using a logistic mixed-effects models for accuracy, and linear mixed-effects models for reaction times. The maximal model included fixed effects for sentence type and concurrent hand motion as well as their interaction. The random effects structure included random intercepts for participants and items. Analyses for reaction time data were performed on the log-transformed values.

The analysis of accuracy (Table B.1) revealed a significant effect of condition (b = 1.44, SE = .3, χ^2 = 19.19, p < .001), such that participants were less accurate for abstract sentences relative to literal ones. There was also a significant negative effect of the concurrent hand motion on accuracy (b = -.24, SE = .1, χ^2 = 5.39, p < .05). There was no interaction. Thus, unlike with the effect of congruency in Experiment 1, the concurrent hand motion did not selectively interfere with the processing of literal sentences, and instead impacted both types. As in Experiment 1, all subsequent analyses excluded incorrect responses.

The complementary pattern was found in the analysis of reaction times (Table B.2): a slowing of responses while engaging in the concurrent hand motion (b = .03, SE = .01, χ^2 = 11.6, p < .001), and faster responses for literal sentences relative to abstract ones (b = -.12, SE = .03, χ^2 = 5.27, p < .05).



(a) (b)

Figure 3.6: Mean accuracy (a) and reaction time (b) as a function of sentence type and concurrent manual task.

3.4.2.2 Gaze Analyses

The averaged time-course of fixation proportions to the 'Literal' and 'Abstract' response options is plotted in Figure 3.7, time-locked to the onset of the sentence-final noun. Several different analyses were conducted, each described in detail below. The first focused on the overall effect of the concurrent hand motion on the relative preference to fixate the 'Abstract' response option, as well as the extent to which this effect varied across time during and immediately following the sentence. Analyses two and three collapse over time, taking binary measures of which response option was fixated first subsequent to the onset of the noun, and whether each item was fixated at least once during the relevant window of analysis.

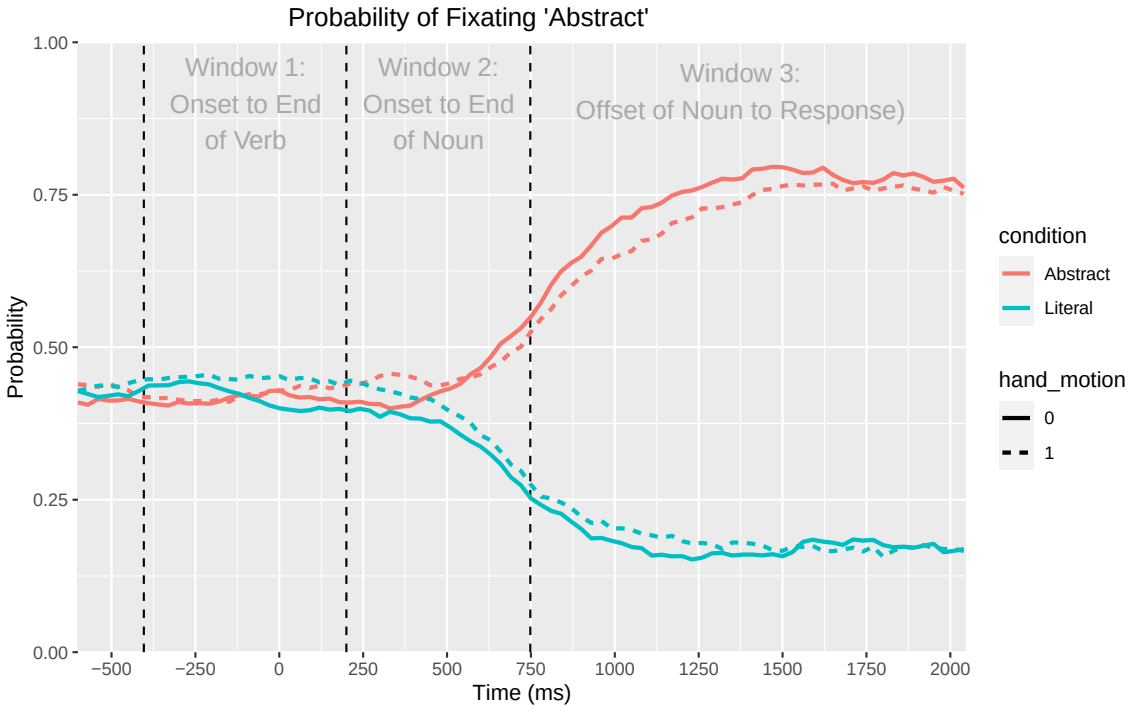


Figure 3.7: Proportion of fixations to the 'Abstract' response box over time, as a function of sentence type and concurrent manual task. Time in this plot is relative to the onset of the sentence-final noun, which disambiguated matched abstract and literal sentences. Dotted lines demarcate the three analysis windows. Note that all sentences were time-locked to the onset of the noun, while the length of the windows on either side varied slightly across items.

In order to consider how interference effects unfold over time while processing a sentence, each trial was divided into three time bins, corresponding to: (1) the onset of the verb to the onset of the noun, (2) the onset to the offset of the noun, and (3) the offset of the noun to the time of response. Region 1 serves as a baseline, allowing for examination of general tendencies to fixate the 'Abstract' versus 'Relative' response options, prior to the sentence-final noun that distinguished literal from matched abstract sentences. Region 2 allows us to consider how eye movements are affected as participants process the noun and reach an interpretation of the sentence. Region 3 allows for the examination of any effects that emerge as participants reach a determination on the sentence type and launch a behavioral response with the foot. These windows were defined at the level of individual items using the onsets and offsets as determined from the recorded materials prior to conducting the experiment. Following standard practice, all time windows were offset by +200ms to allow for the approximate time required to plan and launch a fixation. Time window was treated as a factor using reverse Helmert coding, with the first contrast testing for a difference between window 3 with the mean of windows 1 and 2, while the second contrast tested for a difference between windows 1 and 2.

The dependent measure used here was the "empirical logit," or the log-odds of fixating the "Abstract" versus "Literal" response option (see: Barr, 2008; Brown-Schmidt & Fraundorf, 2015), which was calculated independently for each trial. It should be cautioned that there are potential downsides to using this approach, which involves linear regression on logit-transformed data (see: Donnelly & Verkuilen, 2017). When there is only one target object of interest in a "Visual-World" design, it is preferable to maintain the structure of the raw, binary time-series data and utilize logistic regression. However, if there are two or more objects of interest, given that an individual can only fixate one or the other at a moment in time, it becomes necessary to aggregate data over some period of time. In such cases, the empirical logit provides a reasonable approximation to logistic regression. Furthermore, while three time bins is quite a coarse graining of our fixation data, this ensured a large number of data points in each bin, thereby mitigating the possibility of spurious results. Subsequent analyses using binary data further bolster the reliability of our results.

The data was analyzed using step-wise model comparison of linear mixed effects models, implemented with the *lme4* package in *R*. The random effects structure for all models included random intercepts for participants and items. This random effects structure was first obtained using a backwards model fitting procedure to determine the maximal structure that permitted model convergence. Fixed effects included sentence type, and concurrent hand motion. An omnibus test (Table B.3) showed a significant interaction between condition and time for the second contrast, comparing the final post-noun analysis window with the mean of the previous two windows. This simply corresponds to the increase of fixations to the 'Abstract' response options in window 3 when the sentence was abstract, and complementary decrease when the sentence was literal. This effect is of course not particularly meaningful, except in showing that participants understood the task and indeed look to the 'Abstract' response option reliably for sentence items that were intended and coded as abstract. However, it is worth noting that participants could technically accomplish this task by fixating the same response box on every trial (e.g. only literal) and responding in the *opposite* direction when the sentence does not match the response box. This would be a cognitively-taxing strategy and one we thought unlikely for participants to adopt, but is nonetheless helpful to confirm that it was not.

There was no three-way interaction detected among condition, time, and the hand motion, suggesting that the effect of the hand motion on relative preferences for the literal and abstract did not vary across time bins. As such, we conducted two further analyses: one controlling for time, condition, and their interaction and testing for effects of adding hand motion, and another exploring the effects of condition and hand motion in each time region, to more closely examine the time-course of processing.

In the first analysis controlling for time and sentence condition (Table B.4), model comparison revealed a significant interaction between hand motion and sentence type ($b = .16$, $SE = .05$, $\chi^2 = 11$, $p < .001$). Follow-up tests (Tables B.5, B.6) for a simple effect of the hand motion within each sentence type condition revealed a significant

positive effect for literal sentences (b = .1, SE = .04, χ^2 = 7.63, p < .01), such that participants were more likely to fixate the 'Abstract' response option when doing a concurrent hand motion, while they were *less* likely to do so for abstract sentences (b = -.12, SE = .03, χ^2 = 12.5, p < .001). This is indicated in time window 3 of Figure 3.7 as the lower height of the red dotted line (Abstract/concurrent hand motion) relative to the red solid line (Abstract/no hand motion). This pattern was not predicted, and may require further investigation in order to determine whether it is reliable, and if so, what it reflects. However, we will offer some suggestions in the discussion below.

3.4.2.2.1 Region-by-Region Analysis. The region-by-region analysis began with window 1, corresponding to the period from 200ms following the onset of the verb, to 200ms following the onset of the noun (mean length = 603±168ms). This window served as a baseline for examining any overall biases towards fixating the 'Literal' versus 'Abstract' response option before the sentence could possibly be determined as literal or abstract. Step-wise model comparison (Table B.7) showed no significant differences between sentence conditions, nor a significant effect of the hand motion. Thus, the best fitting model was an intercept-only model, which indicated a general negative bias (b = -.32, SE = .16), where 'Abstract' was coded as 1. In other words, this indicates a slight overall preference to fixate the 'Literal' response option, consistent with the pattern, observed in Figure 3.7, that the probability of fixating 'Abstract' begins below 50%.

Window 2 corresponded to 200ms following the onset of the noun, to 200ms following its offset (mean length = 548±121ms). the first moments of processing during which a participant could potentially determine whether a sentence was literal or abstract. We predicted that, if the concurrent hand motion were to interfere with the comprehension of literal sentences specifically, such effects may first be observed as participants process the disambiguating noun. However, model comparison was unable to detect any effects of condition nor hand motion (Table B.8). This was a surprising result, given that Figure 3.7 clearly shows that fixation probabilities begin to diverge about halfway through window 2. This suggests that window 2 may have been too short to allow the reliable estimation of effects—especially the undoubtedly much smaller effect of the concurrent hand motion, relative to the effect of sentence type. Thus, while it appears from Figure 3.7 that there was a bump in fixations to the 'Abstract' response box for both conditions in this time window, we can draw no strong conclusions about this visual pattern.

Finally, window 3 corresponded to the period beginning 200ms after the offset of the noun, until the point that the participant made a response (mean length = 1107±649ms). This analysis (Table B.9) detected a significant effect of condition (b = -1.49, SE = .06, χ^2 = 226, p < .001), with more fixations to the 'Abstract' option in the abstract condition, as would be expected, and also a significant interaction between condition and hand motion (b = .144, SE = .06, χ^2 = 4.925, p < .05).

Follow-up tests (Tables B.10, B.11) confirmed that the pattern detected in the first analysis, which controlled for time and condition, was driven by the effects seen in time window 3. That is, there were *fewer* fixations to the 'Abstract' response box during window 3 for abstract sentences while participants were engaged in the hand motion.

3.4.2.2.2 First-Fixations Analysis. Figure 3.8 shows the probability that the first fixation on a trial, following the onset of the noun, went to the 'Abstract' option, relative to the 'Literal'. First fixations were calculated by first filtering the data to include only time bins that followed the onset of the final noun in each sentence, which was the point of disambiguation between abstract and literal matched items. This window was offset by 200ms to account for the time required to launch a fixation. We then selected the first row of the data for which there was a fixation to either response option, and recorded which it was. If there were no fixations to either option following the onset of the noun, the trial was discarded. The resulting data was then recoded as a binary variable: 1 if the first fixation was to the "Abstract" option, and 0 if it was to the "Literal" option. These data were then analyzed using logistic mixed effects regression, including random intercepts for subjects and items as before, and step-wise model comparison to test for effects of condition, hand motion, and an interaction.

This analysis (Table B.12) found significant main effects of condition and hand motion. Participants were more likely to make the first fixation to the 'Abstract' option when the sentence was abstract (b = -.23, SE = .07, χ^2 = 9.3, p < .01), though there is a general bias towards making the first fixation to the 'Literal' option. The hand motion made participants reliably more likely to make the first fixation to the 'Abstract' option (b = .19, SE = .07, χ^2 = 7.8, p < .01). However, there was no interaction, indicating that the concurrent hand motion did not selectively influence the processing of one of the two sentence types.

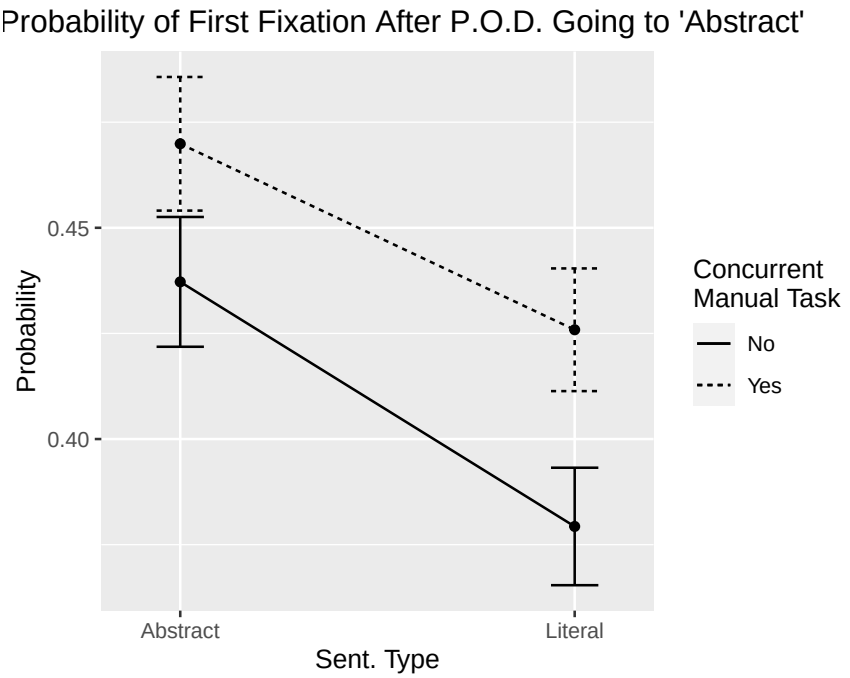


Figure 3.8: Probability of the first fixation following the point of disambiguation (the onset of the final noun) going to the 'Abstract' response option.

3.4.2.2.3 Binary Fixation Analysis. In the final analysis, we compressed the data into a binary measure of whether or not the 'Abstract' response was fixated in each time bin, on each trial, which was analyzed using logistic mixed effects regression. As in the first above analysis involving proportions of fixations, we began with a base model containing fixed effects of sentence type, time and their interaction, and used step-wise model comparison to test for an effect of hand motion and any interactions (Table B.13). Consistent with the first analysis, we again found a significant main effect of hand motion (b = .09, SE = .04, $\chi^2 = 4.29$, $p < .05$) interaction between condition and hand motion (b = .17, SE = .08, $\chi^2 = 4.01$, $p < .05$). Follow-up analyses (Tables B.14, B.15) showed a positive simple effect of the hand motion in the literal condition (b = .16, SE = .05, $\chi^2 = 8.7$, $p < .01$), indicating that participants were more likely overall to fixate the 'Abstract' response option during these sentences if they were engaged in the concurrent manual task. No effect of the manual task was detected for abstract sentences. Therefore, this analysis does not corroborate the pattern observed in the first analysis involving a decreased proportion of fixations to the abstract 'Abstract' option for abstract sentences when engaged in the concurrent manual task.

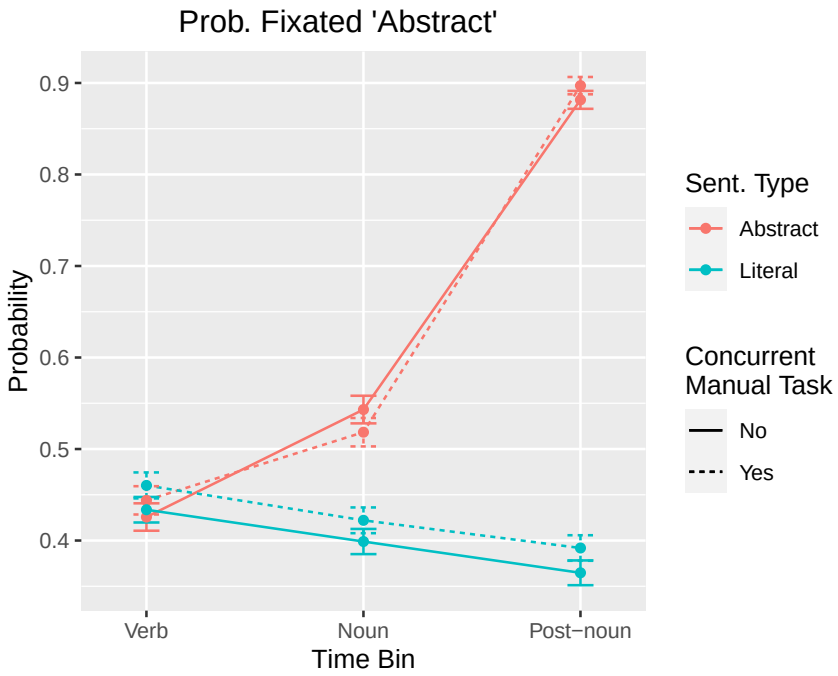


Figure 3.9: The probability that participants fixated the 'Abstract' response option at any time during each analysis window, as a function of sentence type and concurrent manual task.

3.4.3 Discussion

The results of Experiment 2 provide additional support for the notion that the same processes involved with action—in this case, action with the hands—are involved with the processing of language that describes hand-related actions. This was revealed in the fact that, when participants are engaged in a concurrent hand motion while hearing sentences, participants make a relatively greater proportion of fixations to an 'Abstract' response box, are more likely to fixate it at least once during the trial, and are more likely to make the first fixation towards this option, following the point of disambiguation in each sentence.

We find some mixed evidence for the possibility that manual interference impacted both abstract and literal sentences. While the latter were reliably impacted across analyses, the final analysis, which used a binary measure of whether 'Abstract' was fixated on each trial, indicated an interaction between sentence type and hand motion. Follow-up analyses found that the effect of the hand motion on increasing fixations to 'Abstract' was present only for literal sentences. The second analysis, examining first fixations, revealed an overall effect of hand motion on increasing fixations to the 'Abstract' option, but no interaction with condition. Meanwhile, the first analysis showed an interaction and a *negative* effect of hand motion for abstract sentences,

driven by the post-noun time window. This latter effect was in the opposite direction from what was predicted, but given the different patterns of results across these analyses, it is difficult to interpret. As such, it is unclear if in this design *both* literal and abstract sentences were impacted by moving the hands, or only the former.

The region-by-region analysis of fixations was unable to detect significant effects of hand motion in either the first or second time windows, corresponding to processing of the verb and noun, respectively. However, an effect of the hand motion was detected across all three time bins, and appeared driven primarily by changes in fixation proportions in the final time window. This can be taken as evidence that motor activity becomes strongly relevant in the time period shortly *after* a sentence has been completed, and is being interpreted as literal or abstract. However, it should be noted that the region-by-region analysis was also unable to detect the difference in fixations across sentence types in window 2, where Figure 3.7 clearly shows fixations beginning to diverge. This suggests that our region-by-region analysis may have been underpowered, and therefore we should not rule out the possibility that action effects are also relevant in the earliest stages of word processing.

There also was nothing in the analyses indicative of an effect of the hand motion during or shortly after the verb of each sentence. However, we should again hesitate to draw strong conclusions from this null result, as the design of the task may have made an effect at this stage difficult to observe. Because the arrows that indicated the direction in which to respond for a sentence categorized as abstract or literal did not appear until the end of the sentence, the two response boxes were less task-relevant during early stages of the sentence. Because all of our stimuli had the same basic structure and were of similar lengths, participants could quickly become aware of when the response contingencies would appear on each trial. As such, there may simply not have been strong enough motivations to make fixations to *any* response option during the processing of the verb, such that effects could be present at that time, but remain too small to observe in our data.

3.5 Conclusions

Taken together, the results of these two experiments may help to shed light on some mixed results in the literature on sensorimotor grounding, and also raise some new questions. Firstly, Experiment 1 offers yet another replication of the basic action-sentence compatibility effect involving action verbs. Participants were both faster to respond and more accurate when the effector (hand or foot) of the response matched the action verb of the sentence, specifically for literal sentences. This was not an assured outcome, in light of the fact that we used sentence stimuli, as opposed to the large majority of prior work that has used single-word stimuli. If such compatibility effects were specifically linked to the processing of action verbs themselves, and persisted for only a short period of time, we would have been unable to observe

them. Thus, the fact that we do find a compatibility effect indicates that the action described by a sentence as a whole can still elicit a compatibility effect.

There have been quite mixed findings in prior work regarding a potential role of the sensorimotor system in the processing of metaphorical sentences. There have been several demonstrations for somatotopic activation of sensorimotor regions during the reading of metaphorical sentences, as well as a few demonstrations of ACE effects in metaphorical sentences. However, other studies have reported null patterns for metaphorical sentences in both neuroimaging and behavioral work. Our overall pattern of results in Experiment 1 provided no evidence that the comprehension of metaphorical sentences involves sensorimotor systems. Participants did not respond more quickly or accurately with the hands to metaphorical sentences involving hand action verbs, when averaging over all items, nor did any of the psycholinguistic variables explored moderate an effect.

However, the results of Experiment 1 paint a more complex picture than simply claiming that literal sentences involve sensorimotor simulation, while abstract ones do not. An analysis of how the ACE effect was influenced by a number of different psycholinguistic variables revealed substantial variability across items within the literal condition. Literal sentences are not *always* affected by congruency, and instead this effect is moderated by variables such as the frequency and figurativeness of the phrase, and the contextual diversity of the verb. These results add to earlier proposals that sensorimotor grounding or simulation is not an all-or-none phenomenon, but instead highly flexible and context-dependent. It may be the case that our abstract sentences were all *too* abstract to elicit a congruency effect, but the fact that the literal sentences that were rated as relatively more figurative produced diminished congruency effects points to the possibility of a continuum of effects. A metaphorical sentence that is quite novel and not so immediately recognizable as abstract may still produce a congruency effect.

Experiment 2 provided additional support for a functional role of the motor system in the processing of literal sentences. Given our experimental design, this resulted in a pattern that may seem counter-intuitive in light of the foregoing discussion: participants were *more* likely to momentarily entertain an abstract interpretation of literal sentences while moving the hands. While this effect emerged more reliably for literal sentences, it would be misleading to then claim that the motor system is only functionally relevant for the processing of literal sentences. When there is more than one reasonable interpretation of a sentence, at least momentarily, then influencing the likelihood of interpreting the sentence as literal necessarily also influences the likelihood of *other* interpretations. In other words, given that language processing occurs in a highly integrated architecture, the inhibition or priming of any one outcome alters the relative probability of all other outcomes. In this way, we may say that the processing of metaphorical sentences can be *embodied*, without necessarily being grounded in sensorimotor system. That is, concurrent action still plays a role in metaphorical interpretations of sentences, even when that interpretation involves

no reactivation of sensorimotor representations.

Our region-by-region analyses in Experiment 2 provide clear evidence of effects in the period following the offset of the noun until the response. This is consistent with the possibility that the concurrent hand motion influenced a late stage of processing that may be associated with mental imagery, mental simulation, or information integration. However, we are unable to rule out the possibility of earlier effects that occur *during* the processing of each noun, given that our analyses in time window 2 were unable to detect differences between sentence types, indicating that this analysis may have been underpowered.

It should be noted that our failure to detect congruency or interference effects for metaphorical sentences does not necessarily mean that the comprehension of these sentences is not grounded in sensorimotor representations. Instead, it may be that metaphorical sentences are grounded in *different* regions of the brain. For example, the "Words-as-Social-Tools" proposal from Borghi et al. (2019) suggests that metaphorical or abstract language may be grounded in motor regions more involved with social interaction and communication, such as mouth regions. Alternatively, the grounding of abstract languages may be more highly distributed across brain regions, while literal language is more concentrated.

Ultimately, we suggest that it may be time to move on from the debate regarding whether or not any particular type of language depends upon any particular type of grounding. Given that sensorimotor activation is clearly context-dependent and flexible, the processing of any given type of language may appear to involve sensorimotor systems at one moment, but not at another. The more helpful questions going forward may be (1) when is sensorimotor involvement *necessary* for comprehension and when is it unncessary, (2) what aspects of contexts shape the recruitment of sensorimotor information?

3.6 Acknowledgements

I would like to thank Daenna Mabalay, Natalie Cruz, Katherine Crenshaw, Casandra Moua, Ricardo Dionicio, and James Waterford for assistance with data collection and stimulus preparation.

Chapter 4

Context-Sensitive Categorization of Phonemes in Spanish-English Bilinguals

Speech processing presents listeners with a complex categorization problem. In order to recover the intended meaning of an utterance, listeners must first have a way of mapping the continuous dimensions of sound onto discrete linguistic categories, such as phonemes and words, and then onto discrete semantic categories, such as referents and events. But this is no simple feat, as listeners must contend with substantial variability in the production of speech sounds across speakers, contexts, and levels of background noise. Furthermore, even identical speech sounds may point to different meanings, as in the case of homophones.

These issues are even *more* pronounced for bilinguals, who must also contend with inter-language ambiguity. For example, consider that the English "b" is acoustically very similar to the Spanish "p" on the dimension of voice-onset-time, which is a cue to the distinction between voiced and voiceless phonemes that is present in both languages. English speakers tend to produce voiced phonemes such as /b/ with a VOT near 0, while voiceless tokens such as /p/ are produced with a positive VOT; Spanish speakers, on the other hand, tend to produce voiced phonemes with *negative* VOT values, while voiceless phonemes are produced with VOT values near 0 (Lisker & Abramson, 1964).

How bilinguals cope with acoustic ambiguities such as this has been a question of interest to psycholinguists for many years (Elman, Diehl, & Buchwald, 1977; Flege & Eefting, 1987; Williams, 1977), but results remain inconclusive. The purpose of this study was to compare evidence for two distinct possibilities: (1) that bilinguals possess just one set of phonetic categories that is applied cross-linguistically, or (2) that bilinguals possess different sets of phonetic categories for each of their two languages. Given the second possibility, we also considered the extent to which the activation of

the language-specific categories could be moderated by context.

Spanish-English bilinguals conducted a /b/-/p/ phonetic categorization task in the lab. One group of participants saw a set of instructions in English, while another saw instructions in Spanish. Additionally, participants made their categorization judgments by clicking on an image that represented a /b/ or /p/ word in the corresponding language condition (e.g. images representing 'beso' (kiss) and 'peso' (dollar) in the Spanish condition, or 'bear' and 'pear' in the English condition). A group of monolingual English speakers served as control. Categorization data was fit to a Gaussian mixture model that estimated the centroids and relative activation levels of phonetic categories for each participant. We found that Spanish-English bilinguals in the English condition were best fit by models that resembled monolingual English speakers (with low activation of Spanish phonetic categories) while bilinguals in the Spanish condition showed stronger activation of Spanish phonetic categories. These results provide evidence that Spanish-English bilinguals possess *four* phonetic categories corresponding to the typical /b/ and /p/ sounds in each language, and can flexibly adjust the salience of categories in each language as a function of context.

4.1 Modeling Phonetic Categorization

We may think of the speech processing system as implementing a function that maps points in the continuous space of acoustic cues onto phonetic categories. In order to understand how listeners deal with acoustic ambiguity, then, we may hypothesize various possible functions for this mapping, and then consider which function best describes human behavior. One possibility that was considered early in the history of psycholinguistic research was a simple category boundary: listeners may map inputs that fall on one side of the acoustic boundary onto one phonetic category, while inputs that fall on the other side are treated as belonging to an alternative category. Given such a function, we would predict listeners to be insensitive to any acoustic differences that do not span across this boundary. One way to test this hypothesis is to present listeners with acoustic tokens that vary in small increments along an acoustic dimension, such as VOT, and observe the proportion of times that listeners categorize each token as belonging to each category. Early experiments using this method revealed sharp boundaries in categorization responses, such that, for example, English speakers may categorize any token with a VOT value less than 20ms as a /b/ nearly 100% of the time, while tokens with a VOT value greater than 20ms will be categorized as a /p/ nearly 100% of the time.

However, subsequent research using more fine-grained measures, such as eye tracking and mouse-cursor tracking, revealed that listeners were indeed sensitive to within-category variation of acoustic cues. For example, an experiment by McMurray, Tanenhaus, and Aslin (2002) tasked participants with selecting the referent of a spoken word from among four pictures while their eye movements were recorded. Two of the

pictures—the "critical" stimuli—corresponded to words that could be distinguished along the dimension of VOT (e.g. "bear" and "pear"), while two were "distractor" images, with names that were not phonetically similar to each other or to the critical stimuli (e.g. "lamp" and "ship"). The acoustic stimuli on the basis of which participants responded consisted of tokens that varied in VOT value from 0ms (an unambiguous "bear") to 40ms (an unambiguous "pear") in increments of 5ms. Their results showed that acoustic tokens closer to the middle of the VOT continuum resulted in a greater number of fixations to the unchosen alternative. For example, a token with a VOT of 5ms resulted in more fixations to "pear" than a token with a VOT of 0ms, despite that fact that participants identified the word as "bear" nearly 100% of the time in both cases. Results such as this have now conclusively shown that listeners are indeed sensitive to fine-grained changes in acoustic information, and may use this information to guide comprehension.

In light of evidence for sensitivity to within-category acoustic differences, it is no longer defensible to think of the speech processing system as implementing a simple category boundary. A more tenable hypothesis is that the speech processing system can be characterized as a fuzzy category membership function. A very simple version of such a function for a monolingual English speaker, dealing only with the distinction between /b/ and /p/ along the VOT dimension, could consist of two Gaussian membership functions characterized by means, standard deviations, and amplitudes, as shown in the top-left panel of Figure 4.1. In this plot, the red Gaussian corresponds to a membership function for the English /b/, centered on a VOT of 0ms, while the blue Gaussian corresponds to a membership function for the English /p/, centered on a VOT of 40ms. Given any point on the VOT continuum as input, each of the two membership functions will output a value indicating the likelihood that the token belongs to that category. One can then compute the posterior probability of any token along the VOT continuum as belonging to a reference category as the likelihood of the reference function relative to the sum of the likelihoods for both functions. This produces a predicted response curve, shown for an idealized monolingual English speaker and treating /p/ as the reference category, in the top-right panel of Figure 4.1. As this plot shows, an English participant using such a categorization function would be predicted to categorize tokens as /p/ the majority of the time when they fall $>$ 20ms on the VOT continuum, and as /b/ most of the time when they fall $<$ 20ms.

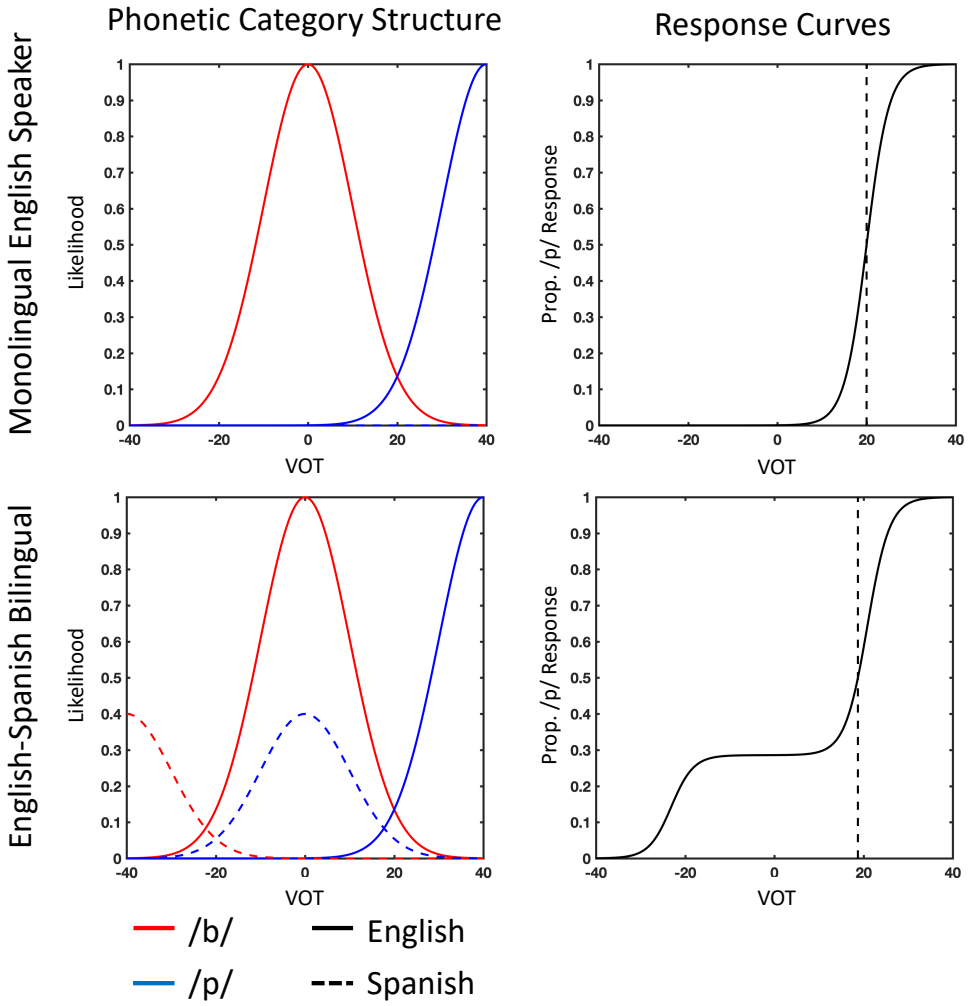


Figure 4.1: Simulated phonetic category structures (left column) and the resultant response curves (right column) for an idealized monolingual English speaker (top row) and an idealized Spanish-English bilingual (bottom row) with partial activation of Spanish phonetic categories. In the right column, dashed vertical lines indicate the /b/-/p/ category boundary.

4.1.1 Modeling Phonetic Categorization in Bilinguals

But what might the membership function look like for an Spanish-English bilingual? There are several possibilites. One is that bilinguals may be well-described using just two membership functions, as expected for monolinguals, but that the centers of these functions may differ from monolinguals. This would mean that bilinguals have just one pair of categories that they use for both languages. Given that the distinction between voiced and voiceless tokens is negatively shifted in Spanish relative to English, we might predict a corresponding negative shift in the centroids of the membership functions. This would result in Spanish-English bilinguals having a sharp

category boundary similar to the top-right panel of Figure 4.1, but with the category boundary falling between the locations observed for typical monolinguals in either English and Spanish. We might also hypothesize that the degree of negative shifting of categories would be dependent upon the relative dominance of English and Spanish for each individual, with those whose first language (L1) is Spanish having a boundary somewhere left of the boundary of an L1 English speaker.

Alternatively, we might hypothesize that Spanish-English bilinguals have *separate* phonetic categories for each language, and thus are best characterized by *four* membership functions, with one pair of /b/ and /p/ membership functions for each language. In this case, we might also think that the relative amplitude of the membership functions for each language to vary according to language experience. This possibility is illustrated in the bottom-left panel of Figure 4.1, with dashed curves corresponding to Spanish phonetic categories. In this plot, the Spanish categories have lower amplitude than the English categories, which may be predicted for a bilingual whose L1 is English. In Bayesian terms, this corresponds to a reduction in the prior probability of observing Spanish categories, relative to English categories. As shown in the bottom-right panel of Figure 4.1, this would result in a very different shape for the predicted behavioral response curves, though the category boundary (the point at which /p/ responses cross 50%) has shifted very little.

4.1.2 Context-Sensitive Categorization

To further complicate this picture, it must also be considered that listeners could flexibly adjust their perception based on the linguistic context (Repp & Liberman, 1987). Language processing is highly interactive (J. B. Falandays, Batzloff, Spevack, & Spivey, 2020; S. C. Spevack, Falandays, Batzloff, & Spivey, 2018), meaning that information corresponding to different "subsystems" of language (e.g. phonetics, syntax, semantics) and even from non-linguistic systems (e.g. vision, motor systems) is rapidly integrated. In the case of bilinguals, classic work by Grosjean and Nicol (2001) established that bilinguals have "language modes," meaning that Spanish-English bilinguals may be able to suppress their knowledge of English when they are in a Spanish context, and vice versa. However, bilinguals also show effects of phonological competition from words in a language other than the one established by the current context (Marian & Spivey, 2003a, 2003b), suggesting that bilinguals may never *fully* suppress either of their languages. Thus, bilinguals are often thought of as changing the relative "activation" of each language, which has been modeled in connectionist networks such as the Bilingual Interactive Activation (BIA; Dijkstra & Van Heuven, 2002) through the use of high-level context nodes for each language that selectively activate words in a corresponding language, and inhibit words in other languages. If indeed contextual cues can modulate the relative activation of English/Spanish linguistic categories, then we might predict language context to influence the shape/boundary of behavioral response curves. If bilinguals are hy-

pothesized to have just one pair of categories (as in the top row of Figure 4.1), these categories could perhaps move around flexibly depending upon the language context. Alternatively, if bilinguals are hypothesized to have two pairs of categories (as in the bottom row of Figure 4.1), language context might influence the relative amplitudes of the membership functions.

Psycholinguists interested in bilingualism have been aware of the possibilities reviewed above for several decades. For example, Williams (1977) also considered how various categorization functions would influence the shape of behavioral response curves and the locations of category boundaries. The results of this study showed that the Spanish-English bilinguals all had steep, monotonic response curves (as in the top-right panel of Figure 4.1), with some bilinguals having a category boundary near the expected location for monolingual Spanish speakers, and others having a boundary near the expected location for monolingual English speakers. Furthermore, boundaries did not appear to shift based on adjusting the expectation of being in an English vs Spanish context. However, it should be considered that this experiment had a quite small sample size (N=8), and the analysis of response curve shapes was purely qualitative, based on visual examination. Another experiment in the same year from Elman et al. (1977) reached a different conclusion, with Spanish-English bilinguals reliably categorizing tokens in the middle of the VOT continuum as /b/ when in an English context, suggesting that their category boundary was positively shifted relative to when they were in a Spanish context. Consistent results were reported by Flege and Eefting (1987), who found a very small (2ms) shift in phonetic category boundaries as a function of language context in Dutch-English bilinguals.

4.1.3 Beyond Category Boundaries

Since the time of the studies reviewed above, research on speech perception in bilinguals seems to have moved on to other issues, such as how bilinguals deal with lexical ambiguity (B. Falandays & Spivey, 2020). As such, the question of how to best characterize the structure and flexibility of the phonetic categorization process in bilinguals remains unresolved. Part of the reason for this may be the heavy reliance on category boundaries in previous work. As the right column of Figure 4.1 shows, alternative structures for representing phonetic categories may result in very different shapes of predicted response curves, while changing category boundaries very little. In the absence of advanced statistical and computational tools for analyzing the shapes of response curves, previous research may have been stalled. Furthermore, even many modern statistical techniques rely on parametric models that are highly constrained in the possible shapes they may take on, which may make them unable to fit the range of response curves that would be possible given an unknown mixture of membership functions.

In the present work, the limitations of previous research were avoided by using the

non-parametric least squares analysis to fit human behavioral responses to simulated response curves. This method allowed for the possibility that categorization responses of Bilinguals are best described by non-monotonic curves that cannot be produced by parameterized models, such as the logistic functions that are often used in related work. Furthermore, the category structures that produced our simulated response curves allowed Spanish and English phonetic categories to vary in their amplitudes, which may reflect the relative "activation" of each language. As a result, we may use the best-fitting model for each participant to infer the activation of English and Spanish categories, which can then be compared across language contexts to examine the flexibility of category representations in bilinguals.

In what follows, we first describe the computational model used to generate possible categorization functions and corresponding behavioral response curves. Then, we describe an experiment deployed to obtain categorization data from Spanish-English bilinguals and monolingual English speakers. Finally, we report the results of our analysis, which provide support for the hypothesis that bilinguals are best characterized as having distinct sets of categories for each language, and that these categories can be flexibly adjusted based on cues to immediate linguistic context.

4.2 Computational Model

We first simulated a large number of possible response curves, derived from mixtures of four Gaussian membership functions that could vary in their means and amplitudes. Each simulation consisted of two pairs of membership functions: one pair corresponding to English /b/ and /p/, and another pair corresponding to Spanish /b/ and /p/. The amplitude for the English pair was varied from [0, 1] in increments of .1, while the amplitudes for the Spanish pair were always equal to 1 minus the English amplitude (e.g. if English categories had an amplitude of .7, Spanish categories had an amplitude of .3). This is intended to capture the possibility that bilinguals may adjust the relative activations of their two languages or, in Bayesian terms, the prior expectation of observing tokens from each language. Importantly, because the amplitude of either language-pair may take on a value of 0, this allows for the possibility that categorization behavior is best described by just two Gaussian membership functions (as in the top-left panel of Figure 4.1), rather than four (as in the bottom-left panel of Figure 4.1).

The means for each category in each language were allowed to vary independently. Means were varied in increments of 5ms along the VOT continuum, with different allowable ranges for each category, corresponding to reasonable values for each category based on human production data (Lisker & Abramson, 1964). The Spanish /b/ category was allowed to vary from [-55, -25] in increments of 5, while the Spanish /p/ mean was allowed to vary from [-15, 15]. The mean for English /b/ was allowed to vary from [-15, 15], while the English /b/ was allowed to vary from [25, 55]. These

resulted in 11 (relative amplitudes) X 7 (English /b/ means) X 7 (Spanish /b/ means) X 7 (English /p/ means) X 7 (Spanish /p/ means) = 26,411 possible parameter combinations. However, when either pair of categories had an amplitude of 0, this pair was effectively non-existent, and thus adjusting the means of the non-existent categories resulted in redundant category structures. After removing these redundant options, there remained 21,707 possible parameter combinations.

From these simulated category membership functions, we produced predicted response curves for the proportion of /p/ responses. This was done by summing the likelihood of English and Spanish /p/ categories, then dividing by the total likelihood summed across all four categories. This results in a posterior probability of categorizing a token as a /p/ (regardless of whether it is a Spanish or English /p/) for each point considered along the VOT continuum. This function was evaluated at nine, evenly spaced points from a VOT of [-20, 40]ms (increments of 7.5ms), corresponding to the acoustic stimuli presented in the human subjects experiment.

4.3 Experimental Method

The human subjects experiment consisted of a 2-alternative forced-choice categorization task, in which participants heard synthetic stop consonants that spanned a VOT continuum. On each trial, participants saw two images which were exemplars of nouns that formed a minimal pair. A minimal pair is a pair of words that differ in just one phonological dimension, which in this case was VOT. Thus, one image was an exemplar of a noun beginning with /b/ (e.g. "bear"), and another corresponding to a noun that began with /p/, but matched the first word in subsequent segments (e.g. "pear"). Participants heard one token at a time, and were instructed to click on the image with the name that began with the consonant they heard. We recorded behavioral responses, reaction times, and mouse-cursor trajectories. Bilingual participants were induced into either an English or Spanish "language set" by virtue of a pre-experiment language survey presented in one of the two languages. Experimental instructions were also presented in the corresponding language, and the images chosen for each language condition corresponded to either English or Spanish minimal pairs.

4.3.1 Participants

142 healthy undergraduate students were recruited from the subject pool of University of California, Merced. Of these participants, 44 were self-reported monolingual English speakers, and 98 were self-reported Spanish-English bilinguals. Bilinguals were randomly assigned to complete the experiment in either a Spanish language context (N = 45) or an English language context (N = 53). Participants provided

informed consent in accordance with IRB protocols and received course credit for their participation. Participation was restricted to those who reported having normal or corrected-to-normal vision and hearing.

4.3.2 Materials

4.3.2.0.1 Auditory Stimuli. The auditory stimuli used in this experiment consisted of nine synthetic consonant-vowel tokens, generated using the Klaat synthesizer in the Praat software package. The VOT of tokens was varied from [-20, 40]ms in increments of 7.5ms, resulting in 9 tokens ranging from a clear /ba/ to a clear /pa/.

4.3.2.0.2 Visual Stimuli. The visual stimuli used in this experiment were simple, colored cartoons representing common English or Spanish nouns that began either with /b/ or /p/. Each image was displayed at a size of 300 X 300 pixels. To ensure that images were reasonable referents for the target nouns, several possible images were first selected for each noun. Then, a team of 3 researchers examined the images and selected the most canonical exemplar. For images corresponding to Spanish nouns, Spanish-English bilinguals were consulted to verify both that the nouns were common ones, familiar to all Spanish speakers, and that the images were good exemplars of these nouns.

For each language, 8 separate images were chosen that began with each phoneme. For each image representing a noun that began with a /b/, a corresponding image was chosen to represent a noun that began with /p/, but was otherwise identical (i.e. "minimal pairs"). For example, in the English condition, an image of a "bear" was matched with an image of a "pear." In the Spanish condition, an image of a "bote" (boat) was matched with an image of a "pote" (jar).

4.3.3 Apparatus

The experiment was implemented using the Psychophysics Toolbox package in Matlab, and run on an Apple iMac computer. Participants used a standard computer mouse to make behavioral responses by clicking the left mouse button. Mouse cursor movements were recorded using custom scripts that sampled the position of the mouse cursor at a rate of 100Hz. Participants wore acoustically-open, high-quality headphones, with the volume set to the maximal level that was comfortable.

4.3.4 Procedure

Participants completed the experiment individually in the lab over 30 minutes. Prior to beginning the experiment, bilingual participants first completed a subset of the LEAP-Q language experience survey. Bilinguals assigned to a Spanish language context completed the survey in Spanish, while those assigned to an English context completed it in English. After completion of the survey, participants were moved to a private testing room to complete the experiment.

The experiment began with participants reading a set of instructions, displayed either in English or Spanish, depending upon the condition. Then, participants were first familiarized with each image in their experimental condition, along with its intended name displayed visually beneath the image. Participants were given the opportunity to review the images/names more than once, if they chose.

Then, participants began a practice phase consisting of 16 trials, exposing them to each matched-pair of images twice. On each trial, one image (e.g. bear) first appeared in the top-left corner of the screen, and its matched image (e.g. pear) appeared in the top-right. After clicking a "start" button at the bottom-center of the screen, a single letter ("b" or "p") appeared in the center of the screen, and participants were instructed to click the image which had a name beginning with that letter.

After completing the practice, participants entered the main phase of the experiment. Participants were presented 3 times with each combination of 9 acoustic tokens, 8 image-pairs, and 2 orderings of the images (e.g. bear in the top-left/pear in the top-right, or vice versa), for a total of 432 trials, in random order. The only difference between the main phase and the practice phase was that, instead of displaying a letter in the center of the screen as in the practice phase, participants heard an acoustic token presented over headphones. Participants were instructed to move their mouse cursor as quickly as possible, while maintaining accuracy, to select the image with the name that corresponded to the acoustic token that they heard.

4.4 Results

4.4.1 Category Boundaries

We first considered the categorization response curves (shown in the aggregate in Figure 4.2), which has been the primary data of interest in related prior work. To recover category boundaries of each participant, the response data for each individual was fit to a four-parameter logistic function. The mean category boundaries were as follows: Monolingual English = 21.44ms ± 4ms; Bilinguals in the English condition = 18.16ms ± 5.87ms; Bilinguals in the Spanish condition = 17.17ms ± 6.14ms.

These data were analyzed using linear regression, with language condition treated as a categorical variable and the monolingual English group as the reference level. This analysis revealed that monolingual English speakers had a significantly different boundary than bilinguals in both the English (b = -3.292, SE =1.11, t = -2.96, p ¡ .01) and Spanish conditions (b = -4.386, SE =1.16, t = -3.79, p ¡ .001). A post-hoc test did not detect a significant difference between the perceptual boundaries for bilinguals in the English versus Spanish conditions (b = -1.09, SE =1.21, t = -.897, p = .372).

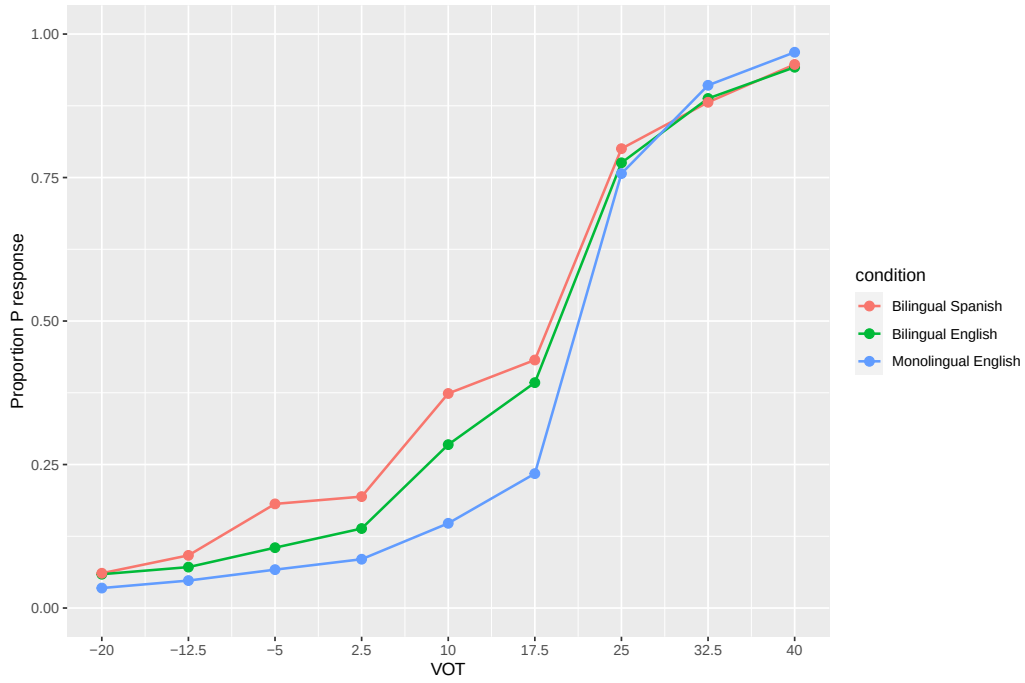


Figure 4.2: Average proportion of /p/ responses as a function of VOT and language condition. Monolingual English speakers are shown in blue. Bilinguals in an English context are shown in green, with bilinguals in a Spanish context shown in red.

4.4.2 Non-parametric Model Fitting

While the previous analysis revealed only a modest difference in category boundaries between monolingual English speakers and Spanish-English bilinguals, and no effects of language condition on bilinguals, this may be due to the limitations of parametric models such as the four-parameter logistic to fit the possible shapes of response curves. The four-parameter logistic is characterized by a floor, a ceiling, a crossover point, and a slope. As such, it would be unable to provide an accurate fit for the non-monotonic response curves that are possible given a categorization function that is a mixture of more than two categories (as in the bottom-left panel of Figure 4.1).

To circumvent this issue, we used the χ^2 goodness-of-fit test to compare the proportion of /p/ responses for each VOT value, for each participant, with the proportions predicted by each of our 21,707 simulated response curves. By selecting the model with the lowest χ^2 value for each participant's data, we are able to infer the centroids of each category (/b/ and /p/) for each language (English or Spanish), as well as the relative activations of each language. Because either language may have an activation of 0, and therefore be functionally absent, this allows for the possibility that some participants are best fit by just two categories, while others are best fit by four.

We next extracted the activations of the English and Spanish categories according to the best-fitting model for each participant, which is displayed in Figure 4.3. This plot shows that monolingual English speakers are clustered in the bottom-right corner, indicating that these participants were best fit by models with little to no activation of Spanish categories, as would be expected. Bilinguals, meanwhile, are best fit by models with greater activation of Spanish categories. Furthermore, there is a noticeable increase in the amount of activation of Spanish categories for those in the Spanish condition (red) compared to those in the English conndition (green). Importantly, while this was a between-subjects design (each participant completed the experiment in only one language "set"), an analysis of our language survey data revealed no significant difference in Spanish proficiency or age-of-acquisition between groups. In fact, bilinguals in the English condition were slightly more proficient and had an earlier age-of-acquisition of Spanish, on average. This suggests that these differences between conditions are not attributable to language experience. To further validate the accuracy of our model-fitting, we also compared the perceptual boundaries obtained from fitting a four-parameter logistic function to each participant's data with the perceptual boundaries corresponding to the best-fitting parameter combination of our four-category categorization functions. The boundaries obtained using these two models were correlated at .95, indicating that our category model provides similar estimates of perceptual boundaries as the more standard procedure.

The mean values for the activation of Spanish categories was as follows: Monolingual English = .11 ± .06; Bilinguals in the English condition = .14 ± .1; Bilinguals in the Spanish condition = .22 ± .15. These data were also analyzed using linear regression, with language condition treated as a categorical variable and the monolingual English group as the reference level. This analysis revealed that monolingual English speakers had a significantly lower activation of Spanish categories than bilinguals in the Spanish condition (b = .111, SE = .023, t = 4.74, p ¡ .001), but did not differ significantly from bilinguals in the English condition (b = .027, SE = .022, t = 1.19, p =.236). A post-hoc test showed that bilinguals in the English condition also had significantly lower activation of Spanish categories than bilinguals in the Spanish condition (b = .084, SE = .025, t = 3.3, p ¡ .01).

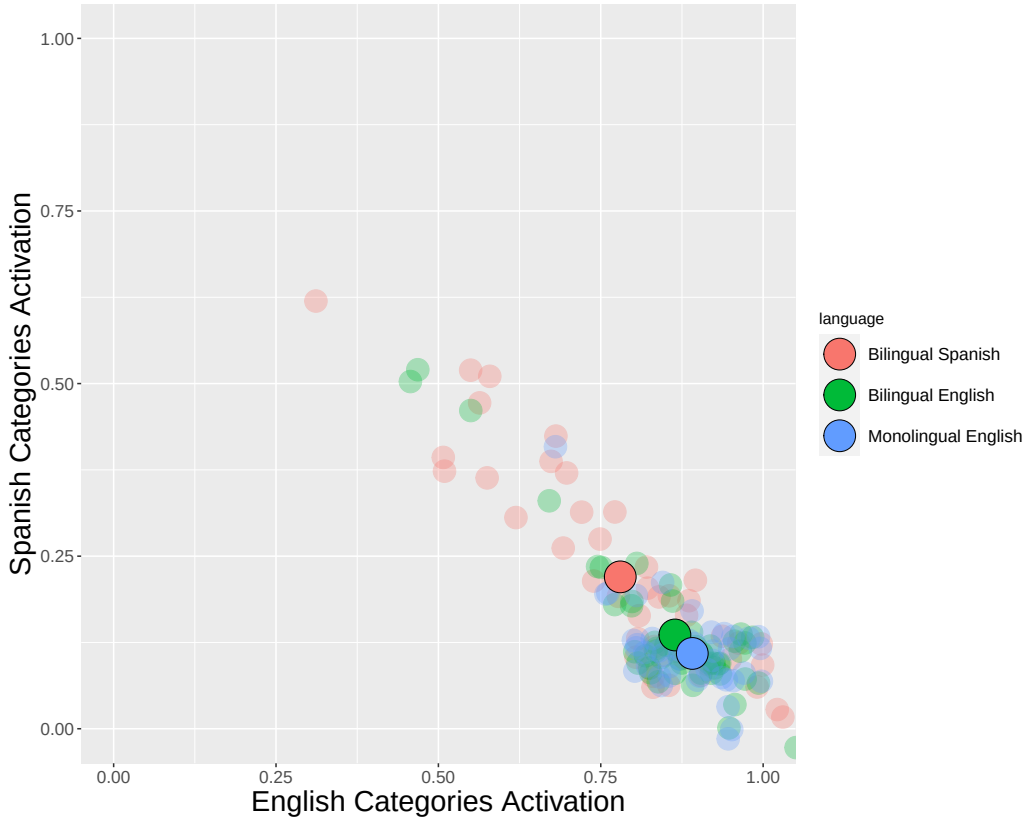


Figure 4.3: The relative activations of English (x-axis) and Spanish (y-axis) categories. Individual points (slightly jittered for visualizability) are shown for each subject, while group means are shown in larger, black-outlined points.

4.5 Discussion

Bilinguals must contend with potential inter-language phonetic ambiguity, such as the fact that the English /b/ is acoustically similar to the Spanish /p/. We considered two major possibilities for how category representations in Spanish-English bilinguals could be constructed to cope with this ambiguity: bilinguals may just have one set of categories for each language, or they may have distinct sets of categories for each language. Furthermore, we considered whether bilinguals may be able to flexibly adjust their categorization functions based on linguistic contexts.

Our human subjects experiment first replicated the qualitative patterns observed in previous work: Spanish-English bilinguals show a perceptual boundary that is between the locations expected for typical monolingual English and Spanish speakers, and this boundary shows small but significant effects of language context. However, relying merely on perceptual boundaries, or other features that can be detected with parametric models such as the four-parameter logistic function, may disguise greater

potential differences between bilinguals and monolinguals, or between language contexts. This is due to the fact that, if bilinguals indeed are best characterized as having distinct categories for each language, this may result in non-monotonic response curves that cannot be well fit by parametric functions.

To deal with this issue, we built a computational model implementing a category membership function with four available categories (a /b/ and /p/ category for each of two languages). By systematically varying the parameters of this model, we simulated a diverse array of possible response curves, and then chose the best-fitting parameter combination for each participant. The resultant fits provided very similar estimates of category boundaries as did the four-parameter logistic function. However, despite only a modest effect of language context on the perceptual boundaries of participants, our analyses revealed that the data from bilinguals in a Spanish language context, relative to an English context, was best fit by categorization functions with stronger activation of Spanish categories. Meanwhile, bilinguals in an English language context were best fit by functions with low levels of activation of Spanish categories, and appeared quite similar to monolingual English speakers in this regard. These results provide converging evidence that bilinguals can be characterized as having distinct phonetic categories for each of their languages, and that the relative activation of categories in each language can be flexibly adjusted based on even weak cues to linguistic context.

Chapter 5

The Emergence of Cultural Attractors: How Dynamic Populations of Learners Achieve Collective Cognitive Alignment

All human groups possess group-specific behavioral repertoires involving cultural variants—things such as tools, linguistic behavior, social norms, religious beliefs, and artistic styles. As cultural variants are observed and copied, they are liable to change over time as a result of noise, errors, and biases in both transmission and interpretation. However, even in the absence of strong selection for specific outcomes, cultural variants may nevertheless converge over successive transmission events toward culture-specific "attractor" points (Sperber, 1996). This effect can be attributed, at least in part, to the fact that individuals within a cultural group often share similar cognitive biases, such that they tend to perceive, remember, and reproduce information in consistent ways (Heyes, 2018). Without this "cognitive alignment," cultural transmission would be far less reliable, and the potential for cumulative cultural evolution would be limited.

But how does cognitive alignment first emerge in initially-uncoordinated, dynamic populations? Current models of cultural evolution usually take cognitive alignment as given. This is implicitly the case in most models based on the mathematics of population genetics or epidemiology, which assume high-fidelity transmission (Acerbi, Mesoudi, & Smolla, 2020; Boyd & Richerson, 1988; Cavalli-Sforza & Feldman, 1981; Mesoudi, 2021), and explicitly the case in most models of cultural attraction, in which any attractors are assumed to be both stable and universally shared throughout the population (Acerbi, Charbonneau, Miton, & Scott-Phillips, 2019; Claidière & Sperber, 2007; Mesoudi, 2021; Rafał, 2018). There are, to our knowledge, no models that demonstrate how attractors arise. This is an important gap in theory in light of

the many cases where culture depends on cognitive biases that are themselves socially acquired (Heyes, 2018; Karmiloff-Smith, 1994), and therefore not guaranteed. Within shifting populations of cognitively-plastic individuals, cognitive alignment may need to be actively and continuously maintained in order for cultural knowledge to be successfully preserved across generations.

In this paper, we develop an agent-based model of the emergence and maintenance of collective cognitive alignment in dynamic populations. Our model adapts an existing model of unsupervised learning of phoneme categories in individual learners (Toscano & McMurray, 2010) to a multi-agent, sociocultural setting wherein individual language learners provide the training input to each other. Agents attempt to use their limited cognitive resources to capture the distribution of sensory signals they observe from neighbors, then use their idiosyncratic perceptual representations to generate new signals. Beginning from a state in which all agents possess a set of randomly distributed categories of uniform probability, under some conditions populations self-organize into signal clusters, which constitute an identifiable set of cultural attractors. These cultural attractors may be thought of as akin to proto-linguistic units, such as a set of phonemes, but also may be taken to represent any culturally-shared repertoire of categories or behaviors. We explore the role of various innate cognitive constraints, levels of transmission error, learning periods, lifespans, population sizes, and network structures to understand when population-level structure may emerge, what properties it is likely to have, and how stable it is.

Our explorations with this model suggest that achieving and maintaining cognitive alignment may depend upon a finely tuned balance of factors at the levels of cognition, development, and demographic structure. We highlight three interesting and potentially counter-intuitive behaviors exhibited by our model that are not accounted for in other models of cultural evolution: First, we find that some noise is beneficial to stabilizing cognitive alignment. Second, we find that long learning times may destabilize and limit the complexity of cultural repertoires, while critical or sensitive periods of learning enhance stability. Third, we find that larger populations develop less complex, but more stable patterns of alignment, and that this effect can be moderated by network structure. These results suggest that additional complexity may be needed in models of cultural evolution to adequately understand how human-level culture can get off the ground and develop. We conclude by highlighting several ways that our model may be extended to complement existing models of cultural evolution and gene-culture co-evolution.

5.1 Why We Need More Models of Cultural Attraction

In research on cultural evolution, there has been a historical and theoretical divide between approaches that emphasize information preservation, and those that emphasize information transformation (Buskell, 2017). The preservative approach can be identified with Darwinian selectionist theories of culture, which tend to focus on the fitness consequences of cultural phenotypes, and to treat transmission as analogous to biological inheritance with noise (Boyd & Richerson, 1988; Cavalli-Sforza & Feldman, 1973, 1981; Dawkins, 1976; Smaldino, 2014). This often reflects a modeling simplification rather than a deep assumption about the intrinsic nature of cultural transmission, as simplifying assumptions are needed to advance theory (Healy, 2017; Smaldino, 2017). However, some researchers have argued that high-fidelity copying is more than just a simplifying assumption but in fact one of the keys to cumulative cultural evolution (H. M. Lewis & Laland, 2012), bolstering this claim with evidence from cross-species studies showing that humans are exceptional- or even *over*-imitators, often copying observed actions even when they are causally irrelevant to an outcome (Hoehl et al., 2019; Horner & Whiten, 2005). In sum, the idea of high-fidelity copying has played a substantial role in explanations of the human capacity for cumulative cultural evolution.

The transformative approach, in contrast, can be identified with Cultural Attractor Theory (CAT), which emphasizes the fact that individuals have potentially-idiosyncratic cognitive biases in how they process and reconstruct cultural variants, such that cultural transmission may not conform to the predictions of a gene-like inheritance system (Claidière, Scott-Phillips, & Sperber, 2014; Scott-Phillips, Blancke, & Heintz, 2018; Sperber, 1996). The distribution of cognitive biases in a population can be thought of as comprising a "cultural attractor landscape," whereby some transformations of variants are more likely than others. An early example of this phenomenon is Bartlett's (1932) classic "War of the Ghosts" study, in which English participants read a Native American (Chinook) folktale and then, after various time delays, attempted to recall the content. Bartlett found that those story elements that were inconsistent with the "cultural schema" of the participants (that is, the narrative patterns with which they were familiar) tended to be forgotten or transformed into more familiar forms, especially as the time delay increased. When culture-specific cognitive biases of this kind are applied iteratively in social transmission, variants may converge towards group-specific equilibria points in the space of possible features, known as cultural attractors. This phenomenon has been demonstrated with transmission chain studies using images (Bartlett, 1932), event descriptions (Mesoudi & Whiten, 2004), music (Ravignani, Delgado, & Kirby, 2016), grammars (Kirby, Cornish, & Smith, 2008), tools (B. Thompson & Griffiths, 2021), and function concepts (Kalish, Griffiths, & Lewandowsky, 2007; for a review see Miton & Charbonneau, 2018).

It is increasingly recognized that there is room, and indeed need, for both approaches (Buskell, 2017; Mesoudi, 2021). Yet, in spite of this nominal consilience, little traction has been gained towards developing a theory that integrates both preservative and transformative factors in cultural evolution. For example, a 2015 review by Acerbi and Mesoudi reported only one known empirical study designed to address both selection and attraction effects simultaneously. This represents a crucial missing link in the literature, given that these two approaches do not deal with neatly-separable scales of analysis (Wimsatt, 1972).

A major barrier towards the fruitful interaction between these two perspectives is a dearth of formal models of cultural attraction. Cultural attractors are said to be statistical abstractions, and therefore to be the primary phenomenon in need of explanation (Scott-Phillips et al., 2018), yet there are no mechanistic models of how cultural attractors form, stabilize, or change over time. The few computational models of cultural attraction that exist instead make the assumption of pre-existing, stable cultural attractor points (Acerbi et al., 2019; Acerbi, Charbonneau, Miton, & Scott-Phillips, 2021; Claidière & Sperber, 2007; Mesoudi, 2021; Rafał, 2018). The cognitive or ecological forces that determine attractor points are assumed to be shared across members of a population from the outset, and stable across generations of individuals. While these models have been useful in showing how the presence of cultural attractors can influence the distribution of cultural variants over time, they are agnostic with respect to how attractors initially form or potentially change over time.

In this paper, we model cultural attractors as arising from the collective alignment of cognitive landscapes within a population (See Fig. 5.1). A cognitive landscape refers to a particular way of parsing the sensory world, storing information, and generating behaviors, which determines the transformation one individual will apply when reproducing a cultural variant from a model. When many individuals within a population develop aligned cognitive landscapes, social transmission becomes more reliable because many different individuals apply convergent transformations to information upon reproduction, allowing cultural variants to cluster into distinct types. In some cases, the alignment of cognitive landscapes in a population may be the result of highly-canalized developmental trajectories driven by genetic evolution. However, many important aspects of human culture rely on cognitive biases that are themselves socially transmitted. As such, our goal in this paper is to offer an account of the emergence of cultural attractors specifically in cases for which there are not yet shared, innate cognitive attractors.

In order to advance this argument, we first present evidence that cultural attractor theory supports a Darwinian view of cultural evolution. Next, we consider several possible mechanisms of attraction, including evolved cognitive biases and shared ecological constraints, but emphasize the importance of collective cognitive alignment through enculturation. Then, we describe how culturally-shared cognitive biases could emerge in a cognitively-dynamic population. Finally, we support our theory with an agent-based model that can account for the emergence of cultural attractors through

the lower-level interactions among cognitive agents, without appealing to selectionist principles.

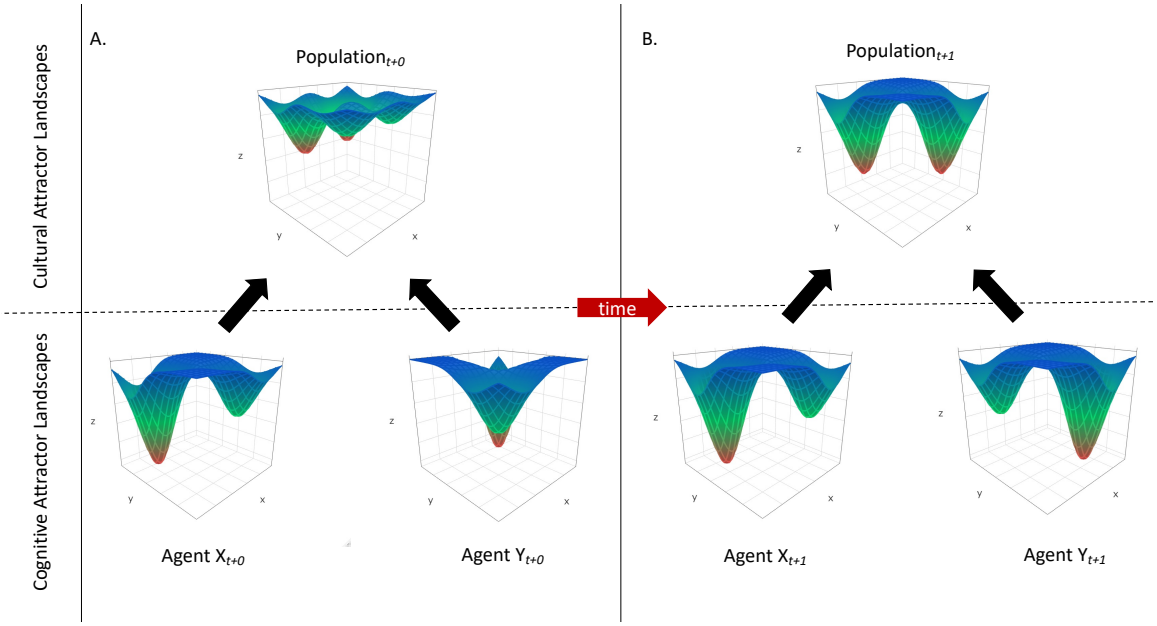


Figure 5.1: A simplified illustration of the feedback loop between cognitive and cultural attractor landscapes. An attractor, generally speaking, is a function describing the rate and direction of change of some variable(s), which can be visualized as a hypersurface. Valleys in an attractor landscape correspond to local equilibria towards which outputs converge over time, with the strength of attraction represented by the steepness of the valley. Here the x- and y-axes may represent any two dimensions of variability in a cultural variant (e.g. length and width of an arrow head; speech cues such as voice-onset-time and fundamental frequency). A cognitive attractor landscape (lower panels) gives the expected transformation that one individual will apply when attempting to reproduce a cultural variant from an observation. In the lower panels we show the cognitive landscapes of two individuals, each containing two attractors of differing location and strength. Multiple cognitive landscapes can be averaged to produce a cultural attractor landscape (upper panels) that gives the expected change in a distribution of cultural variants over the course of multiple transmissions within a population. Panel A. shows a situation in which two individuals possess disaligned cognitive landscapes, resulting in rugged cultural landscape with four weak attractors. Panel B. shows the same two individuals at a later time, with Agent Y having more-closely aligned their cognitive landscape to individual X, resulting in a smoother cultural landscape with two strong attractors.

Table 5.1: Key Concepts

Key Concept	Definition

Cultural Variants	Behaviors or artifacts generated by individuals in a cultural population.
Cognitive Landscape	A cognitive function giving the probability of different outputs (e.g. neural states, behaviors, cultural variants produced) for an individual given some range of inputs. Represents the cumulative effects of sensation, perception, memory, attention, motor control, and any other cognitive processes that shape how an individual responds to stimuli and generates new behaviors.
Cognitive Attractors	Local minima in a cognitive landscape, corresponding to outputs that are more likely for an individual in general, or more likely in response to some particular input.
Cultural Landscape	A function describing the probability of observing different cultural variants at a population level. Represents the aggregate result of a population of cognitive landscapes, plus patterns of social interaction and any ecological factors that influence the observation and reproduction of cultural variants.
Cultural Attractors	Local minima in a cultural landscape, corresponding to high-probability variants for a population in general, or semi-stable equilibria towards which transformations converge given an initial distribution of variants.
Culture-Cognition Feedback Loop	The co-evolution of a cultural landscape with a population of cognitive landscapes. As each generation of individuals learns from exposure to a distribution of cultural variants, this may result in a change to the set of cognitive landscapes, in turn producing a new distribution of cultural variants in the next generation, and so on.
Collective Cognitive Alignment	The convergence of cognitive landscapes within a cultural group in the absence of innately shared cognitive attractors, such that group members tend to perceive, remember, and reconstruct information in convergent ways.

5.1.1 The Role of Cultural Attractors in Darwinian Cultural Evolution and Information Transfer

Cultural attraction theory is often framed as a critique or qualification of Darwinian selectionist models of cultural evolution, in which cultural variants are often modeled as discrete units that are more or less faithfully transmitted (similar to "memes" as described by Dawkins, 1976). But even as CAT challenges the assumption of high-fidelity copying, it simultaneously describes the conditions under which this assumption may be justified: when variants have converged to a cultural attractor point, such that subsequent transmission events no longer incur systematic deviations from a model. Prominent researchers associated with both Darwinian and CAT research camps have pointed to this complementarity between their approaches. Henrich, Boyd, and Richerson (2008) explain that Darwinian models of cultural selection are useful precisely *because* of the existence of cultural attractors (see also: Henrich & Boyd, 2002): in their model, so long as there is more than one attractor present in space of cultural variation, transmission errors will be corrected to some extent and

cultural phenotypes will cluster such that they can effectively be approximated as discrete traits. This stance puts Henrich et al. (2008) in agreement with Claidière et al. (2014), who argued that perfect replication is a special case of attraction: when cultural variants sit at local minima of a stable attractor landscape, there will be no bias in the transformation of the variant over repeated transmissions, allowing pure selection to dominate (see also: Claidière & Sperber, 2007). In this way, the existence of cultural attractors lays the foundation for cumulative culture.

Another way to understand the important role of cultural attractors in Darwinian cultural evolution relates to the capacity for information transfer. Consider that all information transfer presupposes a particular reference frame for distinguishing signal from noise in a continuous physical channel (Fields & Levin, 2020; Von Uexküll, 1934). All information transfer is "transformative" to an extent, in that any sender must apply some function for encoding messages into physical signals, and any receiver must apply some function for decoding messages *from* signals, with both processes inevitably subject to noise, however small. However, information can be preserved when senders and receivers share a reference frame, such that the transformations applied in encoding/decoding are convergent. For example, binary digital signals may be represented as voltages near 0 for *off*, and near 5 for *on*, perhaps using a simple threshold function (i.e. values below 2.5V are treated as *off*, and values above 2.5V are treated as *on*). Given noise, a sender may produce a voltage of $+1$ or -1 on different instances when trying to communicate an *off* message, but in both cases the signal will be compressed into an *off* message by a receiver (with the same reference frame) before passing the message along again, which prevents the accumulation of noise. However, if senders and receivers do not define the same set of signals over the communication channel and/or encode messages into signals using different functions, information will inevitably be destroyed in each instance of transmission. In this light, we may think of cultural attractors as reflecting a shared reference frame that allows cultural information to be preserved and potentially built-upon over time.

Imagine a first individual who invents a dance, focusing primarily on their fancy footwork. An observer with very different cognitive landscape may, frustratingly, fail to appreciate the first dancer's footwork at all, but instead attend to their arm movements, and therefore end up "recreating" a very different dance. A third individual may attend mainly to the second dancer's head movements, and so on. It is not that these individuals are copying inaccurately *per se*, but instead that they do not even agree on what it is they are supposed to copy. If we posit some cognitive function that transforms sensory signals into new behaviors—a cognitive landscape—these individuals have different, but equally valid, functions. In such a situation, there may be social learning occurring in some sense (or at least social influence), but variants would not be expected to cluster in any identifiable way, and indeed it would be hard even to say there *exists* any cultural variant to evolve (a point made also by Claidière & Sperber, 2007).

Conversely, when individuals within a group have highly-aligned cognitive landscapes,

productions of a cultural variant can differ substantially in "surface" characteristics while nonetheless retaining the same culturally-relevant core. Consider a participant in a Western population who is asked to draw a smiley face using red pen on a notepad, a second to copy this image using spraypaint on a wall, and a third to copy the second using lego blocks. In this case, they will all likely recognize each product as instances of the same culturally-shared category, despite variation in the medium. In most respects—except just those few culturally-relevant ones—these could be seen as "low fidelity" copies. However, the cultural core of these productions is not *in* the productions themselves, but instead is an abstract mental category shared across the individuals. When cognitive landscapes are aligned in this way, cultural transmission can occur with sufficient fidelity for selection to act on cultural variants in a Darwinian fashion.

5.1.2 Mechanisms of Convergent Transformation: The Importance of Collective Cognitive Alignment Through Enculturation

From our perspective, the most crucial insight of CAT is the point that social transmission is "reconstructive," meaning that cultural variants are not simply copied, but actively reproduced by individuals, influenced in the process by the memories, biases, and proclivities present in their minds (Claidière et al., 2014; Scott-Phillips et al., 2018; Sperber, 1996). As described by Sperber (1996), reconstructive transmission may produce convergent transformation patterns when there is a "convergence of [...] affective and cognitive processes [...] of many people towards some psychologically attractive type of views in the vast range of possible views." We refer to this convergence as "collective cognitive alignment." In cases where cognitive alignment depends upon enculturation through experience, it becomes possible that cognitive alignment may *fail* to be achieved either within or across generations. This motivates the need for computational models of cultural attraction such as ours, that do not make the assumption of pre-existing attractor points, and instead appeal to a culture-cognition feedback loop.

Collective cognitive alignment through enculturation is not the only possible mechanism of convergent transformation patterns. Shared ecological factors are likely to produce cultural attractors in some cases, ranging from norms of sharing in harsh, isolated climates (Gerkey, 2013) to color categories in environments dominated by correlated spectral patterns (Baronchelli, Gong, Puglisi, & Loreto, 2010). Some attractors may be driven by exogenous motivational factors, such as an imperial edict that results in widespread adoption of a particular hairstyle, upon penalty of death (Morin, 2016). And some cognitive attractors may even be universal as a result of genetic features under strong selection, such as an evolved salience bias for direct eye-contact leading to an increase in viewer-oriented figures over time in a portraiture tradition (Morin, 2013). However, humans exhibit tremendous cultural variation that

cannot be attributed merely to ecological factors. As evidence of the this, we could point to any example of warring, neighboring tribes that distinguish themselves with different cultural markers, languages, customs, and beliefs (Smaldino, 2019). Nor can we attribute this variability to genetic differences between populations, given that there is known to be more genetic variation within human groups than between (Lewontin, 1972). Therefore, we propose that it is critical to explain how cultural attractors may form as a result of the culture-cognition feedback loop in the absence of strong determination by innate biases or shared ecological factors.

5.1.3 The Problem of Collective Cognitive Alignment

The process of cognitively aligning to a cultural reference frame—that is, of acquiring a set of categories and cognitive biases specific to members of a cultural community—is often discussed as a purely individual-level learning process (Ashby & Maddox, 2005; Kuhl, 2000; Toscano & McMurray, 2010). The cultural background that provides the fodder for learning is assumed, at least by many cognitive scientists, to be generally stable. Individuals may vary, but will observe similar training data and ultimately develop similar cognitive landscapes. But cultural environments, and the shared categories associated with them, can change over generational or even intragenerational timescales. As such, cognitive alignment is an ongoing collective coordination problem, in addition to being an individual learning problem.

There exist several computational models of the emergence of category conventions in groups (Baronchelli et al., 2010; Ke, Minett, Au, & Wang, 2002; Kirby, 2001; Puglisi, Baronchelli, & Loreto, 2008; Reali, Chater, & Christiansen, 2018; Skyrms, 2010; Steels, Belpaeme, et al., 2005; reviewed in Kallens, Dale, & Smaldino, 2018). However, these models assume that agents come pre-equipped with a shared set of recognizable and producible cultural variants, such that social transmission has perfect fidelity. In some cases, shared, fixed sets of signal and meaning categories are explicitly pre-defined, as in Kirby's (2001) iterated learning model. Several models have considered the coordination of linguistic labels for perceptual categories (Baronchelli et al., 2010; Gong, Baronchelli, Puglisi, & Loreto, 2011; Puglisi et al., 2008; Steels et al., 2005), allowing perceptual categories to be flexibly adjusted through experience, and for new linguistic labels to be created. However, in these models, the signals (e.g. verbal labels) are still transmitted with perfect fidelity, implying a globally-defined set of signal categories that are available to everyone—a world of Platonic word forms. Even models that represent the possibility of transmission errors (Nowak & Krakauer, 1999; Nowak, Krakauer, & Dress, 1999) treat errors as confusions of one signal category for another, which again presupposes that individuals share a set of signal categories. While this modeling literature has produced many important insights, it does not address cases in which signal categories may be plastic and differ across individuals.

One attempt that begins to address the culture-cognition feedback loop is a model of phonemic evolution by B. Winter and Wedel (2016). In their model, two agents each possessed a mental model of the set of phonemes in their language, represented as labeled clusters of 2-D point exemplars stored in memory. As the two agents communicated by producing signals to each other under the influence of cognitive biases, each agent categorized and stored new exemplars received from their neighbor while prior exemplars decayed in memory. In the process, the agents' labeled clusters of exemplars drifted around the signal space, corresponding to the co-evolution of individuals' perceptual distinctions along with a shared lexicon. While this model is a strong step towards giving due diligence to the issue of cognitive alignment in cultural evolution, Winter and Wedel's (2016) agents *begin* each simulation aligned, and therefore their results can tell us little about how cultural attractors initially emerge. Furthermore, with just two agents interacting in a highly constrained manner, their model cannot address how an attractor landscape is generated and maintained within a dynamic population.

Populations in which cultural attractors emerge often involve non-static sets of individuals. Old members die or leave, while new members are born or arrive from elsewhere. Consider that young learners, by definition, contribute different cognitive biases to the cultural attractor landscape than seasoned "experts," such that deaths and departures of the old and an influx of new learners threaten to alter a cultural attractor landscape in potentially drastic ways. If too many learners enter the population too fast, or many experts suddenly die, a cultural attractor landscape can change or even dissipate (unless there are other stabilizing factors, e.g. mechanisms for external information storage). This is a central point of the "linguistic niche hypothesis," which holds that languages adapt to their learners, in addition to the reverse process (Bentz & Winter, 2014; Dale & Lupyan, 2012; M. L. Lewis & Frank, 2016; Winters, Kirby, & Smith, 2015). For example, it has been proposed that as linguistic populations expand, they may incorporate a greater proportion of adult learners, causing pressures for language to change as a result of different cognitive biases between adults and children (Dale & Lupyan, 2012; Reali et al., 2018). Thus, language (and culture more generally) should not be thought of as information passively transmitted from one generation to the next, but instead as a complex adaptive system, wherein variants are products of individual cognitive landscapes, and individual cognitive landscapes are shaped by experience with other variants (Enfield, 2014; Group" et al., 2009).

In summary, we argue that understanding how cultural attractors can emerge and stabilize in the absence of innate cognitive attractors is an important step towards understanding the capacity for cumulative cultural evolution. Explaining complex processes requires mechanistic formalization (Epstein, 1999; Smaldino, 2017), but any initial formalization is likely to be incomplete, as models tend to accumulate nuance iteratively. Below, we present a model that we believe lays the groundwork for understanding the emergence of cultural attractors in the absence of strong determination from innate biases or shared ecologies. In a population of interacting,

cooperative individuals within a cultural community, it is reasonable to assume that mutual understanding is often, if not always, the goal of communication. Individuals will develop categories based on what is communicated to them, and use those categories to communicate similar information to others. We have argued that the existence of a shared cognitive biases is a prerequisite for treating cultural transmission as inheritance with noise, and so we do not appeal to selectionist principles in developing our theory. Instead, we model the intertwining of cognitive, communicative, developmental, and demographic dynamics. Because many mechanisms that allow for these dynamics are themselves evolved (e.g. learning periods and life cycles, social tendencies, neural structures), a full explanation must eventually reintroduce selection processes, but these we save for future work.

Our model currently offers only a general mechanism by which collective cognitive alignment may emerge through general principles of communication and learning, and should not be taken as mapping precisely onto specific empirical patterns. In other words, ours is a "how-possibly", rather than "how-actually", model (Craver, 2006). We see this model as complementary to the careful historical and anthropological work associated with CAT, which describes distinct instances of cultural attraction and identifies explanatory forces, and suggest that our model may be extended in future work to formalize how specific perturbations or parameters noted in the CAT literature can influence a cultural attractor landscape.

5.2 Model Description

Our model is intended to represent multiple generations in a population of individuals that interact and observe one another, implicitly shaping each other's cognition in the process. The basic requirements for modeling such a system include (1) a population of individuals, (2) a process whereby agents age and die, and new agents are born, (3) a mechanism for individuals to interact and observe one another, and (4) a representation of the systems that shape individuals' perception and production of information (i.e. cognitive landscapes), and (5) a mechanism for updating these representations based on experience (i.e. learning).

We begin with a population of N agents, arranged as nodes in an undirected network where edges represent opportunities for communicative interactions. Four network structures were explored, with the default being fully-connected (more details below). Each agent has an age, which is represented in our model as the number of time steps for which it has been "alive." All agents are initialized with an age of zero.

The model dynamics occurred in discrete time steps (illustrated in Figure 5.2), each of which consisted of two stages: *communication* and *reproduction*. In the communication stage, we iterate through agents in order of their position on the network, giving each a turn to communicate a signal to a randomly selected neighbor. Each

communicator randomly selects one neighbor for interaction (the receiver), and produces a signal, which may be distorted by noise. Receivers then learn something from the observed signal.

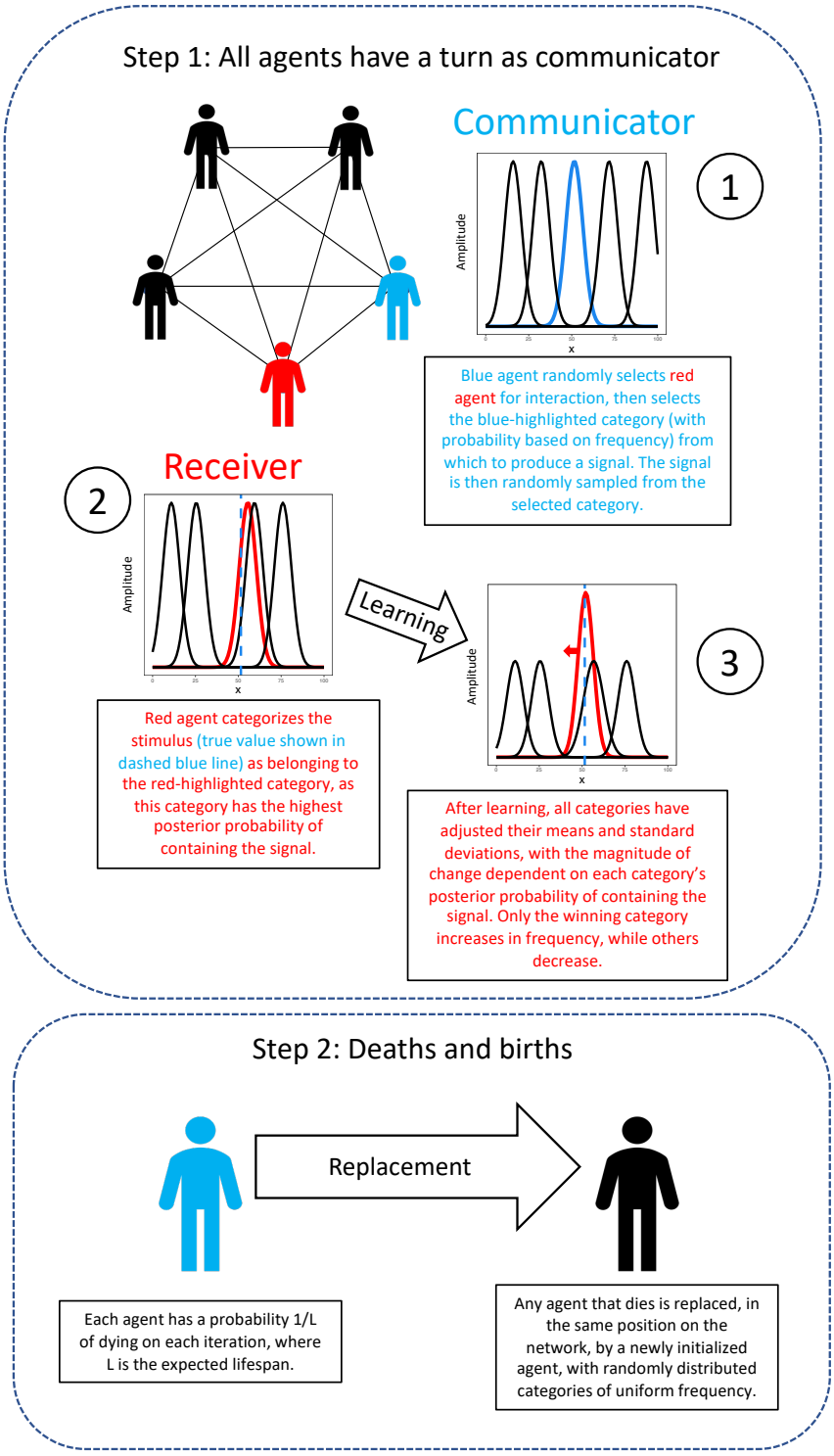


Figure 5.2: An illustration of the model dynamics.

A perceptual signal is some pattern of activity across the sensory receptors of an observer. Patterns of activity across n sensory receptors can be represented as points in an n-dimensional space. While the number of sensory receptors may be large, we assume that we can obtain a projection of this space onto 2-dimensions for plotting, which is commonly done in connectionist models of cognition and neuroimaging work, using mathematical tools such as principal components analysis. Thus, we represent signals as real-valued points on a 2-dimensional $S \times S$ square (we used $S = 100$). These axes could correspond to any featural dimensions which may be extracted by a category learning system, such as the voice-onset-time and fundamental frequency of a speech token (Toscano & McMurray, 2010), or the length and width of an arrow (Henrich et al., 2008).

Signal perception and production are both served by a learned representation of category structures. There are many ways to model category representation and learning. Here, we utilize an unsupervised, 2-dimensional mixture of Gaussians (MOG) model adapted from Toscano and McMurray (2010), which they found to effectively model the acquisition of phoneme categories in English. We expect this algorithm could be replaced by many cognitively-plausible models of categorization, including exemplar-based models (e.g. B. Winter & Wedel, 2016) or neural-network classifiers (e.g. Steels et al., 2005), without changing the overall picture. However, a MOG has useful mathematical properties, and can capture complex distributions with relatively few parameters, so it may be less computationally intensive than other models.

Each agent i possesses in memory a MOG of size $K = 20$, where each category k is defined as a two-dimensional Gaussian distribution with a mean μ_{ik}, standard deviation σ_{ik} (both mean and standard deviation are two-dimensional vectors), a correlation ρ_{ik} between dimensions (though for simplicity, we chose to keep ρ fixed at 0), and an amplitude ϕ_{ik}. The mean of each Gaussian represents the central tendency of the category (similar to prototypes in some theories of categorization), while the standard deviation represents the variability of the category, with smaller standard deviations equating to more specific categories. The amplitude ϕ_{ik} represents the prior probability that a random stimuli is a member of that category. At initialization, the mean of each category for each agent is randomly drawn from a uniform distribution in [[0, 100][0,100]], with a fixed standard deviation of $\sigma_{initial} = 5$. The amplitude of each category is initialized at $1/K$, so that all categories are initially equally probable.

When acting as communicators, agents generate a signal by sampling a category from their MOG, with the probability of selecting each category proportional to the estimated prior probability of observing that category (the amplitude ϕ_{ik}; we describe below how this is estimated through observation). This assumes that agents simply attempt to reproduce the same frequency distribution that they have learned. This is a reasonable starting assumption for the many cultural domains in which imitation and conformity are useful, such as language, but other ways of mapping from memory to production should be explored in future work. We assume that communicators attempt to signal the mean of their selected category, but that noise may distort the

signal that gets received. We used simple Gaussian noise, added independently to each signal dimension, with mean of 0 and standard deviation W.

Upon receiving a signal from a communicator, the receiver agent uses Bayesian inference to categorize the signal and adjust the parameters of their MOG representation in memory. This process is somewhat complicated, and is described in greater detail below (see "Learning"). Essentially, the receiver first maps the signal as a member of the most likely of its own stored categories. It then updates the properties of its categories to reflect this new information.

After each agent has had the opportunity to communicate (not all agents will receive a signal, and some will receive multiple signals, on a given time step), the reproduction stage occurs. Each agent has a probability $1/L$ of dying at each time step, implying an expected lifespan of L time steps. Any agent who dies is removed from the simulation and replaced by a new agent. Newly born agents are initialized in the same way as agents at the beginning of each simulation.

Each simulation was run for 40,000 time steps. Based on piloting, this length appeared sufficient for most of our outcome measures to stabilize. The procedures used to analyze the model are described in detail in the Outcome Measures section below. The code to run this model is available on OSF[1].

5.2.1 Learning

Upon receiving a signal from a neighbor, receiver agents categorize the signal and adjust their category representations using Bayesian inference. Agents first compute the likelihood of the signal belonging to each category j in their MOG, according to a Gaussian likelihood function G:

$$G_{ij}\left(x,y\right)=\phi_{ij}\left(\frac{1}{2\pi\sigma_{ijx}\sigma_{ijy}\sqrt{1-{\rho_{ij}}^{2}}}exp\left(-\frac{1}{2\left(1-{\rho_{ij}}^{2}\right)}\left(\frac{\left(x-\mu_{ijx}\right)^{2}}{{\sigma_{ijx}}^{2}}-\frac{2\rho_{ij}xy}{\sigma_{ijx}\sigma_{ijy}}+\frac{\left(y-\mu_{ij}\right.}{{\sigma_{ijy}}^{2}}\right.\right.\right. \tag{5.1}$$

The likelihood of each category can be thought of as the goodness-of-fit of the signal to each category in the agent's repertoire. In neural network terminology, we can think of the likelihoods as the activation levels of each output node (each category) in response to the input signal. The marginal likelihood M of the signal is the sum

[1] https://osf.io/6bsyx/?view$_o nly = e91d9839ebe441a4841e3d312204e655$

of the likelihoods over all categories in an agent's MOG (or we can think of it as the sum activation at the output layer of a neural network):

$$M_i(x, y) \;=\; \sum_{j=1}^{K} G_{ij}(x, y) \tag{5.2}$$

And the posterior probability P of each category is then calculated as the ratio of the likelihood to the marginal likelihood:

$$P_{ij}(x, y) \;=\; \frac{G_{ij}(x,\ y)}{\sum_{j=1}^{K} G_{ij}(x, y)} \tag{5.3}$$

The posterior probability is the proportional goodness-of-fit of the signal to each category, or the activation of each category scaled by the total activation across all categories. The category with the highest posterior probability (the "argmax") can be thought of as the label an agent applies to a signal, or their "interpretation" of a signal.

The parameters of all categories are then updated using a gradient descent algorithm. This algorithm acts to maximize the marginal likelihood function M by adjusting parameters along the derivative of M with respect to each parameter. More simply stated, agents move their categories around in the 2-D signal space and adjust their shapes such that the signal would be better fit by their MOG, if the agent received the same signal again. Importantly, the magnitude of the adjustment on each category is scaled by its posterior probability. This means only categories that are probable given a signal are moved, while others change little, which prevents all categories from converging to a single point. The learning rules for each parameter are as follows:

$$\triangle\mu_{ijx} \;=\; \eta_{\mu} P_{ij} \frac{1}{(1 - {\rho_{ij}}^2)} \left(\frac{x_{ij} \;-\; \mu_{ijx}}{\sigma_{ijx}^2} - \frac{\rho_{ij} y_{ij}}{\sigma_{ijx}\sigma_{ijy}} \right) \tag{5.4}$$

$$\triangle\sigma_{ijx} \;=\; \eta_{\sigma} P_{ij} \left(\frac{(x_{ij} \;-\; \mu_{ijx})^2}{\sigma_{ijx}^3 \,(1 - {\rho_{ij}}^2)} - \frac{\rho_{ij}\,(x_{ij} \;-\; \mu_{ijx})\,(y_{ij} \;-\; \mu_{ijy})}{\sigma_{ijx}^2 \sigma_{ijy}\,(1 - {\rho_{ij}}^2)} - \frac{1}{\sigma_{ijx}} \right) \tag{5.5}$$

where η represents the learning rate for each parameter. For added simplicity in visualization and the signal production process, correlations between the two dimensions of each category were fixed at 0 and did not update.

Unlike the means and standard deviations, the amplitude (also equivalent to a Bayesian prior) parameter ϕ was updated based on winner-takes-all competition, such that only the category with the highest posterior probability increased in amplitude. Intuitively, this means that agents treat each signal as having actually come from only one category, such that each observation should only increase the estimated base rate of one category. The amplitude of the winning category is updated according to the following learning rule:

$$\triangle\phi_{ij} = \eta_{\phi} P_{ij}(x, y) \tag{5.6}$$

After updating the amplitude of the winning category, the amplitudes across all categories were normalized. This winner-takes all competition increases the amplitude of frequently-heard categories while suppressing unused categories. McMurray, Aslin, and Toscano (2009) showed that this type of competition is crucial for unsupervised learning when the number of categories is unknown; in the absence of winner-takes-all competition, individual learners were unable to detect the correct number of phonetic categories within their training data. As a whole, these learning rules allow agents to begin with a relatively large number of equally-probable categories (e.g. 20), and over time to pare their category representation down into the simplest structure that effectively captures the distribution of signals they observe.

We also explored the effects of a "critical period" in learning. The critical period refers to a period early in life during which the brain is highly plastic and learning is facilitated. The existence of such a period is well established in the literature on language development, and may be an important factor in cumulative culture. This was implemented by turning off learning for an agent after they reached an age C in time steps.

It should be noted that, while our agents use Bayesian inference to categorize signals, we take this to be an algorithmic-level description of cognitive operations, in line with arguments presented by McClelland et al. (2010). The mechanism(s) underlying these inferences could be implemented by a distributed neural network or other system, and hence we need not take a stance with respect to the cognitive reality of Bayesian inference here.

5.2.2 Network Structure

We explored four different network structures, illustrated in Figure 5.3, to examine the ways that connectivity patterns can influence cultural attractor dynamics. All networks were undirected, meaning that links were bidirectional, and network structure was held constant throughout each run. In our baseline model, agents were arranged

in a fully-connected network. This results in the largest possible mean degree of N-1 (here, 49), and the largest possible clustering coefficient—the average proportion of agent i's neighbors who are also connected to each other—of 1. This fully connected network also has the smallest possible average shortest path-length—the average of the minimum edges traversed to connect any two nodes—of 1.

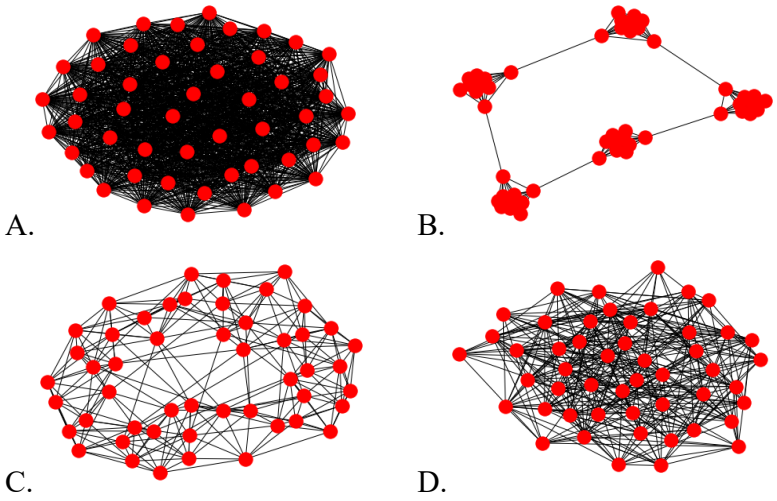


Figure 5.3: Four network structures explored in our model: A. a fully-connected network; B. a connected caveman network; C. a small-world network; D. a realistic social network.

We next considered a connected caveman graph (Watts, 1999), in which agents were first arranged into 5 fully-connected "cliques" of 10 agents each (meaning each agent has 9 neighbors). In each clique, one edge is randomly rewired to a neighboring clique, such that the cliques are ultimately connected in a loop. This network has a near-maximal clustering coefficient of .936. The average shortest path length, however, becomes much longer than the fully-connected network, reaching 3.37.

The third network explored was a small-world network (Watts, 1999), which is formed by connecting each agent to their nearest N neighbors, then randomly rewiring connections with probability P. We used a network where each agent had 10 neighbors and the rewiring probability was set to .1. This resulted in networks with, on the average of 1000 samples, a clustering coefficient of .51 and average shortest-path length of 2.18. All agents had 10 neighbors.

Finally, following methods used by Reali et al. (2018), we explored a "realistic" social network. These networks had a connectivity pattern inspired by empirical patterns seen in modern populations, which have indicated that average nodal degree (i.e. average number of neighbors individuals have) scales with population size such that the clustering coefficient is invariant at a value of $\sim .25$ (Schläpfer et al., 2014). We constructed 20 such networks, which were sampled from randomly across the 100 runs

of the model. These networks had a mean clustering coefficient of .261 and a mean average shortest path-length of 1.7. Agents had an average of 15 neighbors.

5.2.3 Outcome Measures

We analyzed the emergent cultural attractor landscapes along three dimensions: (1) cultural complexity, which we operationalize simply as the number of categories detected at the population level; (2) cultural stability, or the rate of change of the category distribution in signal space; and (3) cognitive alignment, meaning the similarity of cognitive landscapes across individuals. These measures were chosen because of their applicability to a wide array of phenomena in cultural evolution research. First, cultural complexity may relate to the combinatorial possibilities of a cultural repertoire, and measures of complexity are often appealed to in discussions of cumulative cultural evolution. Second, stability may be important for the accumulation of new cultural variants that depend upon existing ones, or for the possibility of inter-generational transfer of information (e.g. if a language changes drastically every generation, communication between individuals of different generations may be disrupted). Third, cognitive alignment may be related to the degree of specialization versus generalization of knowledge in a community, and different domains may benefit from different degrees of alignment (e.g. language is most useful when it is widely shared, while engineering feats may benefit from the joint efforts of individuals with different knowledge).

To obtain our measures, the model was observed every 1000 time steps by generating 500 signals from each agent (using the same method as for communication). Additionally, the state of all agents' MOGs was recorded at the end of each run, in order to characterize cognitive patterns at the agent level. 100 runs were conducted for each parameter setting. To characterize the emergent cultural attractor landscape at the population level, at each time slice of the data we applied the k-means algorithm. To determine the optimal value for k, the partition was calculated at each evaluated time point using values of k ranging from 1-50. We then used the gap statistic (Tibshirani, Walther, & Hastie, 2001) to select the optimal value of k at each timepoint. The optimal value of k was used as an estimate of the complexity of the attractor landscapes.

Next, to examine the stability of the attractor landscape, we adopted a dissimilarity metric for probability distributions known as the earth mover's distance (EMD). The EMD can be understood by imagining different probability distributions as different ways of piling up an amount of dirt (or "earth"). The dissimilarity between two distributions can be thought of as the minimal cost of moving one pile of dirt—a reference distribution—such that it is transformed into a differently-shaped pile of dirt—a target distribution. In this way, the EMD is a type of optimal transport algorithm. While there are many popular similarity metrics to choose from, such

as the Kullback-Leibler divergence or Jensen-Shannon divergence, we selected the EMD because it is symmetrical (unlike KL) and can handle events with probability of 0 (unlike JS). Furthermore, the EMD accounts for the metric space in computing distances. For example, two distributions of the same shape but located in different regions of the signal space will be treated as different under the EMD, but would have a distance of 0 under KL divergence, because the latter does not account for the location of the observations.

Because our signal space is continuous, to compute the EMD we first constructed a discrete probability distribution based on the full set of signal samples at each time point. The signal space was divided into a grid of 20×20 evenly-spaced points (each square being 5×5) and the number of observations in each square was counted, creating a 2-D histogram which was then normalized to sum to 1. We then computed the EMD between the population distribution at each timepoint t to the same population at time t - 1 (therefore there is no measure taken at time 0). This provides a measure of the change in the population distribution over the time between each evaluated timepoint (the model was evaluated every 1000 timesteps).

Finally, to examine the cognitive alignment across agents, we computed the average EMD of the distribution of signals generated by an individual agent to the distribution generated from the rest of the population. Since this is a dissimilarity metric, we will henceforth refer to this measure as cognitive *dis*alignment. At each evaluated timepoint, a 2-D histogram was constructed from the signal samples from each individual agent i in a population of size N, and was compared to another histogram was constructed from the signal samples corresponding to every agent *besides* the focal agent (similar to the "jackknife" resampling technique). Finally, we took the average of these values across agents, which provides a measure of the relative cognitive alignment vs. idiosyncracy, or generalization vs. specialization, in a population.

Table 5.3: Variable model parameters. The values used in the baseline model are presented in bold font.

Parameter	Values Explored	Description
W	**0**, 3, 10	*S.D. of Gaussian noise.* In transmission of a signal, Gaussian noise is added with mean 0 and S.D. = W.
L	5000, **10000**, 15000	*Expected Lifespan.* On each iteration, each agent has probability 1/L of "dying" and being replaced by a new agent.

C	2500, 5000, 10000, 20000, **40000 (length of simulation)**	*Length of the "critical period."* After reaching age C (in time steps), learning is turned off for an agent.
N	10, 25, **50**, 100, 200	*Population size.*
Network type	**Fully-connected**, connected caveman, small world, realistic social network	Four different network structures were explored for connecting agents to neighbors in communication.

Table 5.5: Fixed model parameters.

Parameter	Value	Description
K	20	*Number of Categories in each agent's MOG.*
$\sigma_{initial}$	5	*The S.D. of each category in an agent's MOG upon initialization.*
η_μ	1	*Learning rate for Category means.*
η_σ	1	*Learning rate for Category standard deviations.*
η_ϕ	.001	*Learning rate for Category amplitudes.*
η_ρ	0	*Learning rate for correlation between dimensions.* For simplicity, this value was held constant at 0 such that the two dimensions were uncorrelated.

5.3 Simulation Experiments

In this section, we first present a qualitative analysis of the model dynamics. We then consider three case studies illustrating applications of the model to several areas of inquiry within cultural evolution. First, we consider the effect of transmission noise, which we find has the effect of stabilizing cultural attractor landscapes. Then, we consider the effect of longer lifespans and critical periods in learning, and find that shorter learning times may generate more complex and more stable attractor landscapes. Finally, we consider the effect of population size and network structure. We

find that large populations stabilize and simplify attractor landscapes, while highly cliquish network structures can allow the maintenance of many distinct cultural categories.

5.3.1 Baseline Model: Qualitative Analysis and Visualization

In order to get an intuitive sense of the dynamics of our model, and how the emergent patterns act as cultural attractors, we will first visually analyze the behavior of the model over time on a single representative run (see Table 2 for parameters). Figure 5.4 shows the state of all categories across all agents at nine different time points during a single representative run.

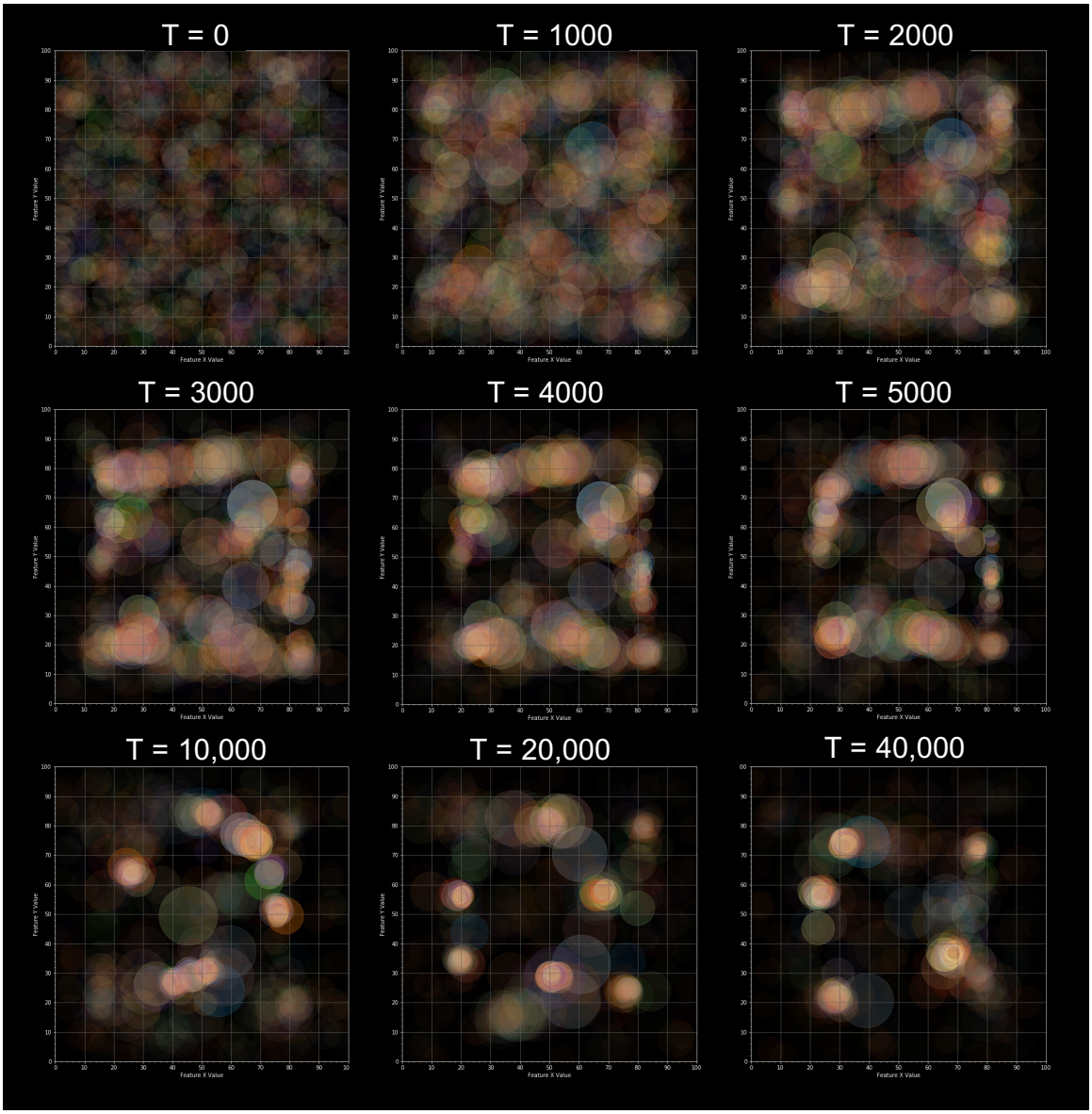


Figure 5.4: The states of all categories across all agents in a population at nine time points of one run. Different colors correspond different agents (here $N = 50$ agents). Each agent has multiple categories (here $K = 20$ categories) in their MOG, which are individual points (20 x 50 = 1000 points total). The size of points is proportional to the SD of the category, and the transparency (alpha value) of the points is proportional to the amplitude of the category, such that low-frequency categories become more transparent. The appearance of fewer points in later time steps is the result both of the alignment of categories across agents, such that points overlap, as well as the fact that most categories in each agent's MOG become suppressed, rendering them transparent in the plot. It should be noted that, because both overlap and amplitude impact the transparency of points, their respective contributes cannot be visually distinguished (i.e. the same visual result can be achieved by fewer overlapping points of greater amplitude, or more overlapping points of lesser amplitude).

The model begins with all agents possessing a set of equal-amplitude categories, uniformly distributed throughout the signal space. Over the first 5,000 time steps, we can see cultural attractors beginning to emerge, as nearby categories are pulled closer and competition at the cognitive level results in some categories getting suppressed, while others increase in amplitude (and therefore, the probability that they will be produced in the future). By 10,000 generations, a clearly distinguishable set of tight clusters have emerged, though there remain some looser clouds of low-amplitude categories, likely driven by new learners entering the population (see Fig. 5.8). At this point the model appears to have reached a dynamical equilibrium, where the qualitative pattern remains the same, but clusters continue to drift around stochastically. Some categories move too near to each other and "merge," while new clusters may occasionally arise in empty regions and others occasionally fade away. Note that categories at the level of agents do not merge. Instead, if two categories become too close to each other, they will compete within an agent's MOG, which can result in one category increasing in amplitude while the other diminishes. On the other hand, the categories detected at the population scale, using the k-means algorithm, do not directly compete, and thus may be described as merging when the algorithm detects two nearby clusters at one time point, but detects only a single cluster at a subsequent time point that encompasses the former two. See the Supplementary Material for a video version of Figure 4.

This dynamical equilibrium is made clear when visualizing the number of clusters that are detected at the population level over time. Figure 5.5 A. shows a time series of the raw number of clusters detected using the k-means algorithm and the gap statistic (with a max k of 50), averaged over 100 runs with the baseline parameter settings. This plots show that our cluster detection algorithm settles at ~ 15 clusters by 20,000 time steps. Figure 5.5 C. reveals that cognitive disalignment also stabilizes within approximately the same time frame. However, Figure 5.5 B. shows that the distribution of categories throughout the signal space continues to change at roughly the same rate over the entirety of each run. Given that the number of categories detected and the average disalignment of agents appear to reach equilibrium by 20,000 timesteps, all subsequent analyses used average values over the final 20,000 timesteps (the second half) of any given run.

We can think of the clusters that form in our model as cultural attractors because these global patterns are precisely what individuals learn to approximate, and thus the clusters are attractor points in cognitive development. Of course, as others have already stated, these cultural attractors are simply statistical aggregates; individual agents do not have direct access to the population-level attractors, but only to unique signals. However, because these clusters correspond to the expected distribution of observations for a random agent (in a fully connected network), these statistical abstractions constitute a real force shaping cognition.

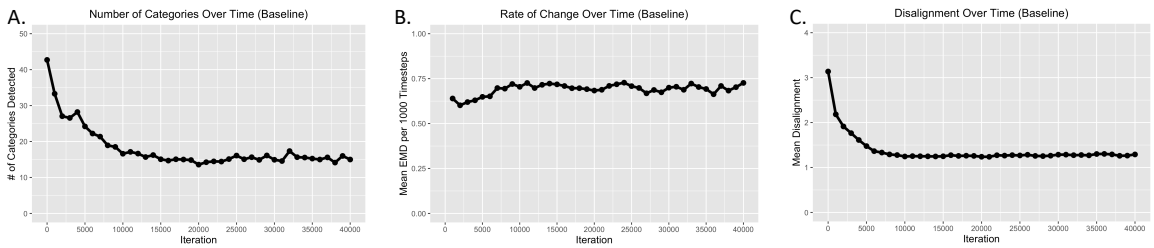


Figure 5.5: Time series, with each point representing the average over 100 runs of the baseline model, of (A.) The complexity of the cultural attractor landscape. (B.) The rate of change of the attractor landscape over time (C.) The cognitive disalignment of agents to the population distribution.

5.3.2 Some Noise is Beneficial for Stabilizing Cultural Attractor Landscapes

We find that as transmission noise is increased, the attractor landscape becomes increasingly stable (Fig. 5.6 B.) This effect is due to the fact that, as noise increases, agents less reliably signal the true mean of their categories, which slows the rate of learning, and therefore the rate of change at the global level. Understandably, increasing noise is also associated with a decrease in the complexity of the attractor landscapes, because when categories become more diffuse, fewer of them can be maintained in the same space (5.6 A.). Some noise (e.g. $W = 5$) also helps to facilitate cognitive alignment in the population (5.6 C.), because the slower-moving targets for learning make it easier for agents to acquire all of the categories in their population. However, the effect of noise on alignment is non-linear: we observe a slight *increase* in disalignment when noise is increased from $W = 5$ to $W = 10$. This suggests that, when $W = 5$, the complexity of the attractor landscape is sufficiently low, and the rate of change sufficiently slow, that agents can effectively align to the population pattern, and therefore further increases in noise will merely reduce the complexity of attractor landscapes at no additional benefit.

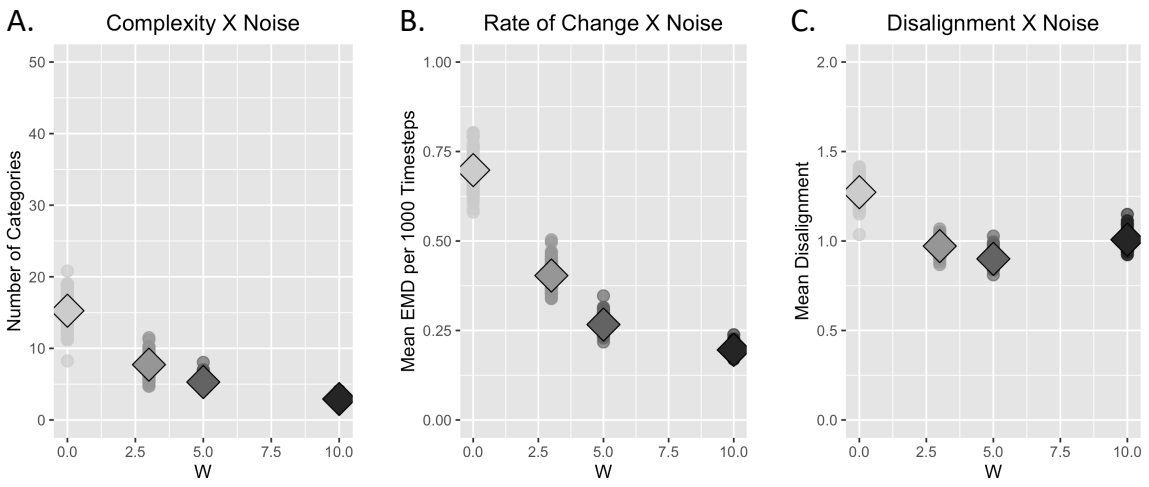


Figure 5.6: The effect of variable Gaussian transmission noise with mean = 0 and SD = W on: (A.) The complexity of the cultural attractor landscape. (B.) The rate of change of the attractor landscape over time (C.) The cognitive disalignment of agents to the population distribution.

5.3.2.1 Discussion

Research on cultural evolution often focuses on the issue of transmission fidelity: transmission noise is generally considered to be a limiting factor for the purposes of cumulative cultural evolution (Nowak et al., 1999), and the success of human populations in developing complex cultural repertoires is often attributed to the high fidelity with which we can transmit information, relative to other species (H. M. Lewis & Laland, 2012). At the same time, many fields outside of cultural evolution have seen a growing recognition of the crucial role that noise can play in complex dynamical systems. This point is exemplified in the literature on "stochastic resonance," which emphasizes that some amount of noise is beneficial for the detection of weak signals in non-linear systems (Gammaitoni, Neri, & Vocca, 2010; McDonnell & Ward, 2011; Wiesenfeld & Moss, 1995). For example, work by Goldman (2004) has shown that the possibility of synaptic transmission failures in the brain can actually enhance the informational efficiency of a synapse.

The behavior of our model with respect to noise suggests a bridge between the literature on transmission fidelity and the work on stochastic resonance. We find that as transmission noise is increased, the attractor landscape becomes increasingly stable over time. These effects are due to the fact that, as noise increases, agents signal the true mean of their categories less reliably, which slows the rate of learning, and therefore the rate of change at the global level. In turn, this helps to promote cognitive alignment across individuals, because the global pattern becomes a slower-moving target for learning. In a domain such as language, cognitive alignment is of crucial

importance, and thus it appears that transmission noise may play a role in the self-organization of linguistic conventions (at least at the level of speech sounds). If there is too *little* noise in transmission, categories may change so rapidly as to create problems using these categories in higher-order systems. For example, lexical categories, which are signalled by combinations of phonemes, may not be possible if phoneme representations are highly unstable in a population.

However, our results should not be taken as contradictory to research suggesting that transmission fidelity is the "key to the build-up of cumulative culture" (H. M. Lewis & Laland, 2012). Rather, we suggest that moderate amounts of noise at the level of behavior/perception promote stable categories that are broadly shared, which counter-intuitively makes these categories able to be signalled with enhanced fidelity. In other words, some within-category noise allows categories to become more distinguishable overall. It is important to reiterate that the attractor landscapes in our model are akin to perceptual distinctions, and should not be confused with higher-order cultural variants that are transmitted *by virtue of* shared perceptual categories. As such, our results suggest that some noise at the level of perception/production may be important for ensuring transmission fidelity at higher levels of abstraction.

5.3.3 Longer Learning Times Can Result in Decreased Complexity of Attractor Landscapes, and Critical Periods Can Enhance Their Stability

While longer learning times intuitively seem necessary in order to acquire more complex knowledge structures, somewhat surprisingly, we find that the complexity of attractor landscapes *decreases* as learning times grow longer. We can see this effect in Fig. 5.7 A., where expected lifespans varied and learning proceeded over the full lifespan. This effect is due to the fact that longer learning times also allow more time for cognitive competition between categories to proceed, which results in more categories becoming suppressed. This suggests that, as agents grow older, they eventually underfit the population distribution, possessing only a subset of the categories that are active at the population level (this is reflected in Fig. 5.8, which shows that older agents conform more poorly to the population distribution than middle-aged agents). Over generations, as new learners are influenced by the behaviors of their older neighbors, this results in a continued decline in the number of categories that are present. However, this comes with the potential benefit of promoting cognitive alignment overall, as agents can more readily fit the global distribution when it is simpler (Fig. 5.7 C.).

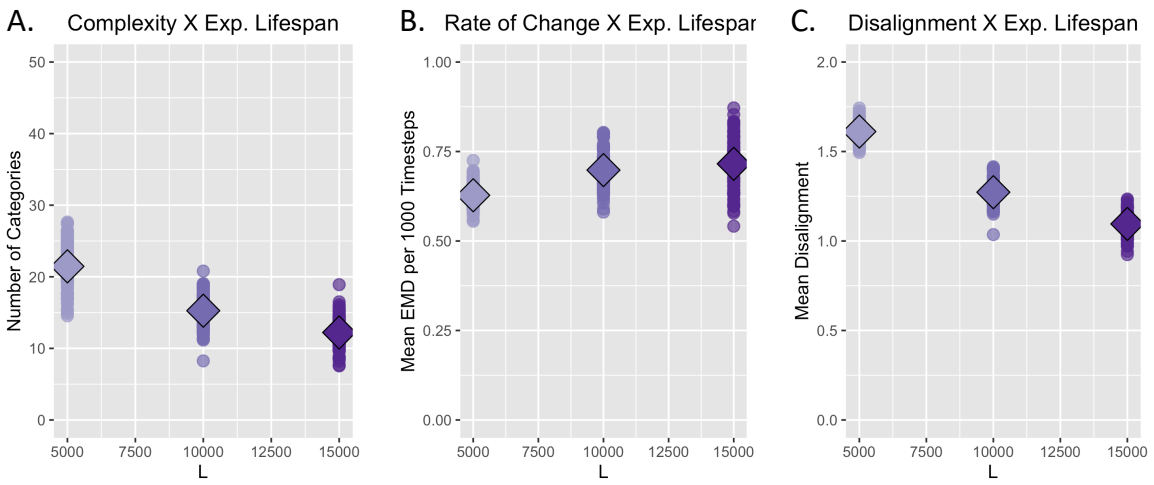


Figure 5.7: The effect of variable lifespans L on: (A.) The complexity of the cultural attractor landscape. (B.) The rate of change of the attractor landscape over time (C.) The cognitive disalignment of agents to the population distribution.

Figure 5.8: Cognitive disalignment with respect to the population clustering pattern by age.

We next considered the effect of adding a critical period of learning, which was implemented by turning off learning after an agent passed C iterations in age. We find that critical periods moderate a trade-off between complexity (5.9 A.) and stability (5.9

B.) of the attractor landscapes. This occurs because, when learning is restricted to a subset of the lifespan, agents who have stopped learning can remain in the population to act as stable models for more recently introduced learners. Figure 5.9 B. shows that the equilibrium value of the rate of change increases as critical periods lengthen. However, if learning times are too short (e.g. 2500 time steps in our model), we observe that agents do not have sufficient time to fit the population distribution, resulting in an increase in cognitive disalignment relative to moderate lengths of critical periods (Fig. 5.9 C.).

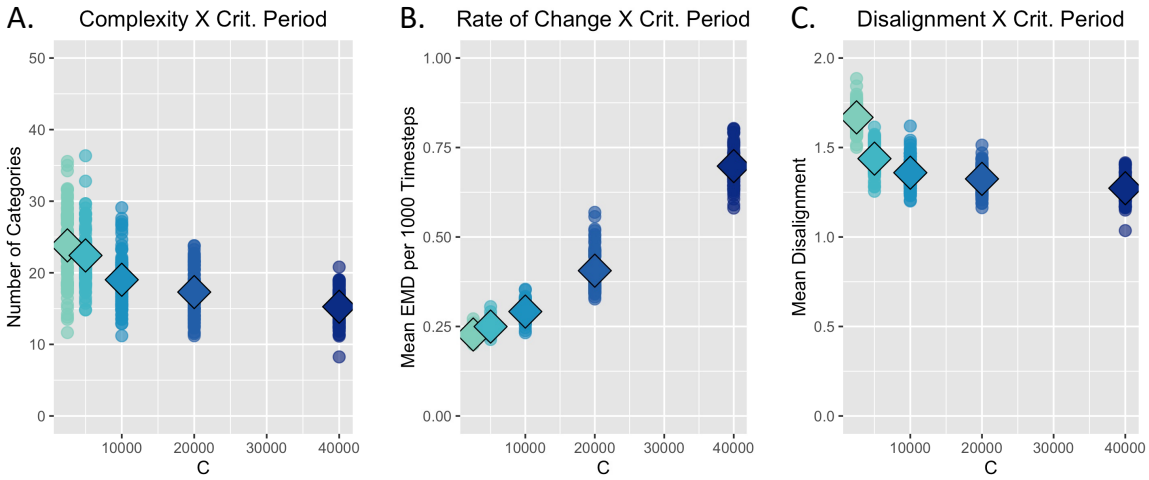


Figure 5.9: As a function of the length of the critical period, C: (A.) The complexity of the cultural attractor landscape. (B.) The rate of change of the attractor landscape over time (C.) The disalignment of agents to the population distribution.

5.3.3.1 Discussion

The typical story told about learning in the research on biological evolution goes something like this: Investment in learning is helpful for adaptation to harsh and/or variable environmental conditions, but time spent learning is costly and detracts from reproductive opportunities. Thus, many organisms exhibit a "sensitive" or "critical" period early in life in which to assess environmental conditions, before committing to an adult phenotype (Frankenhuis & Panchanathan, 2011; Frankenhuis & Walasek, 2020; Panchanathan & Frankenhuis, 2016). In the domain of language, the existence of a critical period is one reason that language acquisition is facilitated in children and harder for adults (Birdsong, 1999; Hakuta, Bialystok, & Wiley, 2003). Such critical periods are generally thought of as a constraint, rather than an adaptation (Hurford, 1991; Komarova & Nowak, 2001). Our findings add complexity to this story, by revealing that as learning times grow longer (e.g. as an adaptation to a complex cultural repertoire), this may cause cultural attractor landscapes to simplify over time. If such a mechanism exists in real groups of cognitive agents, this could help to prevent runaway complexity: if cultural repertoires become complex, this

may select for greater investment in learning, which may in turn result in the cultural repertoire simplifying. Thus, our findings suggest that shorter learning times may not only be selected for due to the cost of learning, but also (likely at the group level) due to a possible role in stabilizing the cultural attractor landscape.

While it is possible that the effect of longer learning times on reducing the number of categories is merely an artifact of our learning algorithm, and may not generalize to real human cognition, there is some reason to think that this may be a real effect. First, we can note that if learning in the brain is Hebbian, neuronal responses that have occurred in the past will increase the tendency for the same response to occur in the future, even if that response is inappropriate with respect to the input (i.e. a categorization error). For example, Japanese speakers may have trouble learning the contrast between /r/ and /l/ phonemes that are present in English, but absent in Japanese, because presentation of either phoneme may simply reinforce the Japanese category that falls somewhere between the English /r/ and /l/ (McClelland, Thomas, McCandliss, & Fiez, 1999). In our model, when a stimulus is categorized as belonging to a high-prior-probability category that is slightly further away from the input value than a lower-prior-probability category, we may consider this a categorization "error," but the winning category will be reinforced nonetheless. Similar effects have been observed in humans, whereby making repeated responses that are in error result in a decrease in participants' abilities to discriminate between perceptual categories (McClelland et al., 1999). This point is further supported by evidence that older adults place a greater weight on lexical frequency when identifying a spoken word among a set of candidate words (i.e. older adults are more likely to identify the spoken word as corresponding to the more-frequent candidate; Revill & Spieler, 2012). Finally, we can note that aging is associated with a decrease in neural resources, which could further limit the number of perceptual distinctions available to an individual (Fjell & Walhovd, 2010). As such, the empirical research on aging and cognitive function suggests that the behavior of our model—a decrease in the complexity of the attractor landscape as learning times increase—is plausible.

Taken together, our results point to interesting trade-offs among stability, complexity, and cognitive alignment with respect to learning times and lifespans. When the problem space is continually changing, longer learning times may be beneficial. At the same time, longer learning times decrease the stability of the learning space, but also may decrease the number of categories to be learned. Critical periods, on the other hand, may not only provide a fitness benefit by minimizing the energy invested in learning, but may also play an important role in the stability and coherence of the cultural variants found in a population. It remains unclear how human developmental trajectories have evolved to balance these complex interactions in a way that allows for cumulative culture, but future explorations with our model, with the addition of representations of fitness and reproduction (allowing for heredity in cognitive capacities), may be able to shed some light on this issue.

5.3.4 Larger Populations Have Simpler, More Stable Cultural Attractor Landscapes, and Network Structure Can Moderate These Effects

Population size and/or density are commonly implicated as important factors in the potential for cumulative cultural evolution, with larger/denser populations being thought to sustain more complex cultural repertoires (Henrich, 2004; Reali et al., 2018). However, in our model we find that larger populations do not tend towards more complex clustering schemes (Fig. 5.10 A.)—in fact, the pattern is quite the reverse, though the number of categories appears to approach a lower asymptote of ~ 10 categories as populations become large. Interestingly, the decrease in complexity that is associated with larger population sizes is not paired with a corresponding decrease in cognitive disalignment (Fig. 5.10 C.). Instead, disalignment shows a slight *positive* relationship with population size. This can be explained by the fact that, although larger populations appear to have simpler emergent category structures, agents in larger populations also have fewer repeat interactions, which is a detriment to cognitive alignment. We also observe that larger populations have slower-changing attractor landscapes (Fig. 5.10 B.). This is because, in smaller populations, individual agents contribute more substantially to the global average. As such, deaths and births of new agents constitute a more significant perturbation in smaller populations, leading to sudden spikes in the rate of change.

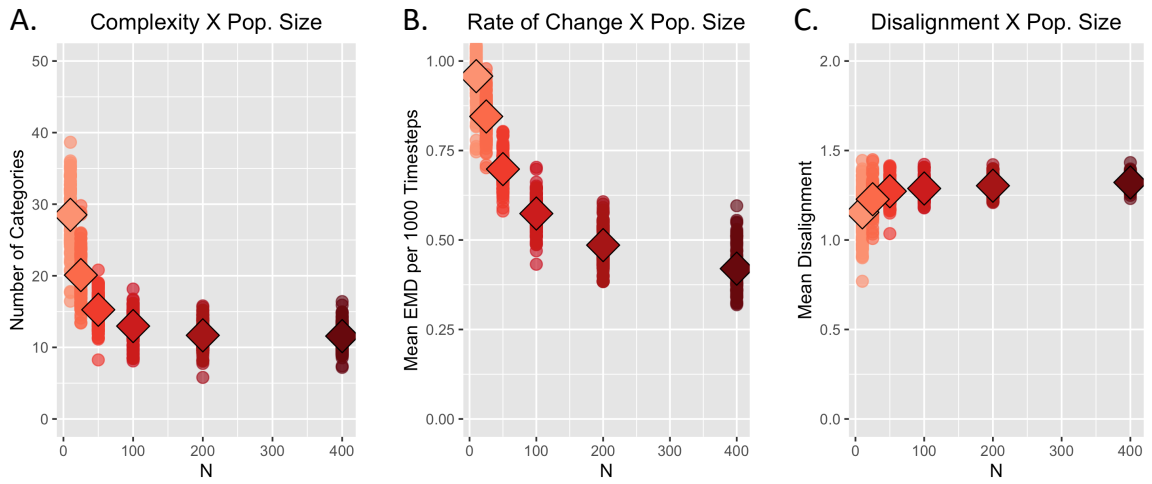


Figure 5.10: The effect of variable population size, N on: (A.) The complexity of the cultural attractor landscape. (B.) The rate of change of the attractor landscape over time (C.) The cognitive disalignment of agents to the population distribution.

Considering network structure, our results show no difference between fully-connected, small-world, or realistic social networks in terms of the complexity of the cultural attractor landscape that forms (Fig. 5.11 A.). However, the connected caveman network differs dramatically from the others, showing far greater complexity. This effect is due

to the fact that the limited connectivity between cliques in the connected caveman network limits the diffusion of conventions, such that each clique tends to converge upon a distinct set of categories, resulting in a much larger number of categories being maintained in the population overall. However, individuals *within* a clique do not actually have more complex cognitive landscapes, relative to individuals embedded in other networks. Thus, this effect is actually due to a decrease in the cognitive alignment of individuals with respect to the global pattern (Fig. 5.10 C.): individuals within a clique conform to each other, but not to others outside of their clique.

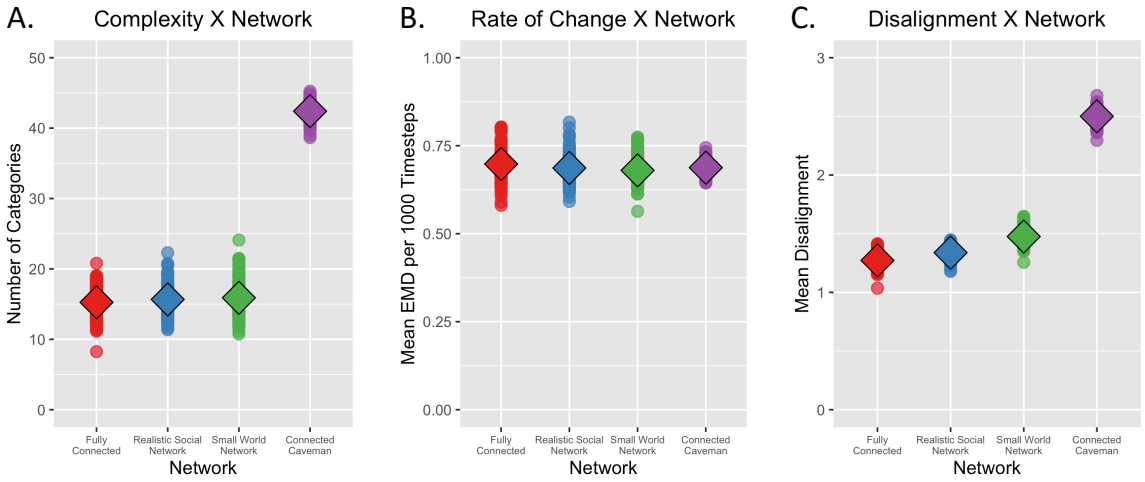


Figure 5.11: As a function of the network type: (A.) The complexity of the cultural attractor landscape. (B.) The rate of change of the attractor landscape over time (C.) The cognitive disalignment of agents to the population distribution.

5.3.4.1 Discussion

Population size and demographic structure are some of the most commonly implicated factors in theories of cultural and linguistic evolution. For example, relating to population size/density, it has been proposed that larger and/or denser populations may be able to sustain more complex skills and technologies (Henrich, 2004), and that larger populations tend to develop larger vocabularies, but simpler grammars (Lupyan & Dale, 2010; Reali et al., 2018). Relating to network structure, the literature on group problem solving suggests that different patterns of connectivity and/or network change are optimal for different types of problems (Lazer & Friedman, 2007; Rulke & Galaskiewicz, 2000; Smart, Huynh, Braines, & Shadbolt, 2010). For example, Lazer and Friedman (2007) showed that, in complex problem spaces where individuals can either independently explore the solution space or copy the solutions of successful neighbors, moderate amounts of network connectivity prove most efficient, because they balance breadth of exploration with the rapid diffusion of "good enough" solutions. Complementing this work, Smolla and Akçay (2019) recently showed that networks and culture may co-evolve, with environments that

select for specialist knowledge resulting in sparse connectivity patterns, such that individuals are repeatedly exposed to the same information and increase their depth of expertise, while environments that select for generalist knowledge result in dense connectivity patterns, for complementary reasons. Other recent results from Cantor et al. (2021) suggest that the relationship between network structure, population size, and diffusion mechanisms are highly complex: networks that perform best in terms of cumulative cultural evolution in one context may perform worst in another.

Some of the results of our case studies are consistent with existing research. For example, as in the work on group problem solving and the role of network structure on cumulative cultural evolution, we find that cliqueish networks (like the connected caveman network) limit the diffusion of conventions, resulting in greater diversity of cognitive landscapes in the population (Derex & Boyd, 2016; Lazer & Friedman, 2007). Other results are complementary to existing research. For example, to the best of our knowledge, there is no model that accounts for the fact that larger populations do not tend to have larger repertoires of perceptual categories (e.g. phoneme inventories, Creanza et al., 2015; Moran, McCloy, & Wright, 2012; though there is some debate here, e.g. Fenk-Oczlon & Pilz, 2021), while they clearly do differ in the complexity of higher-order cultural repertoires that depend on combinations of these categories, such as tools and grammar. While we have not currently explored the possibility of agents constructing artifacts that consist of combinations of elements, our model can be extended to allow for this possibility, as we will discuss in the next section. As such, our model can be integrated with existing models of cultural innovation, and therefore can allow for exploration of the interactions between these two levels of analysis. Finally, our model also produces some behaviors that have not been noted at all in the literature. For example, the role of population size on stabilizing change in attractor landscapes, and the relationship between population size and cognitive alignment, are novel effects, to the best of our knowledge. Thus, our model may suggest interesting new paths for future research, in addition to complementing existing work.

Our explorations with network structure reveal how distinct patterns of cognitive alignment can arise from distinct patterns of connectivity. A network of connectivity determines not only how information flows through a population (i.e. the paths it takes through the network), but also how it is distorted as it flows through that network. Our model focuses on the patterns of distortion, but it is important to note that networks themselves may evolve, reaching different distributions of cognitive landscapes depending upon selection pressures in different domains. This could be a result of preferentially forming connections with those who are cognitively similar (e.g. because interactions are more successful on average), or selectively attending to prestigious or knowledgeable others. Furthermore, networks of interaction for real individuals are better described as multiplex networks.

5.4 Conclusion

Despite some significant debates in the history of cultural evolution research, it is now generally agreed upon that both preservative dynamics (i.e. Darwinian selection) and transformative dynamics (i.e. cultural attraction) are crucial aspects of how culture evolves. We agree with previous claims by Henrich et al. (2008) and Claidière et al. (2014), that cultural attraction effects support Darwinian cultural evolution: when cultural variants cluster around points in the space of possible features, cultural information can be transmitted repeatedly without accumulating random error. Thus, while cumulative cultural evolution may depend upon high-fidelity *transmission*, this does not necessarily imply high-fidelity *copying* mechanisms (Saldana, Fagot, Kirby, Smith, & Claidière, 2019). We attribute this effect largely to collective cognitive alignment, meaning that cultural group members tend to perceive, remember, and reproduce information in consistent ways. To the extent that collective cognitive alignment is maintained through enculturation, whereby each individual "acquires" the cognitive biases of their group through interaction, it becomes possible that collective cognitive alignment may *fail* to be achieved either within or across generations. We have advanced cultural attractor theory by providing a socio-cognitive model of how cultural attractors may form, change, and stabilize in the absence of strongly-determining ecological constraints or innately-shared biases.

Our explorations with this model illustrate that factors at the scale of cognition, development, and demographic structure may interact in complex ways to shape patterns of collective cognitive alignment. First, we found that small amounts of noise in transmission may slow the rate of change of cultural attractor landscapes, promoting cognitive alignment within the population. In this way, noise at the level of perception and/or production may be counterintuitively beneficial for *reducing* errors at the level of cultural categories. Next, we found that longer learning times may result in a reduction of the number of categories at the population level over time, due to competition effects at the cognitive level. At the same time, critical periods of learning help to stabilize attractor landscapes, because older agents remain in the population as "frozen" models for developing agents. Finally, we found that the complexity of cultural attractor landscapes decreases as population size grows larger, approaching a non-zero asymptote. This occurs because individuals in larger populations, in our baseline fully-connected network, have fewer repeat interactions, which makes close alignment more difficult, and as a result, more diffuse categories are maintained in the population. This effect can be mitigated, however, through highly cliquish network structures that make repeat interactions very high, but this can come at the cost of global alignment. Our results offer a preview of the insights that may be gained by introducing more detailed representations of the culture-cognition feedback loop into more models of cultural evolution.

A crucial next step will be to include fitness constraints and selectionist transmission in our model. At present, the cultural attractors in our model are arbitrary and

fitness-neutral. This was an important simplification, since, as we have argued, the organic emergence of a cultural attractor landscape may be a critical *precondition* for Darwinian cultural selection, so an explanation of the emergence of cultural attractors must not ultimately fall back to Darwinian selection. Nonetheless, the cognitive capacities and developmental trajectories that facilitate the emergence of cultural attractors *are* themselves biologically evolved, so natural selection will need to be brought back into the picture in future work. Our model can be extended to incorporate biological inheritance of cognitive priors and/or developmental hyper-parameters, as well as to include fitness constraints, by placing our agents into any type of evolutionary or communicative game. For example, the attractors that emerge could be mapped onto behaviors with immediate survival consequences, or onto frequency-dependent consequences such as when establishing shared systems of reference. We can also allow agents to generate *sequences* of signals, which may provide new insights into the entanglement between perceptual and combinatorial cognition in cultural attractor dynamics.

Integrating theories of Darwinian cultural selection with theories of cultural attraction—and theories of cognition more generally—will benefit from more mechanistic models of the feedback loop between cognitive development and population dynamics. Our model contributes to this theoretical bridge by representing cognitive, dyadic, developmental, and demographic dynamics simultaneously, in order to examine the conditions that either promote or inhibit the self-organization and maintenance of a stable cultural attractor landscape. Viewing cultural attractor landscapes as a complex system of interacting constraints at multiple levels allows for straightforward integration of cultural attractor theory with Darwinian selectionist accounts: fitness-based selection effects can be understood as yet another constraint on the formation of statistical attractor points. We hope this model will be useful for researchers interested in the co-evolution of innate cognitive biases, developmental tendencies, and demographic structure with culture.

Chapter 6

Conclusion: Towards a Science of Mind that Doesn't Rule out Minds

In 2008, the artificial intelligence researcher Luc Steels declared that the symbol-grounding problem had been solved. Steels was referring to the fact that research using autonomous artificial agents has successfully established that these agents can spontaneously coordinate on non-intrinsic symbol-meaning mappings. For example, when a population of agents is initialized with no shared signal-meaning mappings, but by repeatedly pairing up agents in a "guessing game" task in which a speaker agent must produce a name to identify one target item among a set of distractors, and a listener agent tries to select the intended target, these populations can eventually develop their own "proto-language" and achieve a high rate of success in the task. Similarly, research using the iCub robot—which was granted a set of "action primitives" such as *push, pull, grasp* and *release*, a repertoire of available "words", and a recurrent neural network connecting the two sets of categories—has shown that the robot is able to acquire higher-order categories such as *give* (a combination of *grasp, push* and *release*) that are apparently "grounded" in combinations of action primitives (Stramandinoli, Marocco, & Cangelosi, 2012). On the basis of work such as this, Steels concluded that the symbol-grounding problem was a thing of the past.

Unfortunately, Steels' claim was misleading, as it represents a very narrow conception of the symbol-grounding problem. The artificial intelligence research to which he referred merely demonstrates that agents can coordinate on arbitrary mappings between inputs and outputs. But inputs and outputs are still all these agents have, and at no point does this work get us to *intrinsic* meaning of symbols for the agents themselves. A network of coordinated Chinese Rooms has no more meaning than just one (Searle, 1980). In order to truly solve the symbol-grounding problem at the level of language, as Steels thinks we have done, we first need to answer the more fundamental question of how an *individual* cognitive agent can have meaning.

These two levels of symbol grounding—the linguistic and fundamental cognitive

level—are related in important ways. Computational or cognitivist theories of the mind, which understand mental representations as a kind of encoding or correspondence to things in the world, would seem to give us no way out of the symbol-grounding problem at the more fundamental level, and therefore to leave us with only a hollow form of grounding in language. Non-representational theories may seem to avoid the problem at the fundamental level, in showing that many tasks can be solved without the need for symbols at all, instead by an organism "resonating" to flows of energy in the environment. But the cost of adopting a non-representational approach is to make language all the more mysterious. How are we to coordinate on shared linguistic symbols if there are no symbols at all?

In this dissertation, I have tried to pursue a middle-path between representational and non-representational theories of mind: we should have an account that includes a notion of mental representations or symbols, much needed for explanations of language, but also be able to explain how these representations are grounded in perception and action. In chapter two, I presented a reservoir computer model of cognition that attempts to lay the groundwork for this approach. Through local homeostatic mechanisms, a collection of nodes analogous to neurons comes to "resonate" to patterns of input. This process leads to the emergence of "transient localist representations": spontaneous, temporary, and context-dependent mental categories that we may think of as a kind of internal symbol. These representations may count as symbols in the sense that highly similar activation patterns can recur and relate to the same meanings, such as "*stimulus approaching from left, turn right*" or "*dog*" in the subject position. These representations can also be used by the network to generate apparent "predictions" of *specific* upcoming symbols. Given these properties, we can see how such representations may lay the groundwork for linguistic capacities. Nonetheless, these representations are *not* stable encodings of the kind posited by computational theories of mind, and have meaning for the agent only in the context of an ongoing flow of interaction.

In chapter three, I considered one prominent proposal for how linguistic meaning may be grounded: in sensorimotor simulation. Upon closer inspection, this proposal would seem to do little to get us out of the symbol-grounding problem at the fundamental level. Just as the iCub robot can learn the meanings of words for higher-order concepts such as "give" by combining action primitives, humans that understand a word by reactivating sensorimotor regions are simply associating one internal representation to another, never reaching an intrinsic meaning. In other words, theories "grounded cognition" in the sense of Barsalou's Perceptual Symbols System (Barsalou, 1999) are still very much "internalist," and therefore actually ungrounded. The results of the two experiments presented here helps to clarify the limitations of this approach. First, we found that abstract or metaphorical language showed no sign of relying on sensorimotor regions for comprehension. Second, we found that the extent to which even literal language depends upon sensorimotor activity is a matter of context, further expanding the scope of linguistic abilities that is unexplained by the sensorimotor simulation account. Third, our evidence was consistent with the

claim that action-sentence compatibility effects occur in a late stage of processing, and therefore that sensorimotor involvement is not necessarily *functionally* related to language comprehension, but instead a result of mental simulation or imagery after a sentence is understood. While our results certainly leave room for sensorimotor activity to be a part of the story of how linguistic meanings are grounded, they indicate that more is needed.

In chapter four, I investigated the extent to which perceptual categories that underlie language comprehension are flexible or context-sensitive. As discussed in chapter two, one of the benefits of "transient localist representations" may be their context-flexibility: *similar* mental representations can be reused across time if they correspond to adaptive trajectories through a cognitive state space (i.e. trajectories that help to preserve homeostasis), but these mental representations can integrate information from many sources and change on a dime in order to fit the current context. Chapter four demonstrates evidence for such context-flexibility in the phonetic categories of Spanish-English bilinguals. We found that a very subtle manipulation of context—changing a small set of experimental instructions from English to Spanish, and using different sets of visual stimuli—can lead Spanish-English bilinguals to carve up an acoustic continuum in different ways.

Chapter five returned us back to the level of the symbol-grounding problem addressed by Steels—the level at which groups of agents attempt to coordinate on shared signal-meaning mappings. The agent based model in that chapter actually focused on just one side of this problem: the coordination of shared signals. One unrealistic assumption of the artificial intelligence research in this domain is that agents come pre-equipped with a shared repertoire of discrete signals and meanings, even if they initially lack shared *mappings* between them. Our model shows that if one takes the fuzziness and context-sensitivity of human categorization into account, that assumption of prior work appears unjustified. We report some rather counterintuitive results, suggesting that the capacity for human-level language or culture is *not* due to high-fidelity copying and extensive learning, but instead facilitated by transmission noise and short learning times. Our results speak to the way that language is a complex system, requiring the coordination of factors at multiple levels including perception, development and the life cycle, and population size and structure.

Taken as a whole, this body of work may begin to paint a picture of how we can build a theory of cognition that takes *meaning* seriously, and preserves an unbroken "grounding wire" of cognitive models from the level of simple organisms, all the way up to the level of human language evolution. Constructing such a theory does not by any means require that we throw out decades of useful work associated with a computational approach to mind. But it will require reconceptualizing some fundamental concepts in cognitive science, such as "representation," in a way that doesn't fall victim to the symbol-grounding problem. In doing so, we may begin to see many computationalist accounts in a new light, more useful as predictive models or for pragmatic purposes such as developing autonomous artificial agents, but not as true

explanations of cognition. In other cases we may find that computationalist theories can be supplemented with additional mechanisms that suffice for grounding. In any case, I hope that this dissertation may begin to demonstrate that taking meaning seriously is not only a challenge for theories of cognition, but also a mindset that may reveal new insights or paths at every level of analysis.

Bibliography

Acerbi, A., Charbonneau, M., Miton, H., & Scott-Phillips, T. (2019). Cultural stability without copying.

Acerbi, A., Charbonneau, M., Miton, H., & Scott-Phillips, T. (2021). Culture without copying or selection. *Evolutionary Human Sciences*, *3*.

Acerbi, A., & Mesoudi, A. (2015). If we are all cultural darwinians what's the fuss about? clarifying recent disagreements in the field of cultural evolution. *Biology & philosophy*, *30*(4), 481–503.

Acerbi, A., Mesoudi, A., & Smolla, M. (2020). Individual-based models of cultural evolution. a step-by-step guide using r.

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological science*, *17*(9), 814–823.

Ahlberg, D., Dudschig, C., & Kaup, B. (2013). Effector specific response activation during word processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(35)).

Anderson, M., & Chemero, A. (2019). The world well gained. *Andy Clark and his critics*, 161–173.

Anderson, S., & Spivey, M. (2009). The enactment of language: Decades of interactions between linguistic and motor processes. *Language and Cognition*, *1*(1), 87–111.

Andres, M., Finocchiaro, C., Buiatti, M., & Piazza, M. (2015). Contribution of motor representations to action verb processing. *Cognition*, *134*, 174–184.

Aravena, P., Courson, M., Frak, V., Cheylus, A., Paulignan, Y., Deprez, V., & Nazir, T. (2014). Action relevance in linguistic context drives word-induced motor activity. *Frontiers in human neuroscience*, *8*, 163.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol.*, *56*, 149–178.

Aziz-Zadeh, L., Wilson, S., Rizzolatti, G., & Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Current biology*, *16*(18), 1818–1823.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). The celex lexical database (cd-rom).

Baggs, E., & Chemero, A. (2018). Radical embodiment in two directions. *Synthese*, 1–16.

Baggs, E., & Chemero, A. (2019). *The third sense of environment.* New York: Routledge.

Balasubramaniam, R., Riley, M. A., & Turvey, M. (2000). Specificity of postural sway to the demands of a precision task. *Gait & posture*, *11*(1), 12–24.

Baronchelli, A., Gong, T., Puglisi, A., & Loreto, V. (2010). Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences*, *107*(6), 2403–2407.

Barr, D. J. (2008). Analyzing 'visual world'eyetracking data using multilevel logistic regression. *Journal of memory and language*, *59*(4), 457–474.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577–660.

Barsalou, L. (2008). Grounded cognition. *Annu. Rev. Psychol*, *59*, 617–645.

Barsalou, L. (2015). *Processes that mediate stimuli and responses make human action possible.* (The pragmatic turn: toward action-oriented views in cognitive science, 18, 81.)

Barsalou, L. (2016). On staying grounded and avoiding quixotic dead ends. *Psychonomic bulletin review*, *23*(4), 1122–1142.

Barsalou, L., et al. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577–660.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1).

Bentz, C., & Winter, B. (2014). Languages with more second language learners tend to lose nominal case. In *Quantifying language dynamics* (pp. 96–124). Brill.

Bergen, B., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive science*, *31*(5), 733–764.

Bergen, B., & Wheeler, K. (2010). Grammatical aspect and mental simulation. *Brain and language*, *112*(3), 150–158.

Bickhard, M. H. (2009a). Interactivism: A manifesto. *New Ideas in Psychology*, *27*(1), 85–95.

Bickhard, M. H. (2009b). The interactivist model. *Synthese*, *166*(3), 547–591.

Bickhard, M. H. (2016a). The anticipatory brain: Two approaches. In *Fundamental issues of artificial intelligence* (pp. 261–283). Springer.

Bickhard, M. H. (2016b). Inter-and en-activism: Some thoughts and comparisons. *New Ideas in Psychology*, *41*, 23–32.

Bickhard, M. H., & Terveen, L. (1996). *Foundational issues in artificial intelligence and cognitive science: Impasse and solution* (Vol. 109). Elsevier.

Birdsong, D. (1999). *Second language acquisition and the critical period hypothesis.* Routledge.

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of life reviews*, *29*, 120–153.

Borreggine, K., & Kaschak, M. (2006). The action–sentence compatibility effect: It's

all in the timing. *Cognitive Science*, *30*(6), 1097–1112.
Boulenger, V., Hauk, O., & Pulvermüller, F. (2009). Grasping ideas with the motor system: semantic somatotopy in idiom comprehension. *Cerebral cortex*, *19*(8), 1905–1914.
Boulenger, V., Shtyrov, Y., & Pulvermüller, F. (2012). When do you grasp the idea? meg evidence for instantaneous idiom understanding. *Neuroimage*, *59*(4), 3502–3513.
Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago press.
Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, *42*.
Brown-Schmidt, S., & Fraundorf, S. H. (2015). Interpretation of informational questions modulated by joint knowledge and intonational contours. *Journal of Memory and Language*, *84*, 49–74.
Brysbaert, M., & New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, *41*, 977–990.
Buskell, A. (2017). What are cultural attractors? *Biology & Philosophy*, *32*(3), 377–394.
Cantor, M., Chimento, M., Smeele, S. Q., He, P., Papageorgiou, D., Aplin, L. M., & Farine, D. R. (2021). Social network architecture and the tempo of cumulative cultural evolution. *Proceedings of the Royal Society B*, *288*(1946), 20203107.
Casasanto, D., & Gijssels, T. (2015). What makes a metaphor an embodied metaphor? *Linguistics Vanguard*, *1*(1), 327–337.
Cavalli-Sforza, L. L., & Feldman, M. W. (1973). Cultural versus biological inheritance: phenotypic transmission from parents to children.(a theory of the effect of parental phenotypes on children's phenotypes). *American journal of human genetics*, *25*(6), 618.
Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton University Press.
Chatterjee, A. (2010). Disembodying cognition. *Language and cognition*, *2*(1), 79–116.
Citron, F., & Goldberg, A. (2014). Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of cognitive neuroscience*, *26*(11), 2585–2595.
Claidière, N., Scott-Phillips, T. C., & Sperber, D. (2014). How darwinian is cultural evolution? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1642), 20130368.
Claidière, N., & Sperber, D. (2007). The role of attraction in cultural evolution. *Journal of Cognition and Culture*, *7*(1-2), 89–111.
Cohon, R. (2004). Hume's moral philosophy.
Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376.
Creanza, N., Ruhlen, M., Pemberton, T. J., Rosenberg, N. A., Feldman, M. W., & Ramachandran, S. (2015). A comparison of worldwide phonemic and ge-

netic variation in human populations. *Proceedings of the National Academy of Sciences*, *112*(5), 1265–1272.

Dale, R., & Kello, C. T. (2018). "how do humans make sense?" multiscale dynamics and emergent meaning. *New Ideas in Psychology*, *50*, 61–72.

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in complex systems*, *15*(03n04), 1150017.

Dawkins, R. (1976). *The selfish gene*. Oxford university press.

Deitch, D., Rubin, A., & Ziv, Y. (2020). Representational drift in the mouse visual cortex. *bioRxiv*.

Derex, M., & Boyd, R. (2016). Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, *113*(11), 2982–2987.

Desai, R., Binder, J., Conant, L., Mano, Q., & Seidenberg, M. (2011). The neural career of sensory-motor metaphors. *Journal of cognitive neuroscience*, *23*(9), 2376–2386.

Desai, R., Conant, L., Binder, J., Park, H., & Seidenberg, M. (2013). A piece of the action: modulation of sensory-motor regions by action idioms and metaphors. *NeuroImage*, *83*, 862–869.

Diefenbach, C., Rieger, M., Massen, C., & Prinz, W. (2013). Action-sentence compatibility: the role of action effects and timing. *Frontiers in psychology*, *4*, 272.

Dijkstra, A., & Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision.

Donnelly, S., & Verkuilen, J. (2017). Empirical logit analysis is not logistic regression. *Journal of Memory and Language*, *94*, 28–42.

Dove, G. (2016). Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. *Psychonomic bulletin review*, *23*(4), 1109–1121.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Elman, J. L., Diehl, R. L., & Buchwald, S. E. (1977). Perceptual switching in bilinguals. *The Journal of the acoustical Society of America*, *62*(4), 971–974.

Enfield, N. J. (2014). *Natural causes of language: Frames, biases, and cultural transmission*. Language Science Press.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*(5), 41–60.

Falandays, B., & Spivey, M. J. (2020). 2 theory visualizations for bilingual models of lexical ambiguity resolution. *Bilingual Lexical Ambiguity Resolution*, 17.

Falandays, J., Batzloff, B., Spevack, S., & Spivey, M. (2018). Interactionism in language: From neural networks to bodies to dyads. *Language, Cognition and Neuroscience*, 1–16.

Falandays, J., Brown-Schmidt, S., & Toscano, J. (2020). Long-lasting gradient activation of referents during spoken language processing. *Journal of Memory and Language*, *112*, 104088.

Falandays, J., & Spivey, M. (2019). Abstract meanings may be more dynamic, due to their sociality: Comment on" words as social tools: Language, sociality and

inner grounding in abstract concepts" by anna m. *Borghi et al. Physics of life reviews*, *29*, 175–177.

Falandays, J. B., Batzloff, B. J., Spevack, S. C., & Spivey, M. J. (2020). Interactionism in language: from neural networks to bodies to dyads. *Language, Cognition and Neuroscience*, *35*(5), 543–558.

Falandays, J. B., Nguyen, B., & Spivey, M. J. (2021). Is prediction nothing more than multi-scale pattern completion of the future? *Brain Research*, 147578.

Feiten, T. E. (2020). Mind after uexküll: a foray into the worlds of ecological psychologists and enactivists. *Frontiers in psychology*, *11*, 480.

Fenk-Oczlon, G., & Pilz, J. (2021). Linguistic complexity: Relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. *Frontiers in Communication*, *6*, 66.

Fields, C., & Levin, M. (2020). How do living systems create meaning? *Philosophies*, *5*(4), 36.

Fiorillo, C., Kim, J., & Hong, S. (2014). The meaning of spikes from the neuron's point of view: predictive homeostasis generates the appearance of randomness. *Frontiers in Computational Neuroscience*, *8*, 49.

Fischer, M., & Zwaan, R. (2008). Embodied language: A review of the role of the motor system in language comprehension. *The Quarterly Journal of Experimental Psychology*, *61*(6), 825–850.

Fjell, A. M., & Walhovd, K. B. (2010). Structural brain changes in aging: courses, causes and cognitive consequences. *Rev Neurosci*, *21*(3), 187–221.

Flege, J. E., & Eefting, W. (1987). Cross-language switching in stop consonant perception and production by dutch speakers of english. *Speech Communication*, *6*(3), 185–202.

Francis, W. N., Kucera, H., Kučera, H., & Mackie, A. W. (1982). *Frequency analysis of english usage: Lexicon and grammar*. Houghton Mifflin.

Frankenhuis, W. E., & Panchanathan, K. (2011). Individual differences in developmental plasticity may result from stochastic sampling. *Perspectives on Psychological Science*, *6*(4), 336–347.

Frankenhuis, W. E., & Walasek, N. (2020). Modeling the evolution of sensitive periods. *Developmental cognitive neuroscience*, *41*, 100715.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Gammaitoni, L., Neri, I., & Vocca, H. (2010). The benefits of noise and nonlinearity: extracting energy from random vibrations. *Chemical Physics*, *375*(2-3), 435–438.

Gerkey, D. (2013). Cooperation in context: Public goods games and post-soviet collectives in kamchatka, russia. *Current Anthropology*, *54*(2), 144–176.

Gibbs, R. (1993). Why idioms are not dead metaphors. In *Idioms: Processing, structure, and interpretation* (p. 57–77).

Gijssels, T., Ivry, R., & Casasanto, D. (2018). tdcs to premotor cortex changes action verb understanding: Complementary effects of inhibitory and excitatory stimulation. *Scientific reports*, *8*(1), 1–7.

Glenberg, A., & Kaschak, M. (2002). Grounding language in action. *Psychonomic*

bulletin review, *9*(3), 558–565.
Goldman, M. S. (2004). Enhancement of information transmission efficiency by synaptic failures. *Neural computation*, *16*(6), 1137–1162.
Goldstein, A., Arzouan, Y., & Faust, M. (2012). Killing a novel metaphor and reviving a dead one: Erp correlates of metaphor conventionalization. *Brain and language*, *123*(2), 137–142.
Gong, T., Baronchelli, A., Puglisi, A., & Loreto, V. (2011). Exploring the roles of complex networks in linguistic categorization. *Artificial life*, *18*(1), 107–121.
Grosjean, F., & Nicol, J. (2001). The bilingual's language modes. *The bilingualism reader*, 428–449.
Group", [U+FFFD]., Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., ... others (2009). Language is a complex adaptive system: Position paper. *Language learning*, *59*, 1–26.
Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical evidence: A test of the critical-period hypothesis for second-language acquisition. *Psychological science*, *14*(1), 31–38.
Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346.
Hauk, O. (2016). Only time will tell–why temporal information is essential for our neuroscientific understanding of semantics. *Psychonomic Bulletin & Review*, *23*(4), 1072–1079.
Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307.
Hauk, O., & Tschentscher, N. (2013). The body of evidence: what can neuroscience tell us about embodied semantics? *Frontiers in psychology*, *4*, 50.
Healy, K. (2017). Fuck nuance. *Sociological Theory*, *35*(2), 118–127.
Henrich, J. (2004). Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses: the tasmanian case. *American antiquity*, 197–214.
Henrich, J., & Boyd, R. (2002). On modeling cognition and culture. *Journal of Cognition and Culture*, *2*(2), 87–112.
Henrich, J., Boyd, R., & Richerson, P. J. (2008). Five misunderstandings about cultural evolution. *Human Nature*, *19*(2), 119–137.
Heras-Escribano, M. (2019). Pragmatism, enactivism, and ecological psychology: towards a unified approach to post-cognitivism. *Synthese*, 1–27.
Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.
Hoehl, S., Keupp, S., Schleihauf, H., McGuigan, N., Buttelmann, D., & Whiten, A. (2019). 'over-imitation': A review and appraisal of a decade of research. *Developmental Review*, *51*, 90–108.
Hoffmeyer, J., & Emmeche, C. (2014). Code-duality and the semiotics of nature. In *On semiotic modeling* (pp. 117–166). De Gruyter Mouton.
Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (pan troglodytes) and children (homo sapiens). *Animal cognition*, *8*(3), 164–181.

Hotton, S., & Yoshimi, J. (2010). The dynamics of embodied cognition. *International Journal of Bifurcation and Chaos*, *20*(04), 943–972.

Huette, S., & Anderson, S. (2012). Negation without symbols: The importance of recurrence and context in linguistic negation. *Journal of Integrative Neuroscience*, *11*(03), 295–312.

Hurford, J. R. (1991). The evolution of the critical period for language acquisition. *Cognition*, *40*(3), 159–201.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kallens, P. A. C., Dale, R., & Smaldino, P. E. (2018). Cultural evolution of categorization. *Cognitive Systems Research*, *52*, 765–774.

Karmiloff-Smith, B. A. (1994). Beyond modularity: A developmental perspective on cognitive science. *European journal of disorders of communication*, *29*(1), 95–105.

Ke, J., Minett, J. W., Au, C.-P., & Wang, W. S.-Y. (2002). Self-organization and selection in the emergence of vocabulary. *Complexity*, *7*(3), 41–54.

Kello, C. T. (2013). Critical branching neural networks. *Psychological review*, *120*(1), 230.

Kello, C. T., Brown, G. D., Ferrer-i Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in cognitive sciences*, *14*(5), 223–232.

Kello, C. T., Kerster, B., & Johnson, E. (2011). Critical branching neural computation, neural avalanches, and 1/f scaling. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Kemmerer, D. (2015). Are the motor features of verb meanings represented in the precentral motor cortices? yes, but within the context of a flexible, multilevel architecture for conceptual knowledge. *Psychonomic Bulletin Review*, *22*(4), 1068–1075.

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence, and future directions. *Cortex*, *48*, 805–825.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Komarova, N. L., & Nowak, M. A. (2001). Natural selection of the critical period for language acquisition. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *268*(1472), 1189–1196.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, *97*(22), 11850–11857.

Lacey, S., Stilla, R., Deshpande, G., Zhao, S., Stephens, C., McCormick, K., ...

Sathian, K. (2017). Engagement of the left extrastriate body area during body-part metaphor comprehension. *Brain and language*, *166*, 1–18.

Lacey, S., Stilla, R., & Sathian, K. (2012). Metaphorically feeling: comprehending textural metaphors activates somatosensory cortex. *Brain and language*, *120*(3), 416–421.

Lakoff, G. (1987). The death of dead metaphor. *Metaphor and symbol*, *2*(2), 143–147.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Lauro, L., Mattavelli, G., Papagno, C., & Tettamanti, M. (2013). She runs, the road runs, my mind runs, bad blood runs between us: Literal and figurative motion verbs: An fmri study. *NeuroImage*, *83*, 361–371.

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative science quarterly*, *52*(4), 667–694.

Lewis, H. M., & Laland, K. N. (2012). Transmission fidelity is the key to the build-up of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2171–2180.

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182–195.

Lewontin, R. C. (1972). The apportionment of human diversity. In *Evolutionary biology* (pp. 381–398). Springer.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

Lukosevicius, M., Jaeger, H., & Schrauwen, B. (2012). Reservoir computing trends. *KI-Künstliche Intelligenz*, *26*(4), 365–371.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559.

MacLean, J. N. (2003). Activity-independent homeostasis in rhythmically active neurons. *Neuron*, *37*(1), 109–120.

Mahon, B. (2015). What is embodied about cognition? *Language, cognition and neuroscience*, *30*(4), 420–429.

Mahon, B., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of physiology-Paris*, *102*(1-3), 59–70.

Marian, V., & Spivey, M. (2003a). Bilingual and monolingual processing of competing lexical items. *Applied Psycholinguistics*, *24*(2), 173.

Marian, V., & Spivey, M. (2003b). Competing activation in bilingual language processing: Within-and between-language competition. *Bilingualism*, *6*(2), 97.

Matlock, T. (2004). Fictive motion as cognitive simulation. *Memory & cognition*, *32*(8), 1389–1400.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, *14*(8), 348–356.

McClelland, J. L., Thomas, A. G., McCandliss, B. D., & Fiez, J. A. (1999). Understanding failures of learning: Hebbian learning, competition for representational space, and some preliminary experimental data. *Progress in brain research*, *121*,

75–80.
McDonnell, M. D., & Ward, L. M. (2011). The benefits of noise in neural systems: bridging theory and experiment. *Nature Reviews Neuroscience*, *12*(7), 415–425.
McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42.
Mesoudi, A. (2021). *Cultural evolution*. University of Chicago Press.
Mesoudi, A., & Whiten, A. (2004). The hierarchical transformation of event knowledge in human cultural transmission. *Journal of cognition and culture*, *4*(1), 1–24.
Meteyard, L., Cuadrado, S., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*(7), 788–804.
Mirabella, G., Iaconelli, S., Spadacenta, S., Federico, P., & Gallese, V. (2012). Processing of hand-related verbs specifically affects the planning and execution of arm reaching movements. *PLoS One*, *7*(4).
Miton, H., & Charbonneau, M. (2018). Cumulative culture in the laboratory: Methodological and theoretical challenges. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1879), 20180677.
Moran, S., McCloy, D., & Wright, R. (2012). Revisiting population size vs. phoneme inventory size. *Language*, 877–893.
Morey, R. D., Kaschak, M. P., Díez-Álamo, A. M., Glenberg, A. M., Zwaan, R. A., Lakens, D., ... others (2021). A pre-registered, multi-lab non-replication of the action-sentence compatibility effect (ace). *Psychonomic bulletin & review*, 1–14.
Morin, O. (2013). How portraits turned their eyes upon us: visual preferences and demographic change in cultural evolution. *Evolution and Human Behavior*, *34*(3), 222–229.
Morin, O. (2016). *How traditions live and die*. Oxford University Press.
Müller, C. (2009). *Metaphors dead and alive, sleeping and waking: A dynamic view*. University of Chicago Press.
Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences*, *96*(14), 8028–8033.
Nowak, M. A., Krakauer, D. C., & Dress, A. (1999). An error limit for the evolution of language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *266*(1433), 2131–2136.
Onnis, L., Farmer, T., Baroni, M., Christiansen, M., & Spivey, M. (2008). Generalizable distributional regularities aid fluent language processing: The case of semantic valence tendencies. *Italian Journal of Linguistics*, *20*(1), 125–152.
O'Leary, T., & Wyllie, D. (2011). Neuronal homeostasis: time for a change? *The Journal of Physiology*, *589*(20), 4811–4826.
Panchanathan, K., & Frankenhuis, W. E. (2016). The evolution of sensitive periods in a model of incremental development. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1823), 20152439.
Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. *Journal of Experimental Psychology: Learning,*

Memory, and Cognition, *40*(1), 275.

Pobric, G., Mashal, N., Faust, M., & Lavidor, M. (2008). The role of the right cerebral hemisphere in processing novel metaphoric expressions: a transcranial magnetic stimulation study. *Journal of cognitive neuroscience*, *20*(1), 170–181.

Puglisi, A., Baronchelli, A., & Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences*, *105*(23), 7936–7940.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature reviews neuroscience*, *6*(7), 576–582.

Quadflieg, S., Etzel, J., Gazzola, V., Keysers, C., Schubert, T., Waiter, G., & Macrae, C. (2011). Puddles, parties, and professors: Linking word categorization to neural patterns of visuospatial coding. *Journal of Cognitive Neuroscience*, *23*(10), 2636–2649.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners.* (ArXiv Preprint .)

Rafał, M. (2018). The relationship between the accuracy of cultural transmission and the strength of cultural attractors. *Social Evolution & History*, *17*(2).

Raja, V. (2018). A theory of resonance: Towards an ecological cognitive architecture. *Minds and Machines*, *28*(1), 29–51.

Raja, V. (2021). Resonance and radical embodiment. *Synthese*, *199*(1), 113–141.

Raposo, A., Moss, H., Stamatakis, E., & Tyler, L. (2009). Modulation of motor and premotor cortices by actions, action words and action sentences. *Neuropsychologia*, *47*(2), 388–396.

Ravignani, A., Delgado, T., & Kirby, S. (2016). Musical evolution in the lab exhibits rhythmic universals. *Nature Human Behaviour*, *1*(1), 1–7.

Reali, F., Chater, N., & Christiansen, M. H. (2018). Simpler grammar, larger vocabulary: How population size affects language. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1871), 20172586.

Repp, B. H., & Liberman, A. (1987). Phonetic category boundaries are flexible. In S. N. Harnad (Ed.), *Categorical perception.* New York: Cambridge University Press.

Revill, K. P., & Spieler, D. H. (2012). The effect of lexical frequency on spoken word recognition in young and older listeners. *Psychology and aging*, *27*(1), 80.

Richards, B. A., & Lillicrap, T. P. (2021). The brain-computer metaphor debate is useless: A matter of semantics. *Frontiers in Computer Science*, 11.

Richardson, D., & Matlock, T. (2007). The integration of figurative language and static depictions: An eye movement study of fictive motion. *Cognition*, *102*(1), 129–138.

Richardson, D., Spivey, M., Barsalou, L., & McRae, K. (2003). Spatial representations activated during real[U+2010]time comprehension of verbs. *Cognitive science*, *27*(5), 767–780.

Richardson, D. C., Spivey, M. J., & Cheung, J. (2001). Motor representations in memory and mental models: Embodiement in cognition. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 23).

Rodny, J. J., Shea, T. M., & Kello, C. T. (2017). Transient localist representations

in critical branching networks. *Language, Cognition and Neuroscience*, *32*(3), 330–341.

Rosen, R. (2012). Anticipatory systems. In *Anticipatory systems* (pp. 313–370). Springer.

Rule, M. E., O'Leary, T., & Harvey, C. D. (2019). Causes and consequences of representational drift. *Current opinion in neurobiology*, *58*, 141–147.

Rulke, D. L., & Galaskiewicz, J. (2000). Distribution of knowledge, group network structure, and group performance. *Management Science*, *46*(5), 612–625.

Rüschemeyer, S., Brass, M., & Friederici, A. (2007). Comprehending prehending: neural correlates of processing verbs with motor stems. *Journal of cognitive neuroscience*, *19*(5), 855–865.

Saldana, C., Fagot, J., Kirby, S., Smith, K., & Claidière, N. (2019). High-fidelity copying is not necessarily the key to cumulative cultural evolution: a study in monkeys and children. *Proceedings of the Royal Society B*, *286*(1904), 20190729.

Santana, E., & Vega, M. (2011). Metaphors are embodied, and so are their literal counterparts. *Frontiers in psychology*, *2*, 90.

Schläpfer, M., Bettencourt, L. M., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., ... Ratti, C. (2014). The scaling of human interactions with city size. *Journal of the Royal Society Interface*, *11*(98), 20130789.

Schoonover, C. E., Ohashi, S. N., Axel, R., & Fink, A. J. (2020). Representational drift in primary olfactory cortex. *bioRxiv*.

Schubert, T. (2005). Your highness: vertical positions as perceptual symbols of power. *Journal of personality and social psychology*, *89*(1), 1.

Scorolli, C., & Borghi, A. (2007). Sentence comprehension and action: Effector specific modulation of the motor system. *Brain research*, 119–124.

Scott-Phillips, T., Blancke, S., & Heintz, C. (2018). Four misunderstandings about cultural attraction. *Evolutionary Anthropology: Issues, News, and Reviews*, *27*(4), 162–173.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424.

Shebani, Z., & Pulvermüller, F. (2013). Moving the hands and feet specifically impairs working memory for arm-and leg-related action words. *Cortex*, *49*(1), 222–231.

Skyrms, B. (2010). The flow of information in signaling games. *Philosophical Studies*, *147*(1), 155–165.

Smaldino, P. E. (2014). The cultural evolution of emergent group-level traits. *Behavioral and Brain Sciences*, *37*(3), 243–95.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. *Computational social psychology*, 311–331.

Smaldino, P. E. (2019). Social identity and cooperation in cultural evolution. *Behavioural processes*, *161*, 108–116.

Smart, P. R., Huynh, T. D., Braines, D., & Shadbolt, N. (2010). Dynamic networks and distributed problem-solving.

Smolla, M., & Akçay, E. (2019). Cultural selection shapes network structure. *Science advances*, *5*(8), eaaw0609.

Sperber, D. (1996). Explaining culture: A naturalistic approach. *Cambridge, MA:*

Cambridge.

Spevack, S., Falandays, J., Batzloff, B., & Spivey, M. (2018). Interactivity of language. *Language and Linguistics Compass*, *12*(7), 12282.

Spevack, S. C., Falandays, J. B., Batzloff, B., & Spivey, M. J. (2018). Interactivity of language. *Language and Linguistics Compass*, *12*(7), e12282.

Steels, L. (2008). The symbol grounding problem has been solved. so what's next. *Symbols and embodiment: Debates on meaning and cognition*, 223–244.

Steels, L., Belpaeme, T., et al. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences*, *28*(4), 469–488.

Stramandinoli, F., Marocco, D., & Cangelosi, A. (2012). The grounding of higher order concepts in action and language: a cognitive robotics model. *Neural Networks*, *32*, 165–173.

Strozyk, J., Dudschig, C., & Kaup, B. (2019). Do i need to have my hands free to understand hand-related language? investigating the functional relevance of experiential simulations. *Psychological research*, *83*(3), 406–418.

Szary, J., Kerster, B., & Kello, C. T. (2011). What makes a brain smart? reservoir computing as an approach for general intelligence. In *International conference on artificial general intelligence* (pp. 407–413).

Thompson, B., & Griffiths, T. L. (2021). Human biases limit cumulative innovation. *Proceedings of the Royal Society B*, *288*(1946), 20202752.

Thompson, E. (2010). *Mind in life*. Harvard University Press.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Tomasino, B., Fink, G., Sparing, R., Dafotakis, M., & Weiss, P. (2008). Action verbs and the primary motor cortex: a comparative tms study of silent reading, frequency judgments, and motor imagery. *Neuropsychologia*, *46*(7), 1915–1926.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, *34*(3), 434–464.

Tsao, T., & Tsao, D. Y. (2021). A topological solution to object segmentation and tracking. *arXiv preprint arXiv:2107.02036*.

Tschentscher, N., Hauk, O., Fischer, M., & Pulvermüller, F. (2012). You can count on the motor cortex: finger counting habits modulate motor cortex activation evoked by numbers. *Neuroimage*, *59*(4), 3139–3148.

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human perception and performance*, *24*(3), 830.

Turrigiano, G., & Nelson, S. (2004). Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, *5*(2), 97–107.

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind, revised edition: Cognitive science and human experience*. MIT press.

Vega, M., Moreno, V., & Castillo, D. (2013). The comprehension of action-related sentences may cause interference rather than facilitation on matching actions.

Psychological research, *77*(1), 20–30.
Von Uexküll, J. (1934). A stroll through the worlds of animals and men: A picture book of invisible worlds. *Semiotica*, *89*(4), 319–391.
Wiesenfeld, K., & Moss, F. (1995). Stochastic resonance and the benefits of noise: from ice ages to crayfish and squids. *Nature*, *373*(6509), 33–36.
Willems, R., Hagoort, P., & Casasanto, D. (2010). Body-specific representations of action verbs: Neural evidence from right-and left-handers. *Psychological Science*, *21*(1), 67–74.
Willems, R., Labruna, L., D'Esposito, M., Ivry, R., & Casasanto, D. (2011). A functional role for the motor system in language understanding: evidence from theta-burst transcranial magnetic stimulation. *Psychological science*, *22*(7), 849–854.
Williams, L. (1977). The perception of stop consonant voicing by spanish-english bilinguals. *Perception & Psychophysics*, *21*(4), 289–297.
Wimsatt, W. C. (1972). Complexity and organization. In *Psa: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1972, pp. 67–86).
Winter, A., Dudschig, C., & Kaup, B. (2022). The action-sentence compatibility effect (ace) a benchmark finding for embodiment-a meta-analysis.
Winter, B., & Wedel, A. (2016). The co-evolution of speech and the lexicon: The interaction of functional pressures, redundancy, and category variation. *Topics in cognitive science*, *8*(2), 503–513.
Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*(3), 415–449.
Yang, J., & Shu, H. (2016). Involvement of the motor system in comprehension of non-literal action language: a meta-analysis study. *Brain topography*, *29*(1), 94–107.
Yee, E., Chrysikou, E., Hoffman, E., & Thompson-Schill, S. (2013). Manual experience shapes object representations. *Psychological science*, *24*(6), 909–919.
Yoshimi, J. (2012). Active internalism and open dynamical systems. *Philosophical Psychology*, *25*(1), 1–24.
Zwaan, R. (2014). Embodiment and language comprehension: reframing the discussion. *Trends in cognitive sciences*, *18*(5), 229–234.

Appendix A

Chapter 3: Experiment 1 Model Tables

	Model 1	Model 2	Model 3	Model 4
(Intercept)	2.54***	2.23***	2.28***	2.23***
	(0.12)	(0.16)	(0.16)	(0.16)
conditionLiteral		0.60**	0.60**	0.72**
		(0.22)	(0.22)	(0.23)
congruencyincongruent			−0.10	−0.01
			(0.05)	(0.07)
conditionLiteral:congruencyincongruent				−0.22*
				(0.11)
AIC	9546.22	9540.83	9539.79	9537.99
BIC	9569.15	9571.39	9577.99	9583.83
Log Likelihood	−4770.11	−4766.41	−4764.90	−4762.99
Num. obs.	15374	15374	15374	15374
Num. groups: trialID	179	179	179	179
Num. groups: subject	90	90	90	90
Var: trialID (Intercept)	2.00	1.89	1.89	1.89
Var: subject (Intercept)	0.13	0.13	0.13	0.13

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A.1: Accuracy model comparison results

	Model 1	Model 2
(Intercept)	2.92***	3.05***
	(0.16)	(0.16)
congruencyincongruent		−0.25**
		(0.09)
AIC	4051.84	4045.74
BIC	4072.72	4073.58
Log Likelihood	−2022.92	−2018.87
Num. obs.	7796	7796
Num. groups: subject	90	90
Num. groups: trialID	90	90
Var: subject (Intercept)	0.47	0.47
Var: trialID (Intercept)	1.44	1.43

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A.2: Accuracy model comparison results: Simple effects for literal sentences

	Model 1	Model 2
(Intercept)	2.33***	2.33***
	(0.19)	(0.19)
congruencyincongruent		−0.01
		(0.07)
AIC	5313.39	5315.37
BIC	5334.19	5343.10
Log Likelihood	−2653.70	−2653.68
Num. obs.	7578	7578
Num. groups: subject	90	90
Num. groups: trialID	89	89
Var: subject (Intercept)	0.34	0.34
Var: trialID (Intercept)	2.58	2.58

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A.3: Accuracy model comparison results:: Simple effects for abstract sentences

	Model 1	Model 2	Model 3	Model 4
(Intercept)	6.64***	6.67***	6.66***	6.67***
	(0.03)	(0.03)	(0.03)	(0.03)
conditionLiteral		−0.06*	−0.06*	−0.07**
		(0.02)	(0.02)	(0.03)
congruencyincongruent			0.03***	0.01
			(0.01)	(0.01)
conditionLiteral:congruencyincongruent				0.03*
				(0.01)
AIC	15562.43	15558.85	15548.93	15546.73
BIC	15592.43	15596.35	15593.93	15599.23
Log Likelihood	−7777.22	−7774.42	−7768.47	−7766.37
Num. obs.	13361	13361	13361	13361
Num. groups: trialID	179	179	179	179
Num. groups: subject	90	90	90	90
Var: trialID (Intercept)	0.03	0.03	0.03	0.03
Var: subject (Intercept)	0.06	0.06	0.06	0.06
Var: Residual	0.18	0.18	0.18	0.18

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A.4: Reaction time model comparison results.

	Model 1	Model 2
(Intercept)	6.61***	6.60***
	(0.03)	(0.03)
congruencyincongruent		0.04***
		(0.01)
AIC	8284.18	8270.15
BIC	8311.63	8304.46
Log Likelihood	−4138.09	−4130.07
Num. obs.	7062	7062
Num. groups: subject	90	90
Num. groups: trialID	90	90
Var: subject (Intercept)	0.06	0.06
Var: trialID (Intercept)	0.02	0.02
Var: Residual	0.18	0.18

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A.5: Reaction time model comparison results: Simple effects for literal sentences

	Model 1	Model 2
(Intercept)	6.68***	6.67***
	(0.03)	(0.03)
congruencyincongruent		0.01
		(0.01)
AIC	7216.99	7218.23
BIC	7243.98	7251.97
Log Likelihood	−3604.50	−3604.12
Num. obs.	6299	6299
Num. groups: subject	90	90
Num. groups: trialID	89	89
Var: subject (Intercept)	0.06	0.06
Var: trialID (Intercept)	0.03	0.03
Var: Residual	0.17	0.17

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A.6: Reaction time model comparison results: Simple effects for abstract sentences

	Model 1
(Intercept)	6.81***
	(0.16)
conditionLiteral	−0.08
	(0.22)
congruencyincongruent	0.04
	(0.12)
Lg10PF	0.01
	(0.02)
fam_PC	0.01
	(0.02)
figurativeness	−0.22***
	(0.03)
noun_congruency_bias	−0.07**
	(0.02)
verb_congruency_bias	0.05
	(0.15)
NounLg10CD	−0.03*
	(0.01)
VerbLg10CD	0.03
	(0.02)
vviq	−0.00
	(0.03)
conditionLiteral:congruencyincongruent	−0.19
	(0.16)
conditionLiteral:vviq	0.01
	(0.01)
congruencyincongruent:vviq	0.01
	(0.01)
conditionLiteral:VerbLg10CD	−0.03
	(0.02)
congruencyincongruent:VerbLg10CD	−0.01
	(0.01)
conditionLiteral:NounLg10CD	0.00
	(0.02)
congruencyincongruent:NounLg10CD	0.01
	(0.01)
conditionLiteral:verb_congruency_bias	−0.01
	(0.21)
congruencyincongruent:verb_congruency_bias	−0.07
	(0.28)
conditionLiteral:noun_congruency_bias	0.04
	(0.03)
congruencyincongruent:noun_congruency_bias	0.10*
	(0.05)
conditionLiteral:figurativeness	0.37***
	(0.05)
congruencyincongruent:figurativeness	0.00
	(0.02)
conditionLiteral:fam_PC	−0.10***
	(0.02)
congruencyincongruent:fam_PC	−0.00
	(0.01)
conditionLiteral:Lg10PF	−0.05
	(0.03)
congruencyincongruent:Lg10PF	0.01
	(0.01)
conditionLiteral:congruencyincongruent:vviq	−0.02
	(0.01)
conditionLiteral:congruencyincongruent:VerbLg10CD	0.05**
	(0.02)
conditionLiteral:congruencyincongruent:NounLg10CD	0.00
	(0.02)
conditionLiteral:congruencyincongruent:verb_congruency_bias	−0.07
	(0.39)
conditionLiteral:congruencyincongruent:noun_congruency_bias	−0.09
	(0.05)
conditionLiteral:congruencyincongruent:figurativeness	−0.12**
	(0.04)
conditionLiteral:congruencyincongruent:fam_PC	−0.01
	(0.02)
conditionLiteral:congruencyincongruent:Lg10PF	0.04*
	(0.02)
AIC	15392.91
BIC	15685.42
Log Likelihood	−7657.46
Num. obs.	13361
Num. groups: trialID	179
Num. groups: subject	90
Var: trialID (Intercept)	0.01
Var: subject (Intercept)	0.06
Var: Residual	0.18

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A.7: Summary of omnibus model testing for effects of psycholinguistic variables on congruency effect.

	Model 1
(Intercept)	6.73***
	(0.04)
congruencyincongruent	−0.07**
	(0.03)
VerbLg10CD	0.02
	(0.02)
figurativeness	0.17***
	(0.04)
Lg10PF	−0.09***
	(0.02)
congruencyincongruent:VerbLg10CD	0.02*
	(0.01)
congruencyincongruent:figurativeness	−0.13***
	(0.03)
congruencyincongruent:Lg10PF	0.07***
	(0.01)
AIC	8176.42
BIC	8251.91
Log Likelihood	−4077.21
Num. obs.	7062
Num. groups: subject	90
Num. groups: trialID	90
Var: subject (Intercept)	0.06
Var: trialID (Intercept)	0.01
Var: Residual	0.17

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A.8: Summary of omnibus model testing for effects of psycholinguistic variables on congruency effect: literal sentences.

	Model 1
(Intercept)	6.85***
	(0.04)
congruencyincongruent	0.01
	(0.02)
VerbLg10CD	0.03
	(0.02)
figurativeness	−0.20***
	(0.03)
Lg10PF	0.01
	(0.02)
congruencyincongruent:VerbLg10CD	−0.01
	(0.01)
congruencyincongruent:figurativeness	−0.01
	(0.02)
congruencyincongruent:Lg10PF	0.01
	(0.01)
AIC	7192.17
BIC	7266.40
Log Likelihood	−3585.08
Num. obs.	6299
Num. groups: subject	90
Num. groups: trialID	89
Var: subject (Intercept)	0.06
Var: trialID (Intercept)	0.02
Var: Residual	0.17

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A.9: Summary of omnibus model testing for effects of psycholinguistic variables on congruency effect: abstract sentences.

Appendix B

Chapter 3: Experiment 2 Model Tables

	Model 1	Model 2	Model 3	Model 4
(Intercept)	2.87^{***}	2.14^{***}	2.26^{***}	2.23^{***}
	(0.19)	(0.22)	(0.23)	(0.23)
conditionLiteral		1.43^{***}	1.44^{***}	1.56^{***}
		(0.31)	(0.30)	(0.33)
hand_motion1			-0.24^{*}	-0.18
			(0.10)	(0.12)
conditionLiteral:hand_motion1				-0.22
				(0.22)
AIC	2938.61	2921.43	2918.04	2919.12
BIC	2958.28	2947.65	2950.82	2958.46
Log Likelihood	-1466.31	-1456.71	-1454.02	-1453.56
Num. obs.	5200	5200	5200	5200
Num. groups: trialID	90	90	90	90
Num. groups: subject	61	61	61	61
Var: trialID (Intercept)	2.28	1.69	1.68	1.68
Var: subject (Intercept)	0.25	0.25	0.25	0.25

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.1: Accuracy model comparison results.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	7.07^{***}	7.13^{***}	7.11^{***}	7.11^{***}
	(0.03)	(0.04)	(0.04)	(0.04)
conditionLiteral		-0.12^{***}	-0.12^{***}	-0.11^{**}
		(0.03)	(0.03)	(0.04)
hand_motion1			0.03^{*}	0.03
			(0.01)	(0.02)
conditionLiteral:hand_motion1				-0.01
				(0.03)
AIC	5566.84	5557.22	5553.95	5555.91
BIC	5592.59	5589.41	5592.58	5600.98
Log Likelihood	-2779.42	-2773.61	-2770.98	-2770.95
Num. obs.	4622	4622	4622	4622
Num. groups: trialID	90	90	90	90
Num. groups: subject	61	61	61	61
Var: trialID (Intercept)	0.02	0.02	0.02	0.02
Var: subject (Intercept)	0.04	0.04	0.04	0.04
Var: Residual	0.18	0.18	0.18	0.18

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.2: Reaction time model comparison results.

	Model 1
(Intercept)	0.04
	(0.12)
conditionLiteral	−0.52***
	(0.05)
hand_motion1	−0.07
	(0.05)
tbin1	−0.09
	(0.09)
tbin2	0.93***
	(0.05)
conditionLiteral:hand_motion1	0.16*
	(0.06)
conditionLiteral:tbin1	0.13
	(0.12)
conditionLiteral:tbin2	−1.53***
	(0.08)
hand_motion1:tbin1	0.04
	(0.13)
hand_motion1:tbin2	−0.08
	(0.08)
conditionLiteral:hand_motion1:tbin1	−0.06
	(0.18)
conditionLiteral:hand_motion1:tbin2	−0.00
	(0.11)
AIC	70219.80
BIC	70332.85
Log Likelihood	−35094.90
Num. obs.	13852
Num. groups: trialID	90
Num. groups: subject	61
Var: trialID (Intercept)	0.02
Var: subject (Intercept)	0.75
Var: Residual	9.22

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.3: Summary of omnibus model testing for effects of condition, hand-motion, and time-bin on log odds of fixating 'Abstract' versus 'Literal'

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	0.01	0.02	0.06	0.04	0.04
	(0.11)	(0.12)	(0.12)	(0.12)	(0.12)
tbin1	−0.07	−0.07	−0.07	−0.07	−0.09
	(0.06)	(0.06)	(0.06)	(0.08)	(0.09)
tbin2	0.89***	0.89***	0.90***	0.94***	0.93***
	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)
conditionLiteral	−0.44***	−0.44***	−0.52***	−0.52***	−0.52***
	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)
tbin1:conditionLiteral	0.09	0.09	0.10	0.10	0.13
	(0.09)	(0.09)	(0.09)	(0.09)	(0.12)
tbin2:conditionLiteral	−1.53***	−1.53***	−1.53***	−1.53***	−1.53***
	(0.05)	(0.05)	(0.05)	(0.05)	(0.08)
hand_motion1		−0.02	−0.10**	−0.07	−0.07
		(0.02)	(0.03)	(0.04)	(0.05)
conditionLiteral:hand_motion1			0.16***	0.16**	0.16*
			(0.05)	(0.05)	(0.06)
tbin1:hand_motion1				0.01	0.04
				(0.09)	(0.13)
tbin2:hand_motion1				−0.09	−0.08
				(0.05)	(0.08)
tbin1:conditionLiteral:hand_motion1					−0.06
					(0.18)
tbin2:conditionLiteral:hand_motion1					−0.00
					(0.11)
AIC	70222.48	70223.51	70214.50	70215.93	70219.80
BIC	70290.31	70298.87	70297.40	70313.90	70332.85
Log Likelihood	−35102.24	−35101.75	−35096.25	−35094.96	−35094.90
Num. obs.	13852	13852	13852	13852	13852
Num. groups: trialID	90	90	90	90	90
Num. groups: subject	61	61	61	61	61
Var: trialID (Intercept)	0.02	0.02	0.02	0.02	0.02
Var: subject (Intercept)	0.75	0.75	0.75	0.75	0.75
Var: Residual	9.23	9.23	9.22	9.22	9.22

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.4: Model comparison: Effect of hand-motion on log odds of fixating 'Abstract' versus 'Literal', controlling for time bin and sentence type.

	Model 1	Model 2	Model 3
(Intercept)	-0.46^{***}	-0.51^{***}	-0.52^{***}
	(0.12)	(0.12)	(0.13)
tbin1	0.04	0.04	0.05
	(0.06)	(0.06)	(0.09)
tbin2	-0.70^{***}	-0.70^{***}	-0.68^{***}
	(0.04)	(0.04)	(0.05)
hand_motion1		0.10^{**}	0.11^{**}
		(0.04)	(0.04)
tbin1:hand_motion1			-0.01
			(0.12)
tbin2:hand_motion1			-0.05
			(0.07)
AIC	38255.00	38249.37	38252.95
BIC	38296.52	38297.81	38315.23
Log Likelihood	-19121.50	-19117.69	-19117.47
Num. obs.	7485	7485	7485
Num. groups: subject	61	61	61
Num. groups: trialID	45	45	45
Var: subject (Intercept)	0.86	0.87	0.87
Var: trialID (Intercept)	0.02	0.02	0.02
Var: Residual	8.85	8.84	8.84

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.5: Simple effect of hand motion on log odds of fixating 'Abstract' versus 'Literal' in the literal condition, controlling for time bin.

	Model 1	Model 2	Model 3
(Intercept)	0.00	0.06	0.04
	(0.11)	(0.11)	(0.11)
tbin1	−0.08	−0.09	−0.12
	(0.07)	(0.07)	(0.09)
tbin2	0.96***	0.97***	1.02***
	(0.04)	(0.04)	(0.06)
hand_motion1		−0.12***	−0.08
		(0.03)	(0.05)
tbin1:hand_motion1			0.07
			(0.13)
tbin2:hand_motion1			−0.10
			(0.08)
AIC	31899.29	31888.80	31890.59
BIC	31939.84	31936.12	31951.42
Log Likelihood	−15943.64	−15937.40	−15936.30
Num. obs.	6367	6367	6367
Num. groups: subject	61	61	61
Num. groups: trialID	45	45	45
Var: subject (Intercept)	0.63	0.63	0.63
Var: trialID (Intercept)	0.03	0.03	0.03
Var: Residual	9.30	9.28	9.28

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.6: Simple effect of hand motion on log odds of fixating 'Abstract' versus 'Literal' for abstract sentences, controlling for time bin.

	Model 1	Model 2	Model 3
(Intercept)	−0.32*	−0.32	−0.36*
	(0.16)	(0.17)	(0.17)
conditionLiteral		−0.01	−0.02
		(0.08)	(0.08)
hand_motion1			0.10
			(0.06)
AIC	24352.26	24354.24	24353.67
BIC	24378.01	24386.43	24392.30
Log Likelihood	−12172.13	−12172.12	−12170.84
Num. obs.	4622	4622	4622
Num. groups: trialID	90	90	90
Num. groups: subject	61	61	61
Var: trialID (Intercept)	0.06	0.06	0.06
Var: subject (Intercept)	1.44	1.44	1.45
Var: Residual	7.46	7.46	7.46

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.7: Effects of sentence type and hand motion in analysis window 1 (during the verb).

	Model 1	Model 2	Model 3	Model 4
(Intercept)	−0.33*	−0.35*	−0.39*	−0.34*
	(0.15)	(0.16)	(0.16)	(0.16)
conditionLiteral		0.04	0.03	−0.06
		(0.07)	(0.07)	(0.09)
hand_motion1			0.07	−0.03
			(0.06)	(0.08)
conditionLiteral:hand_motion1				0.20
				(0.12)
AIC	24182.73	24184.40	24184.83	24184.07
BIC	24208.48	24216.59	24223.46	24229.14
Log Likelihood	−12087.36	−12087.20	−12086.41	−12085.03
Num. obs.	4622	4622	4622	4622
Num. groups: trialID	90	90	90	90
Num. groups: subject	61	61	61	61
Var: trialID (Intercept)	0.02	0.02	0.03	0.03
Var: subject (Intercept)	1.33	1.33	1.34	1.34
Var: Residual	7.95	7.95	7.94	7.94

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.8: Effects of sentence type and hand motion on log odds of fixating 'Abstract' versus 'Literal' in analysis window 2 (during the noun).

	Model 1	Model 2	Model 3	Model 4
(Intercept)	−0.12	0.59***	0.61***	0.64***
	(0.11)	(0.08)	(0.08)	(0.08)
conditionLiteral		−1.41***	−1.41***	−1.49***
		(0.04)	(0.04)	(0.06)
hand_motion1			−0.04	−0.11*
			(0.03)	(0.04)
conditionLiteral:hand_motion1				0.14*
				(0.06)
AIC	21555.30	21331.27	21331.77	21328.84
BIC	21581.04	21363.45	21370.38	21373.89
Log Likelihood	−10773.65	−10660.64	−10659.88	−10657.42
Num. obs.	4608	4608	4608	4608
Num. groups: trialID	90	90	90	90
Num. groups: subject	61	61	61	61
Var: trialID (Intercept)	0.54	0.02	0.02	0.02
Var: subject (Intercept)	0.29	0.28	0.28	0.27
Var: Residual	11.36	11.37	11.36	11.35

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.9: Effects of sentence type and hand motion on log odds of fixating 'Abstract' versus 'Literal' in analysis window 3 (after the noun).

	Model 1	Model 2
(Intercept)	−0.39*	−0.48**
	(0.17)	(0.17)
hand_motion1		0.18*
		(0.09)
AIC	13143.87	13141.19
BIC	13167.17	13170.31
Log Likelihood	−6567.94	−6565.60
Num. obs.	2498	2498
Num. groups: subject	61	61
Num. groups: trialID	45	45
Var: subject (Intercept)	1.52	1.54
Var: trialID (Intercept)	0.02	0.03
Var: Residual	7.62	7.60

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.10: Simple effect hand motion on log odds of fixating 'Abstract' versus 'Literal' for literal sentences in analysis window 3 (after the noun).

	Model 1	Model 2
(Intercept)	0.64^{***}	0.71^{***}
	(0.07)	(0.07)
hand_motion1		-0.14^{**}
		(0.05)
AIC	9443.44	9435.85
BIC	9466.08	9464.14
Log Likelihood	−4717.72	−4712.92
Num. obs.	2119	2119
Num. groups: subject	61	61
Num. groups: trialID	45	45
Var: subject (Intercept)	0.17	0.18
Var: trialID (Intercept)	0.03	0.03
Var: Residual	11.52	11.46

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.11: Simple effect hand motion on log odds of fixating 'Abstract' versus 'Literal' for abstract sentences in analysis window 3 (after the noun).

	Model 1	Model 2	Model 3	Model 4
(Intercept)	-0.42^{*}	−0.29	-0.39^{*}	−0.34
	(0.17)	(0.17)	(0.17)	(0.18)
conditionLiteral		-0.23^{**}	-0.23^{**}	-0.31^{**}
		(0.07)	(0.07)	(0.10)
hand_motion1			0.19^{**}	0.11
			(0.07)	(0.10)
conditionLiteral:hand_motion1				0.16
				(0.14)
AIC	5197.35	5190.05	5184.26	5184.90
BIC	5216.53	5215.63	5216.22	5223.26
Log Likelihood	−2595.68	−2591.03	−2587.13	−2586.45
Num. obs.	4417	4417	4417	4417
Num. groups: trialID	90	90	90	90
Num. groups: subject	61	61	61	61
Var: trialID (Intercept)	0.03	0.01	0.01	0.01
Var: subject (Intercept)	1.57	1.56	1.56	1.57

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.12: Effects of condition, hand-motion, and their interaction on the log odds that the first fixation following P.O.D. was to 'Abstract' relative to 'Literal.'

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	-0.46^{*}	-0.51^{**}	-0.46^{*}	-0.48^{*}	-0.49^{*}
	(0.20)	(0.20)	(0.20)	(0.20)	(0.20)
tbin2	0.51^{***}	0.51^{***}	0.51^{***}	0.56^{***}	0.61^{***}
	(0.07)	(0.07)	(0.07)	(0.08)	(0.10)
tbin3	3.13^{***}	3.13^{***}	3.13^{***}	3.11^{***}	3.04^{***}
	(0.10)	(0.10)	(0.10)	(0.11)	(0.13)
conditionLiteral	0.13	0.13	0.05	0.06	0.08
	(0.09)	(0.09)	(0.10)	(0.10)	(0.11)
tbin2:conditionLiteral	-0.70^{***}	-0.70^{***}	-0.70^{***}	-0.70^{***}	-0.80^{***}
	(0.10)	(0.10)	(0.10)	(0.10)	(0.13)
tbin3:conditionLiteral	-3.50^{***}	-3.50^{***}	-3.50^{***}	-3.50^{***}	-3.41^{***}
	(0.12)	(0.12)	(0.12)	(0.12)	(0.16)
hand_motion1		0.09^{*}	-0.02	0.02	0.04
		(0.04)	(0.07)	(0.08)	(0.10)
conditionLiteral:hand_motion1			0.17^{*}	0.15	0.11
			(0.08)	(0.09)	(0.14)
tbin2:hand_motion1				-0.11	-0.22
				(0.10)	(0.14)
tbin3:hand_motion1				0.04	0.20
				(0.11)	(0.19)
tbin2:conditionLiteral:hand_motion1					0.20
					(0.19)
tbin3:conditionLiteral:hand_motion1					-0.20
					(0.23)
AIC	14391.26	14388.97	14386.95	14388.61	14389.54
BIC	14451.55	14456.79	14462.32	14479.05	14495.05
Log Likelihood	-7187.63	-7185.48	-7183.48	-7182.31	-7180.77
Num. obs.	13852	13852	13852	13852	13852
Num. groups: trialID	90	90	90	90	90
Num. groups: subject	61	61	61	61	61
Var: trialID (Intercept)	0.08	0.08	0.08	0.08	0.08
Var: subject (Intercept)	2.05	2.05	2.05	2.05	2.06

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.13: Effects of condition, hand-motion, and time on the log odds of fixating 'Abstract' at least once during analysis window.

	Model 1	Model 2	Model 3
(Intercept)	-0.31	-0.39^{*}	-0.39
	(0.20)	(0.20)	(0.20)
tbin2	-0.19^{**}	-0.19^{**}	-0.18^{*}
	(0.06)	(0.06)	(0.09)
tbin3	-0.37^{***}	-0.37^{***}	-0.37^{***}
	(0.07)	(0.07)	(0.09)
hand_motion1		0.16^{**}	0.16
		(0.05)	(0.09)
tbin2:hand_motion1			-0.02
			(0.13)
tbin3:hand_motion1			0.00
			(0.13)
AIC	8500.39	8493.69	8497.67
BIC	8534.99	8535.22	8553.03
Log Likelihood	-4245.19	-4240.85	-4240.83
Num. obs.	7485	7485	7485
Num. groups: subject	61	61	61
Num. groups: trialID	45	45	45
Var: subject (Intercept)	2.08	2.09	2.09
Var: trialID (Intercept)	0.08	0.08	0.08

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.14: Simple effects of hand-motion and time on the log odds of fixating 'Abstract' at least once during analysis window, within literal sentences.

	Model 1	Model 2	Model 3
(Intercept)	-0.50^{*}	-0.50^{*}	-0.53^{*}
	(0.21)	(0.21)	(0.21)
tbin2	0.53^{***}	0.53^{***}	0.64^{***}
	(0.07)	(0.07)	(0.10)
tbin3	3.26^{***}	3.26^{***}	3.17^{***}
	(0.11)	(0.11)	(0.14)
hand_motion1		−0.01	0.05
		(0.07)	(0.10)
tbin2:hand_motion1			−0.23
			(0.14)
tbin3:hand_motion1			0.18
			(0.19)
AIC	5895.18	5897.13	5895.90
BIC	5928.97	5937.68	5949.97
Log Likelihood	−2942.59	−2942.56	−2939.95
Num. obs.	6367	6367	6367
Num. groups: subject	61	61	61
Num. groups: trialID	45	45	45
Var: subject (Intercept)	2.25	2.25	2.26
Var: trialID (Intercept)	0.10	0.10	0.10

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table B.15: Simple effects of hand-motion and time on the log odds of fixating 'Abstract' at least once during analysis window, within abstract sentences.