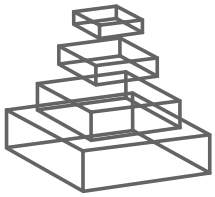
frontiers

RESEARCH TOPICS

# 50 YEARS AFTER THE PERCEPTRON, 25 YEARS AFTER PDP: NEURAL COMPUTATION IN LANGUAGE SCIENCES

Topic Editors
Julien Mayor, Pablo Gomez, Franklin Chang and Gary Lupyan

frontiers in PSYCHOLOGY



# frontiers




Cover image provided by Ibbl sarl, Lausanne CH

**ISSN** 1664-8714
**ISBN** 978-2-88919-257-1
**DOI** 10.3389/978-2-88919-257-1


## ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.
All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!
Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

# 50 YEARS AFTER THE PERCEPTRON, 25 YEARS AFTER PDP: NEURAL COMPUTATION IN LANGUAGE SCIENCES

Topic Editors:
**Julien Mayor,** University of Geneva, Switzerland
**Pablo Gomez,** DePaul, USA
**Franklin Chang,** University of Liverpool, United Kingdom
**Gary Lupyan,** University of Wisconsin, USA

This Research Topic aims to showcase the state of the art in language research while celebrating the 25th anniversary of the tremendously influential work of the PDP group, and the 50th anniversary of the perceptron. Although PDP models are often the gold standard to which new models are compared, the scope of this Research Topic is not constrained to connectionist models. Instead, we aimed to create a landmark forum in which experts in the field define the state of the art and future directions of the psychological processes underlying language learning and use, broadly defined. We thus called for papers involving computational modeling and original research as well as technical, philosophical, or historical discussions pertaining to models of cognition. We especially encouraged submissions aimed at contrasting different computational frameworks, and their relationship to imaging and behavioral data.

# Table of Contents


**04** ***Connectionism Coming of Age: Legacy and Future Challenges***
Julien Mayor, Pablo Gomez, Franklin Chang and Gary Lupyan

**07** ***Recurrent Temporal Networks and Language Acquisition—From Corticostriatal Neurophysiology to Reservoir Computing***
Peter Ford Dominey

**21** ***The Stability-Plasticity Dilemma: Investigating the Continuum From Catastrophic Forgetting to Age-Limited Learning Effects***
Martial Mermillod, Aurélia Bugaiska and Patrick Bonin

**24** ***An Amodal Shared Resource Model of Language-Mediated Visual Attention***
Alastair Charles Smith, Padraic Monaghan and Falk Huettig

**40** ***Integrating Probabilistic Models of Perception and Interactive Neural Networks: A Historical and Tutorial Review***
James L. McClelland

**65** ***Modeling Language and Cognition With Deep Unsupervised Learning: A Tutorial Overview***
Marco Zorzi, Alberto Testolin and Ivilin Peev Stoianov

**79** ***Spoken Word Recognition Without a Trace***
Thomas Hannagan, James S. Magnuson and Jonathan Grainger

**96** ***Deep Generative Learning of Location-Invariant Visual Word Recognition***
Maria Grazia Di Bono and Marco Zorzi

**106** ***A Computational Model to Investigate Assumptions in the Headturn Preference Procedure***
Christina Bergmann, Louis ten Bosch, Paula Fikkert and Lou Boves

**121** ***Experience and Generalization in a Connectionist Model of Mandarin Chinese Relative Clause Processing***
Yaling Hsiao and Maryellen C. MacDonald

**140** ***Self-Organizing Map Models of Language Acquisition***
Ping Li and Xiaowei Zhao

**155** ***Context, Cortex, and Associations: A Connectionist Developmental Approach to Verbal Analogies***
Pavlos Kollias and James L. McClelland

**169** ***Beyond Modeling Abstractions: Learning Nouns Over Developmental Time in Atypical Populations and Individuals***
Clare E. Sims, Savannah M. Schilling and Eliana Colunga

frontiers in PSYCHOLOGY

EDITORIAL
published: 04 March 2014
doi: 10.3389/fpsyg.2014.00187

# Connectionism coming of age: legacy and future challenges

***Julien Mayor[1]*, Pablo Gomez[2], Franklin Chang[3] and Gary Lupyan[4]***

[1] *Department of Psychology and Educational Sciences, University of Geneva, Genève, Switzerland*
[2] *Department of Psychology, De Paul University, Chicago, IL, USA*
[3] *Department of Psychological Sciences, University of Liverpool, Liverpool, UK*
[4] *Department of Psychology, University of Wisconsin, Madison, WI, USA*
**Correspondence: julien.mayor@unige.ch*

***Edited by:***
*Manuel Carreiras, Basque Center on Cognition, Brain and Language, Spain*

**Keywords: recurrent networks, interactive processing, probabilistic cognition, computational modeling, language acquisition, language processing, speech perception, computational linguistics**

## ABOUT 50 YEARS AFTER THE INTRODUCTION OF THE PERCEPTRON AND SOME 25 YEARS AFTER THE INTRODUCTION OF PDP MODELS, WHERE ARE WE NOW?

In 1986, Rumelhart and McClelland took the cognitive science community by storm with the Parallel Distributed Processing (PDP) framework. Rather than abstracting from the biological substrate as was sought by the "information processing" paradigms of the 1970s, connectionism, as it has come to be called, embraced it. An immediate appeal of the connectionist agenda was its aim: to construct at the algorithmic level models of cognition that were compatible with their implementation in the biological substrate.

The PDP group argued that this could be achieved by turning to networks of artificial neurons, originally introduced by McCulloch and Pitts (1943) which the group showed were able to provide insights into a wide range of psychological domains, from categorization, to perception, to memory, to language. This work built on an earlier formulation by Rosenblatt (1958) who introduced a simple type of feed-forward neural network called the perceptron. Perceptrons were limited to solving simple linearly-separable problems and although networks composed of perceptrons were known to be able to compute any Boolean function (including XOR, Minsky and Papert, 1969), there was no effective way of training such networks. In 1986, Rumelhart, Hinton and Williams introduced the back-propagation algorithm, providing an effective way of training multi-layered neural networks, which could easily learn non linearly-separable functions. In addition to providing the field with an effective learning algorithm, the PDP group published a series of demonstrations of how long standing questions in cognitive psychology could be elegantly solved using simple learning rules, distributed representations, and interactive processing.

To take a classic example, consider the word-superiority effect, in which people can detect letters within a word faster than individual letters or letters within a non-word (Reicher, 1969). This result is difficult to square with serial "information-processing" theories of cognition that were dominant at the time (how could someone recognize "R" before "FRIEND" if recognizing the word required recognizing the letters?). Accounting for such findings demanded a framework which could naturally accommodate interactive processes within a bidirectional flow of information. The so-called "Interactive-activation model" (McClelland and Rumelhart, 1981) provided just such a framework.

The connectionist paradigm was not without its critics. The principal critiques can be divided into three classes. First, some neuroscientists (Crick, 1989) questioned the biological plausibility of backpropagation, when they failed to observe experimentally complex and differentiated back-propagating signals that are required to learn in multi-layered neural networks. A second critique concerned stability-plasticity of the learned representations in these models. Some phenomena require the ability to rapidly learn new information, but sometimes newly learned knowledge overwrites previously learned information (catastrophic interference; McCloskey and Cohen, 1989). Third, representing spatial and temporal invariance—something that apparently came easily to people—was difficult for models, e.g., recognizing that the letter "T" in "TOM" was the "same" as the "T" in "POT." This invariance problem was typically solved by multiplying a large number of hard-wired units that were space- or time-locked (see e.g., McClelland and Elman, 1986). Finally, critics pointed out that the networks were incapable of learning true rules on which a number of human behavioral, namely language-learning was thought to depend (e.g., Marcus, 2003; cf. Fodor and Pylyshyn, 1988; Seidenberg, 1999).

The connectionist approach has embraced these challenges: Although some connectionist models continue to rely on back-propagation, others have moved to more biologically realistic learning rules (Giese and Poggio, 2003; Masquelier and Thorpe, 2007). Far from being a critical flaw of connectionism, the phenomenon of catastrophic interference (Mermillod et al., 2013) proved to be a feature that led to the development of complementary learning systems (McClelland et al., 1995).

Progress has also been made on the invariance problem. For example, within the speech domain representing the similarity between similar speech sounds regardless of their location within a word has been addressed in the past by Grossberg and Myers (2000) and Norris (1994) and this issue presents a new more streamlined and computationally efficient model (Hannagan et al., 2013). An especially powerful approach to solving the location invariance problem in the visual domain is presented by Di Bono and Zorzi (2013), also in this issue.

A key challenge for connectionism is to explain the learning of abstract structural representations. The use of recurrent networks (Elman, 1990; Dominey, 2013) and self-organizing maps, has captured important aspects of language learning (e.g., Mayor and Plunkett, 2010; Li and Zhao, 2013), while work on deep learning (Hinton and Salakhutdinov, 2006) has made it possible to model the emergence of structured and abstract representations within multi-layered hierarchical networks (Zorzi et al., 2013). The work on verbal analogies by Kollias and McClelland (2013) continues to address the challenges of modeling more abstract representations, but truly understanding how neural architectures give rise to symbolic cognition is a gap that remains. Although learning and representing formal language rules may not be completely outside of the abilities of neural networks (e.g., Chang, 2009), it seems clear that understanding human cognition requires understanding how we solve these symbolic problems (Clark and Karmiloff-Smith, 1993; Lupyan, 2013). Future generations of connectionist modelers may wish to fill this gap and in so doing provide a fuller picture of how neural networks give rise to intelligence of the sort enables us to ponder the very workings of our cognition.

## WHAT'S NEXT?

The articles assembled in this issue demonstrate the range of topics currently addressed by connectionist models: from word learning in atypical populations (Sims et al., 2013), to sentence processing (Hsiao and MacDonald, 2013), to multimodal processing (Bergmann et al., 2013), to interactions between language and vision (Smith et al., 2013). We expect this diversity to continue to increase. We also hope to see greater increasing integration between connectionism and a computationally similar but philosophically distinct models employing Bayesian inference. Although the computational similarities between these two approaches have been previously recognized (McClelland, 1998), detailed tutorials like the one contained in this volume (McClelland, 2013) provide new clarity on the relationship between these two approaches.

The influence of theoretical constructs introduced by the connectionist approach have become part and parcel of cognitive science (although they are now often not accompanied by the label "connectionism" or "PDP"). The distributed representations that challenge classical symbolic models and which emerge naturally in neural networks are now no longer theoretical constructs and can be directly observed in the brain (Kriegeskorte et al., 2008; Chang et al., 2010). Evidence for rapid warping of these representations by task demands (of the sort described by e.g., McClelland and Rogers, 2003) is also being confirmed through modern neuroimaging (e.g., Çukur et al., 2013) [1]. Many connectionist models have stressed prediction as a way of learning structure and statistical inputs (e.g., Dell and Chang, 2014). This too finds wide support in contemporary neuroscience (Friston, 2010) leading some to even argue that prediction is the unifying feature of all cognitive and perceptual processes (Clark, 2013, for review). Interactive processing—another core feature of the connectionist paradigm—has become similarly foundational. The interplay between bottom-up and top-down information is now recognized to be critical from everything as simple as simply detecting the presence of a visual stimulus, to consciousness itself (e.g., Dehaene et al., 2003; Gilbert and Sigman, 2007; Lupyan and Ward, 2013).

Finally, contemporary neural networks, most notably those utilizing so called deep-learning, have found success in solving practical problems such as image and speech recognition, and natural language processing. For example, algorithms based on the deep-learning approach are now used by Google to extract high-level features from images with, in some cases, above-human performance (Ciresan et al., 2011; Le et al., 2011).

[1] It is useful to note that the methods that make these analyses possible, most notably multi-voxel pattern analyses (MVPA, e.g., Norman et al., 2006) and "representational dissimilarity matrices" (Kriegeskorte et al., 2008) are adaptations of methods developed for analyzing dynamics of artificial neural networks.

## ACKNOWLEDGMENTS

This work is supported by the Swiss National Science Foundation Grant 131700 awarded to Julien Mayor. Franklin Chang is supported by Leverhulme Trust Research Project Grant (RPG-158).

## REFERENCES

Bergmann, C., ten Bosch, L., Fikkert, P., and Boves, L. (2013). A computational model to investigate assumptions in the headturn preference procedure. *Front. Psychol.* 4:676. doi: 10.3389/fpsyg.2013.00676

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* 13, 1428–1432. doi: 10.1038/nn.2641

Chang, F. (2009). Learning to order words: a connectionist model of heavy NP shift and accessibility effects in Japanese and English. *J. Mem. Lang.* 61, 374–397. doi: 10.1016/j.jml.2009.07.006

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. (2011). "Flexible, high performance convolutional neural networks for image classification," in *International Joint Conference on Artificial Intelligence IJCAI-2011.* (Barcelona), 1237–1242.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Clark, A., and Karmiloff-Smith, A. (1993). The cognizer's innards: a psychological and philosophical perspective on the development of thought. *Mind Lang.* 8, 487–519. doi: 10.1111/j.1468-0017.1993.tb00299.x

Crick, F. (1989). The recent excitement about neural networks. *Nature* 337, 129–132. doi: 10.1038/337129a0

Çukur, T., Nishimoto, S., Huth, A. G., and Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nat. Neurosci.* 16, 763–770. doi: 10.1038/nn.3381

Dehaene, S., Sergent, C., and Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8520–8525. doi: 10.1073/pnas.1332574100

Dell, G. S., and Chang, F. (2014). The P-Chain: Relating sentence production and its disorders to comprehension and acquisition. *Philos. Trans. R. Soc. B Biol. Sci.* 369, 20120394. doi: 10.1098/rstb.2012.0394

Di Bono, M. G., and Zorzi, M. (2013). Deep generative learning of location-invariant visual word recognition. *Front. Psychol.* 4:635. doi: 10.3389/fpsyg.2013.00635

Dominey, P. F. (2013). Recurrent temporal networks and language acquisition—from corticostriatal neurophysiology to reservoir computing. *Front. Psychol.* 4:500. doi: 10.3389/fpsyg.2013.00500

Elman, J. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–212. doi: 10.1207/s15516709cog1402_1

Fodor, J. A., and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71. doi: 10.1016/0010-0277(88)90031-5

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Giese, M., and Poggio, T. (2003). Neural mechanisms for the recognition of biological movements and action. *Nat. Rev. Neurosci.* 4, 179–192. doi: 10.1038/nrn1057

Gilbert, C. D., and Sigman, M. (2007). Brain states: top-down influences in sensory processing. *Neuron* 54, 677–696. doi: 10.1016/j.neuron.2007.05.019

Grossberg, S., and Myers, C. W. (2000). The resonant dynamics of speech perception: interword integration and duration-dependent backward effects. *Psychol. Rev.* 107, 735. doi: 10.1037/0033-295X.107.4.735

Hannagan, T., Magnuson, J. S., and Grainger, J. (2013). Spoken word recognition without a TRACE. *Front. Psychol.* 4:563. doi: 10.3389/fpsyg.2013.00563

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hsiao, Y., and MacDonald, M. C. (2013). Experience and generalization in a connectionist model of Mandarin Chinese relative clause processing. *Front. Psychol.* 4:767. doi: 10.3389/fpsyg.2013.00767

Kollias, P., and McClelland, J. L. (2013). Context, cortex, and associations: a connectionist developmental approach to verbal analogies. *Front. Psychol.* 4:857. doi: 10.3389/fpsyg.2013.00857

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141. doi: 10.1016/j.neuron.2008.10.043

Le, Q. V., Ranzato, M.'A., Monga, R., Devin, M., Chen, K., Corrado, G. S., et al. (2011). Building high-level features using large scale unsupervised learning. *arXiv preprint arXiv:*1112.6209.

Li, P., and Zhao, X. (2013). Self-organizing map models of language acquisition. *Front. Psychol.* 4:828. doi: 10.3389/fpsyg.2013.00828

Lupyan, G. (2013). The difficulties of executing simple algorithms: why brains make mistakes computers don't. *Cognition* 129, 615–636. doi: 10.1016/j.cognition.2013.08.015

Lupyan, G., and Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proc. Natl. Acad. Sci. U.S.A.* 110, 14196–14201. doi: 10.1073/pnas.1303312110

Marcus, G. F. (2003). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge: The MIT Press.

Masquelier, T., and Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput. Biol.* 3:e31. doi: 10.1371/journal.pcbi.0030031

Mayor, J., and Plunkett, K. (2010). A neurocomputational model of taxonomic responding and fast mapping in early word learning. *Psychol. Rev.* 117, 1–31. doi: 10.1037/a0018130

McClelland, J. L. (1998). "Connectionist models and Bayesian inference," in *Rational Models of Cognition*, eds M. Oaksford and N. Chater (Oxford: Oxford University Press), 21–53.

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review. *Front. Psychol.* 4:503. doi: 10.3389/fpsyg.2013.00503

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning-systems in the hippocampus and neocortex - insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102, 419–457.

McClelland, J. L., and Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nat. Rev. Neurosci.* 4, 310–322. doi: 10.1038/nrn1076

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychol. Rev.* 88, 375–407. doi: 10.1037/0033-295X.88.5.375

McCloskey, M., and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. *Psychol. Learn. Motiv.* 24, 109–164. doi: 10.1016/S0079-7421(08)60536-8

McCulloch, W. S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133. doi: 10.1007/BF02478259

Mermillod, M., Bugaiska, A., and Bonin, P. (2013). The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Front. Psychol.* 4:504. doi: 10.3389/fpsyg.2013.00504

Minsky, M. L., and Papert, S. A. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430. doi: 10.1016/j.tics.2006.07.005

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition* 52, 189–234. doi: 10.1016/0010-0277(94)90043-4

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *J. Exp. Psychol.* 81, 275. doi: 10.1037/h0027768

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386. doi: 10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors *Nature* 323, 9. doi: 10.1038/323533a0

Rumelhart, D. E., and McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Vol. 1. Cambridge, MA: Foundations MIT Press.

Seidenberg, M. S. (1999). Do infants learn grammar with algebra or statistics? *Science* 284, 434–435. author reply: 436–437.

Sims, C. E., Schilling, S. M., and Colunga, E. (2013). Beyond modeling abstractions: learning nouns over developmental time in atypical populations and individuals. *Front. Psychol.* 4:871. doi: 10.3389/fpsyg.2013.00871

Smith, A. C., Monaghan, P., and Huettig, F. (2013). An amodal shared resource model of language-mediated visual attention. *Front. Psychol.* 4:528. doi: 10.3389/fpsyg.2013.00528

Zorzi, M., Testolin, A., and Stoianov, I. P. (2013). Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Front. Psychol.* 4:515. doi: 10.3389/fpsyg.2013.00515

*Received: 13 February 2014; accepted: 15 February 2014; published online: 04 March 2014.*

*Citation: Mayor J, Gomez P, Chang F and Lupyan G (2014) Connectionism coming of age: legacy and future challenges. Front. Psychol. **5**:187. doi: 10.3389/fpsyg.2014.00187 This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 05 August 2013
doi: 10.3389/fpsyg.2013.00500

# Recurrent temporal networks and language acquisition—from corticostriatal neurophysiology to reservoir computing

***Peter F. Dominey****

*Robot Cognition Laboratory, Centre National de la Recherche Scientifique and INSERM Stem Cell and Brain Research Institute, Bron Cedex, France*

***Edited by:***
*Franklin Chang, University of Liverpool, UK*

***Reviewed by:***
*Matthias Schlesewsky, Johannes Gutenberg University Mainz, Germany*
*Hartmut Fitz, Max Planck Institute for Psycholinguistics, Netherlands*

****Correspondence:***
*Peter F. Dominey, Robot Cognition Laboratory, Centre National de la Recherche Scientifique and INSERM Stem Cell and Brain Research Institute, 18 Avenue Doyen Lepine, 69675 Bron Cedex, France*
*e-mail: peter.dominey@inserm.fr*

One of the most paradoxical aspects of human language is that it is so unlike any other form of behavior in the animal world, yet at the same time, it has developed in a species that is not far removed from ancestral species that do not possess language. While aspects of non-human primate and avian interaction clearly constitute communication, this communication appears distinct from the rich, combinatorial and abstract quality of human language. So how does the human primate brain allow for language? In an effort to answer this question, a line of research has been developed that attempts to build a language processing capability based in part on the gross neuroanatomy of the corticostriatal system of the human brain. This paper situates this research program in its historical context, that begins with the primate oculomotor system and sensorimotor sequencing, and passes, via recent advances in reservoir computing to provide insight into the open questions, and possible approaches, for future research that attempts to model language processing. One novel and useful idea from this research is that the overlap of cortical projections onto common regions in the striatum allows for adaptive binding of cortical signals from distinct circuits, under the control of dopamine, which has a strong adaptive advantage. A second idea is that recurrent cortical networks with fixed connections can represent arbitrary sequential and temporal structure, which is the basis of the reservoir computing framework. Finally, bringing these notions together, a relatively simple mechanism can be built for learning the grammatical constructions, as the mappings from surface structure of sentences to their meaning. This research suggests that the components of language that link conceptual structure to grammatical structure may be much simpler that has been proposed in other research programs. It also suggests that part of the residual complexity is in the conceptual system itself.

**Keywords: reservoir computing, recurrent network, P600, grammatical construction, striatum**

## INTRODUCTION

We begin with the neurophysiological basis of orienting behavior, which provides the framework that leads to language. In a dynamic and changing world, filled with predators and prey, the ability to rapidly orient one's spatial attention to the right place is a question of survival. In mammals with mobile heads (e.g., cats) and primates with mobile eyes (monkeys and man), the ability to orient allows these animals to control their attention to the environment with high precision, and with a temporal reactivity on the scale of hundreds of milliseconds—fractions of a second. In the 1980's research in the oculomotor system of the cat and macaque monkey reached a certain height of completion, and the neural circuits that processed information from visual input to motor response were specified at a fairly high level of detail [reviewed in Dominey and Arbib (1992)]. This represented an important phase in cognitive neuroscience, because the same circuits that specified motor control and spatial attention in the oculomotor system were templates for parallel circuits that would provide part of the basis for higher cognitive function and language.

In this context, one of the major architectural properties of the primate brain is the massive organized projection from cortex to striatum (Selemon and Goldman-Rakic, 1985; Lehericy et al., 2004). Essentially all of neocortex projects in a topographically organized manner to the striatum, through the pallidum to the thalamus and back to cortex (Ilinsky et al., 1985), thus yielding what can be considered as a set of largely distinct and segregated corticostriatal circuits or loops (see **Figure 1**), dedicated to distinct functions, including control of different motor systems such as the oculomotor system, and the limbic reward system (Alexander et al., 1986). This paper will argue that this notion can be extended to a cortico-striatal language circuit (Dominey and Inui, 2009; Dominey et al., 2009).

The closed loop structure provides a feedback of the results of the outcome of the system back into cortex. Such feedback connections have been demonstrated to play an important role in memory and sequence processing (Jaeger and Haas, 2004; Jaeger et al., 2007).

At the same time that the functional neuroanatomy of the oculomotor loop had been quite well characterized and modeled

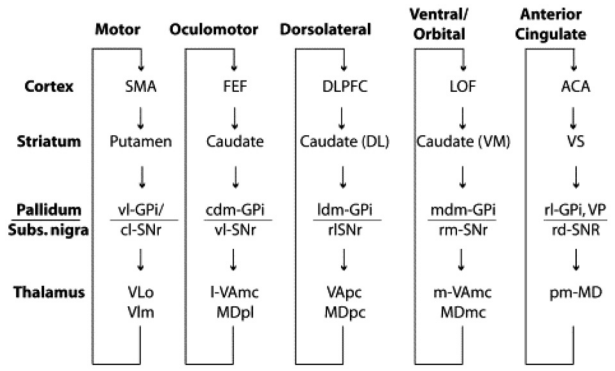



**FIGURE 1 | Parallel organization of the cortico-striato-thalamo-cortical loops.** From Grahn et al. (2009), modified from Alexander et al. (1986).

in a neurophysiologically realistic manner (Dominey and Arbib, 1992), the mechanisms for dopamine modulated plasticity in the corticostriatal synapse (Calabresi et al., 1992) that could lead to adaptive behavior were also being characterized (Robbins et al., 1990; Reading et al., 1991). For example, Reading and Robins demonstrated how the lateral caudate-putamen is required for the learning of arbitrary stimulus-response associations (Reading et al., 1991), which were also impaired in the absence of cortico-striatal dopamine (Robbins et al., 1990).

This inspired us to consider that the cortico-striatal junction could be used as a convergence point where information from different modalities could be functionally linked by dopamine-modulated cortico-striatal plasticity (Dominey et al., 1995). Indeed, while the "central dogma" of the corticostriatal system presents a parallel and segregated set of loops as illustrated in **Figure 1**, from the beginning this was known to be a simplification (Selemon and Goldman-Rakic, 1985), as in fact, the projections from cortex to striatum display a more complex topography. While the main and most dense projections follow the parallel circuit concept, more diffuse projections form larger territories, leading to large overlap of the different circuits (Selemon and Goldman-Rakic, 1985). These overlaps provide a crucial function—they allow the adaptive binding together of cortical signals from different functional circuits. Thus, for example, visual features from infero-temporal (IT) cortex can become linked to direction eye movements (saccades) to different locations in space. We modeled this framework by extending the oculomotor model so that the oculomotor region of the caudate received inputs from the oculomotor frontal eye fields, consistent with the parallel circuits in **Figure 1**, and in addition it received projections from the inferior temporal cortex, consistent with known neuroanatomy (Selemon and Goldman-Rakic, 1985), which code the features of visual stimuli. The resulting model provided the first account of how reward-related dopamine could strengthen corticostriatal synapses binding stimulus coding to behavioral response coding (Dominey et al., 1995). The relevance of this historical interlude into the functional neuroanatomy of the corticostriatal system will soon become apparent, as we make the link from associative learning, to sequence learning to language.

## SERIAL, TEMPORAL AND ABSTRACT STRUCTURE AND THE INITIAL STATE

Twenty-five years ago, Barone and Joseph (1989) studied neural activity in the prefrontal cortex (PFC) of monkeys that had been trained to perform a simple task that involved watching the presentation of a visual sequence on a response button board, and then after a short delay, reproducing the sequence by touching the buttons on the board in the same order that they were presented. They observed that neurons in the dorsolateral prefrontal cortex (DLPFC) displayed two characteristic responses to stimuli in the sequence task. First, as had previously been observed, the neurons were spatially selective, with preferences for stimuli in particular locations in the retinal image. The second characteristic was new, and revolutionary: many of these neurons also displayed a "sequence rank" effect, that is, they had preferences for stimuli that had appeared first, second or last in the input sequence. Thus, the spatial selectivity in many neurons was modulated by the rank or order of the element in the sequence. This indicated that DLPFC embodies a mechanism for discriminating the order of items in a perceptual sequence.

In an effort to understand how this order-sensitive property could result from principal characteristics of the PFC, we recalled that a second major architectural property of the primate brain (the first being the massive cortico-striatal projection) is the abundance of local connectivity in cortico-cortical connections, or recurrent connections, particularly in the frontal cortex (Goldman-Rakic, 1987). Recurrent connections in neural networks provide known dynamical system properties, and indeed in the context of Elman's simple recurrent network (SRN) the power of recurrent connections in language-related processing was revealed (Elman, 1990, 1991). Intuitively, recurrent connections allow information from past events to remain coded, circulating through these connections, and thus allowing the past to influence the coding of new inputs. This provides an intrinsic sequence coding capability.

The use of recurrent connections in the context of synaptic adaptation also unveiled the immense technical challenge of determining how a given recurrent connection contributed to error in the network response, since over multiple time steps the state of activation in the recurrent network changes dynamically (Pearlmutter, 1995). The solution developed by Elman was to limit the simulation of cycles through the recurrent net to one or two time steps. This provided a dramatic simplification of the learning algorithm while preserving the essential property of recurrent connections. This introduced a significant limitation, however, with respect to the processing of time.

A principal objective of computational neuroscience is to simulate and explain neural activity over the time course of the behavioral experiment. Thus, in Barone and Joseph's sequenceing task, stimuli are presented for a certain duration, and the subsequent execution of the sequence by the animal unfolds in time, including the reaction times for the individual responses. The simplification in the SRN renders such realistic treatment of time impossible, as the time between successive sequence elements is fixed by the learning algorithm.

In order to circumvent the technical challenges of recurrent learning in time, we chose an alternative approach. We decided

to fix the connection strengths of the recurrent connections at the outset so as to provide the simulated PFC network with a dynamic structure that would retain a trace of previous inputs via the recurrent connections. The resulting patterns of activity in the cortical network could then be associated with the corresponding behavioral outputs by reward-related (dopaminergic) plasticity in the corticostriatal synapses (Dominey, 1995; Dominey et al., 1995).

The resulting system is illustrated in **Figure 2**. The principal characteristics are the presence of fixed recurrent connections in the PFC network (corresponding to the DLPFC in **Figure 1**), and modifiable connections between these PFC neurons and the neurons in the striatum (caudate nucleus—CD), which form an associative memory, associating dynamic states in the recurrent network with the desired output response. It is noteworthy that this combination of fixed recurrent connections, and modifiable connections to "readout" neurons was the first characterization of what has now come to be known as the reservoir principle in reservoir computing (Maass et al., 2002; Lukosevicius and Jaeger, 2009). The resulting network displayed a number of interesting properties.

First, it was able to explain the behavior of monkeys in the Barone and Joseph sequence learning task, and more interestingly, the neural activity in simulated PFC neurons displayed the same combination of spatial and sequence rank coding properties as those recorded in the monkey PFC (Dominey et al., 1995). In particular, the simulated PFC neurons were spatially selective, and this spatial selectivity was modulated by the rank of the spatial target in the sequence. This was a computational neuroscience success.

Second, the model displayed a robust sequence learning capability. **Figure 3** illustrates the dynamic activity within PFC neurons during the presentation and replay of a 25 element sequence. One can observe that the pattern of activation in PFC (recurrent network) neurons displays a rich dynamic behavior, and that indeed, the states in PFC corresponding to the different elements in the sequence are indeed separable, as revealed by the observation that the cosines of the state vectors are never equal to unity (i.e., the state vectors are never identical). In this context, the model could account for (Dominey, 1998a) and predict (Dominey, 1998b) human sequence learning behavior in the well studied serial reaction time (SRT) task. Because the connections from cortex to striatum are modified by learning, neurons in the striatum become activated with reduced reaction times in cases where learning is significant. That is, when responding to visual inputs presented in a well-learned sequence, stronger learned cortico-striatal connections lead to faster activation of the striatal response neurons, leading to a reduced number of simulation time steps for generating the model output. For elements presented in a random sequence, there was no learning effect, and significantly more time steps were required to generate the response in the striatal neurons. Details can be found in Dominey (1998a,b).

While the model thus provided a robust learning for serial and temporal structure, it failed to learn what we called abstract structure. Serial and abstract structure are defined such that the two sequences ABCBAC and DEFEDF have different serial structure (i.e., the serial order of specific sequence elements), but identical abstract structure (i.e., the relations between repeating elements within the sequence), following an internal repetitive

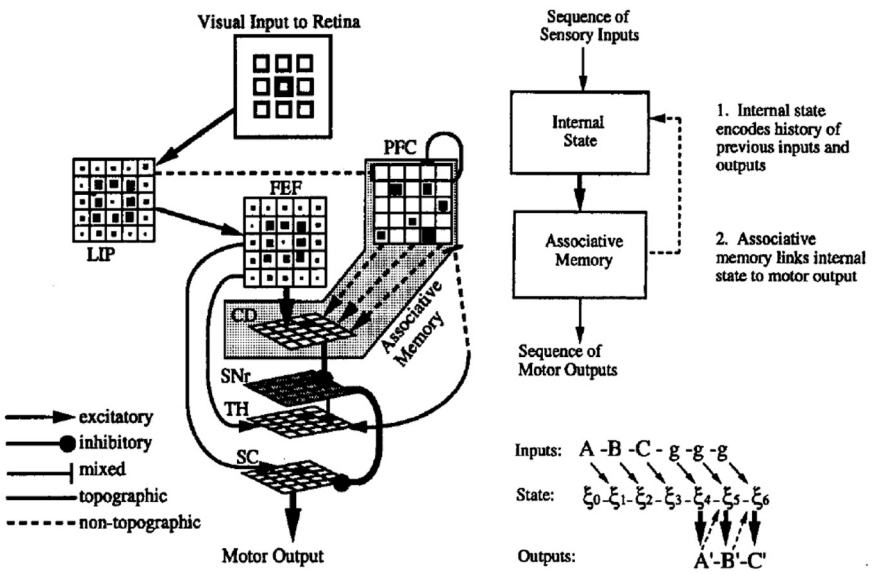



**FIGURE 2 | Model of cortico-striatal system for sensorimotor sequence learning. Left**—neuroanatomical structure of model. Visual input to simulated retina projects to lateral interparietal cortex (LIP) and frontal fields (FEF), and prefrontal cortex (PFC) (via mixed connections). PFC has recurrent connections, rendering it a rich dynamical system, and projects with modifiable connections to the caudate nucleus of the striatum (CD), which serves to activate the motor superior colliculus (SC) via the oculomotor circuit. **Right**—synthetic view. Recurrent PFC network encodes internal state, and projects with modifiable connects to associative memory. Feedback connections from associative memory to internal state allow state to be influenced by the results of the learned associations. From Dominey (1995).

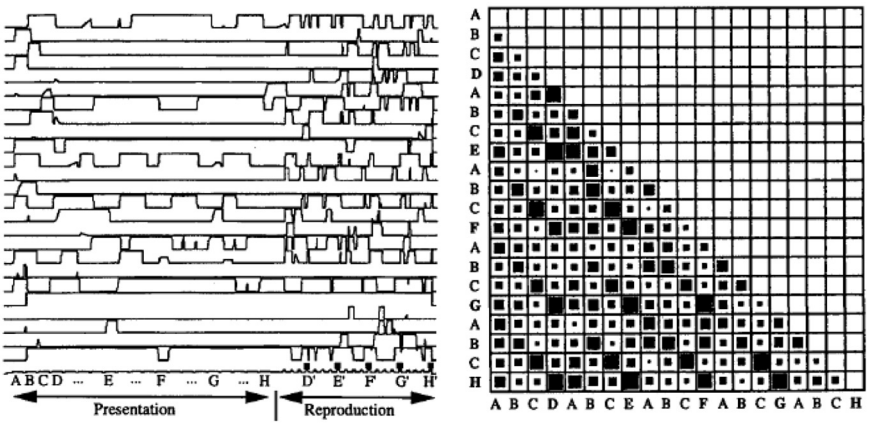



**FIGURE 3 | Neural activity during complex sequence processing.** **Left**—time trace of activity in 25 PFC neurons during presentation and subsequent replay of a complex sequence of order 4. **Right**—vector cosines of PFC state vector activity during the response choice for each of the 25 responses in the output sequence execution. Cosine represented spatially with 0 as empty and 1 as fully filled case. Note that the cases (and subsequent diagonals in the matrix) corresponding to choices of D, E, F, G, and H have relatively high cosines, indicating that the PFC states are similar, but not identical, for these elements. From Dominey (1995).

pattern 123213 (Dominey et al., 1998). In order to account for learning such abstract structure, a system would need additional processing mechanisms in order to detect that the current element in a sequence is a repetition of an earlier element, and then to "recode" the sequence in terms of this pattern of repeating elements (Dominey et al., 1998).

We introduced these modifications (Dominey et al., 1998), and the resulting hybrid model, illustrated in **Figure 4**, could thus learn serial, temporal, and abstract structure of perceptual-motor sequences. To demonstrate the importance of this system in helping to characterize the human initial state in language learning, we chose three landmark papers that defined infants' abilities to implicitly learn the serial (Saffran et al., 1996), temporal (Nazzi et al., 1998), and abstract structure (Marcus et al., 1999) of sound sequences. Saffran et al. demonstrated that in minutes infants could learn the sequential structure of syllable sequences, and detect new sequences of the same syllables that violated the learned structure (Saffran et al., 1996). Nazzi et al. similarly demonstrated that infants are sensitive to the rhythmic structure (stress-timed, syllable-timed, and mora-timed) of language stimuli, and can learn to discriminate between distinct classes in minutes (Nazzi et al., 1998). Finally, Marcus et al. demonstrated that infants can just as quickly learn to discriminate abstract structures of syllable sequences like ABA vs. ABB, where A and B represent variables that can be filled in by new syllables (Marcus et al., 1999). That is, the children could recognize a totally new sequence with syllables that they had never heard (i.e., from a new domain) as fitting with the learned rule ABA or ABB. This was an important finding as it indicates infants can generalize over variables in these sequences. These authors argued that the innate ability to discriminate serial, temporal and abstract structure could contribute to the initial state in language learning.

In a series of simulation studies, we replicated these human demonstrations of learning serial (Saffran et al., 1996), temporal (Nazzi et al., 1998), and abstract structure (Marcus et al., 1999) of sound sequences in the hybrid model. Serial and temporal structure were learned by the simpler temporal recurrent network (TRN), and the abstract structure was learned by the abstract recurrent network (ARN) which required a working memory and recognition capability to detect and represent the repetitive structure of the abstract sequences (Dominey and Ramus, 2000). This was an important step in the developing argument about the possible neural mechanisms of language learning.

Subsequent research has suggested that children in the Saffran task may have been picking up on unintended cues related to chunk strength (Perruchet and Pacton, 2006). The TRN has the property that previous inputs influence the state of the recurrent network and thus influence how subsequent input will be processed. Any kind of sequential structure that can be expressed in these terms should lead to learning effects in the TRN. Similarly, Marcus et al.'s (1999) claim that SRN-like models cannot account for their abstract sequence learning results has been challenged. In the Dienes SRN-based model (Dienes et al., 1999), an additional layer was added to allow the mapping of the new domain onto the learned domain, and multiple presentations of the novel stimuli (for adaptation) are required. Likewise, Chang (2002) demonstrated that the standard SRN fails to generalize on an identity construction (related to the ABA construction of Marcus), while his dual path model successfully generalizes. From this perspective it appears that without additional task specific adaptations, Marcus's claim remains intact.

## FROM SEQUENCE LEARNING TO LANGUAGE—GRAMMATICAL CONSTRUCTION LEARNING

The notion of abstract rule-based structure suggested a possible link to language processing. In order to test the model in a language processing task, we identified a task that had been developed by Caplan et al. (1985) in which aphasic subjects listened to sentences and then had to indicate the corresponding meaning by pointing to images depicting the agent, object and recipient

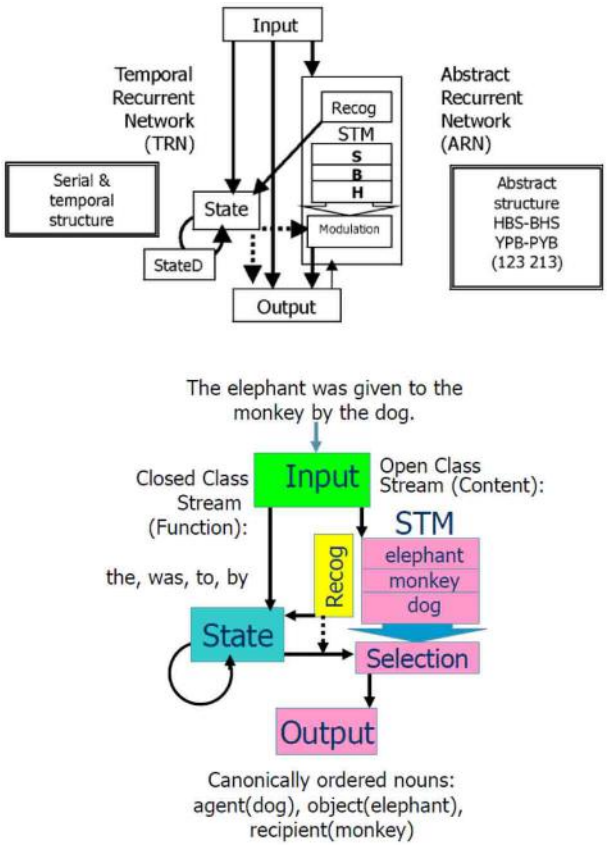



**FIGURE 4 | Combined Abstract Temporal Recurrent Network (ATRN) model. Above:** The temporal recurrent network (TRN) exploits recurrent network dynamics in the recurrently connected neurons in State and StateD layers, to encode serial and temporal structure. To encode the abstract structure common to "isomorphic" sequences such as HBSBHS and YPBPYB, the abstract recurrent network (ARN) stores the N previous elements of the current sequence in a short term memory (STM). The Recognition function compares the current sequence element response generated in Output to the previous elements coded in STM to detect the abstract repetitive structure. This abstracted coding is represented in the recurrent State. In the learned expression of abstract structure knowledge, the contents of STM are selectively provided to the output stream by the activation of Modulation neurons by State neurons. From Dominey et al. (2003). **Below:** Illustration of how the model accommodates language input.

(i.e., who, gave what to whom), always in that "canonical" order. Thus, in the formal task description the "input sequence" is the sequence of words in the sentence, and the "output sequence" is the sequence agent, object, and recipient, corresponding to the meaning in terms of thematic role assignment. The only cues available for determining "who did what to whom" were the word order and grammatical marking, so this is considered a task of syntactic comprehension.

This approach is consistent with the cue competition model of language (Li and Macwhinney, 2013), which holds that across languages, a limited set of cues including the configuration of grammatical function words (closed class morphology in general), word order and prosody are used to encode the grammatical structure that allows thematic role assignment to take place. We thus implement the cue competition hypothesis (Bates et al., 1982, 1991) focusing on word order and grammatical morphology. In our modeling, the notion is that the sequence of closed class words forms a pattern of activity within the recurrent network, and that this pattern can be associated with the corresponding thematic role specification.

In performing the Caplan task, when faced with an example sentence: "The elephant was given to the monkey by the rabbit," after hearing this sentence, the experimental subject was required to indicate the meaning by pointing to images depicting the rabbit, elephant, and monkey (corresponding to agent, object, recipient) in that order. Thus, the Caplan task identifies an excellent challenge for language modeling: Given an input sentence, generate a standardized representation of the meaning (i.e., identify the agent, object, and recipient, always in that "canonical" order). The question now is—how can we reformulate this task so as to be processed by our abstract sequencing model. The general notion is that sentence type should correspond to abstract structure. So the Caplan task involves learning nine different abstract structures. Considering our example sentence, if we replace the words with symbols then this becomes an abstract sequence task, where the input is of the form: a E b c d a M e a R, and the corresponding output is R E M (for rabbit, elephant monkey), where upper case letters indicate nouns, and lower class elements indicate all other lexical categories.

We imposed a lexical categorization process at the level of the input processing, with open class words going to the ARN and closed class words going to the TRN, illustrated in the lower panel of **Figure 4**. Interestingly across languages, these lexical categories tend to have acoustic and distributional signatures that can be used by infants to perform lexical categorization in a process of prosodic bootstrapping (Morgan and Demuth, 1996). Connectionist models have been shown to be able to learn to distinguish open and closed class words from distributional regularities (e.g., Elman, 1990; Chang, 2002). We observed that for French and English, the TRN could encode differences in the prosodic structure of open vs. closed class words in order to perform the lexical categorization between these word classes (Blanc et al., 2003). This provides a demonstration of self-coherence of language in that the most crucial and basic information (i.e., lexical category) is coded at or near the perceptual level.

We thus performed this conversion of the nine sentence types to these abstract sequences. Following the Caplan protocol, five distinct sentences were generated for each sentence type, by replacing the nouns with new nouns. During training, the input sentence was presented to the model, and then in continuation the correctly ordered nouns were presented (i.e., in the agent, object, recipient order). As illustrated in **Figure 4**, the open class words stored in the STM during the sentence input were then compared with the "response" open class elements. This comparison allowed the sequence of correctly ordered nouns to be "recoded" in terms of their respective matching with the nouns stored in the STM. This recoding became the abstract structure that was learned. That is, for each of the nine sentence types, the model learned the reordering of the nouns from their input order in the sentence, to the output "canonical" order agent, object, recipient.

Thus, after training the model could be exposed to a new sentence (with new nouns) that was legal with respect to the learned sentence forms, and it could correctly process the new sentence (reordering the nouns in the agent, object, and recipient order).

## NEURAL IMPLEMENTATION OF GRAMMATICAL CONSTRUCTIONS

Looking at the model in **Figure 4**, there is nothing "language specific" about it. Indeed, we proposed that this same model can be used to process abstract sequences, and sentences. This lead to the "audacious" proposition of an "equivalence hypothesis" (Dominey et al., 2003; Dominey, 2005) stating that a common neural system would participate in aspects of processing sentences and abstract non-linguistic sequences. We found strong correlational support for this hypothesis, observing that in agrammatic aphasic patients with left peri-sylivan (Broca's region) lesions, there was a significant correlation between performance in the nine sentence-type Caplan task of syntactic comprehension (described above and modeled), and a task of abstract sequence processing (Dominey et al., 2003). In a further test of this hypothesis we determined whether the left anterior negativity (LAN), an ERP component related to morphosyntactic processing that can reliably be elicited around 350–500 ms after grammatical functions words (Brown et al., 1999) could be elicited by the grammatical "function symbols" in our non-linguistic sequences. Subjects processed sequences with the abstract structures ABC*x*BAC and ABC*y*ABC where x and y, respectively indicated the non-canonical (complex) vs. canonical (simple) rule. We observed a LAN effect for the function symbol which signaled the more complex structure mapping (Hoen and Dominey, 2000), consistent with data from sentence processing. The link between abstract structure and grammatical structure was further revealed when we demonstrated that agrammatic aphasics trained on an abstract structure that corresponds to the remapping of a relativized structure to a canonical structure demonstrated post-test improvement in sentence processing that was specific to the relativized sentences (Hoen et al., 2003). Continuing to test the equivalence hypothesis, we subsequently examined brain activity during sentence and abstract sequence processing with fMRI, and revealed a common network including the dorsal pars opercularis territory of Broca's area for sentence and abstract sequence processing, with additional activation of the ventral pars triangularis region of Broca's area only for sentence processing (Hoen et al., 2006). Thus, the fMRI results confirmed the model's prediction that a common brain system would account for the structural remapping processing aspects of sentence and abstract sequence processing.

Interestingly, this neural computational mechanism appeared capable of providing a neurophysiological grounding of the notion of grammatical construction processing. In this framework, language is considered as a structured inventory of mappings between sentence surface structure and meanings, referred to as grammatical constructions (Goldberg, 1995; Tomasello, 2003). If grammatical constructions are mappings from sentence structure to meaning, then the language system must be able to (a) identify the construction type for a given sentence, and (b) use this information to extract the meaning from the sentence, based on the identified construction. Our thematic role assignment model implements this notion of grammatical constructions. Word order and closed class structure are integrated in the recurrent network, satisfying (a) and this integrated representation is associated, by learning, with the appropriate mapping of open class elements onto their roles in the sentence, satisfying (b). This integration of word order and closed class structure corresponds to an implementation of the cue competition model. We demonstrated the robustness of this model, and provided further support for the cue competition model by testing the neural model with three distinct languages—English, French and Japanese, each with a distinct set of relevant cues. In each of these languages, a universal property holds—the mapping of sentence to meaning is fully specified by the pattern of open and closed class words unique to each grammatical construction type. The model was thus able to learn 38 distinct English constructions, 26 Japanese constructions, and nine French constructions based on the Caplan task. The model thus integrated results from human neurophysiology and behavior into a coherent framework, with a cross-linguistic validation.

Consistent with human neurophysiology, a central premise in our modeling is that the pattern of grammatical function words is represented in a recurrent cortical network, and that via plasticity in the corticostriatal synapses, the system can learn specific constructions, including constructions in different languages. **Figure 5** illustrates a mapping of the neural computations onto human brain anatomy.

We (Dominey and Inui, 2009; Dominey et al., 2009) attempted to reconcile the corticostriatal model with mainstream neurophysiological models of language processing (Friederici, 2002; Hagoort, 2005) in more detail. Lexical categorization takes place in the temporal cortex, allowing for distinct processing of grammatical function words and semantic content words. Closed class elements are processed in a recurrent frontal network (Inui et al., 2007) corresponding to BA47. The pattern of closed class words forms a characteristic representation in the recurrent network, which can become associated with the appropriate mapping of open class elements onto their respective thematic roles through corticostriatal plasticity. The resulting activity then implements this mapping via the thalamo-cortical projection to the dorsal-prefrontal area BA44/6. Thus, in the inferior frontal gyrus, we consider a transition from syntactic integration in BA47, word level semantics in BA45, and sentence level integration in BA44/6, which is to a certain extent consistent with a similar gradient of processing in the model proposed by Hagoort (2005). This allocation of brain functions to the neuroanatomical regions in **Figure 5** should be considered as tentative, and potentially could be replaced by different allocation of functions. The more solid proposal and contribution of this work is the demonstration that a recurrent cortical network, likely in Broca's region, can integrate multiple cues (here word order and closed class structure) consistent with the cue-competition model, and through corticostriatal plasticity this representation can implement grammatical constructions as mappings from sentence structure to meaning, consistent with the emerging role of the corticostriatal system in language processing.

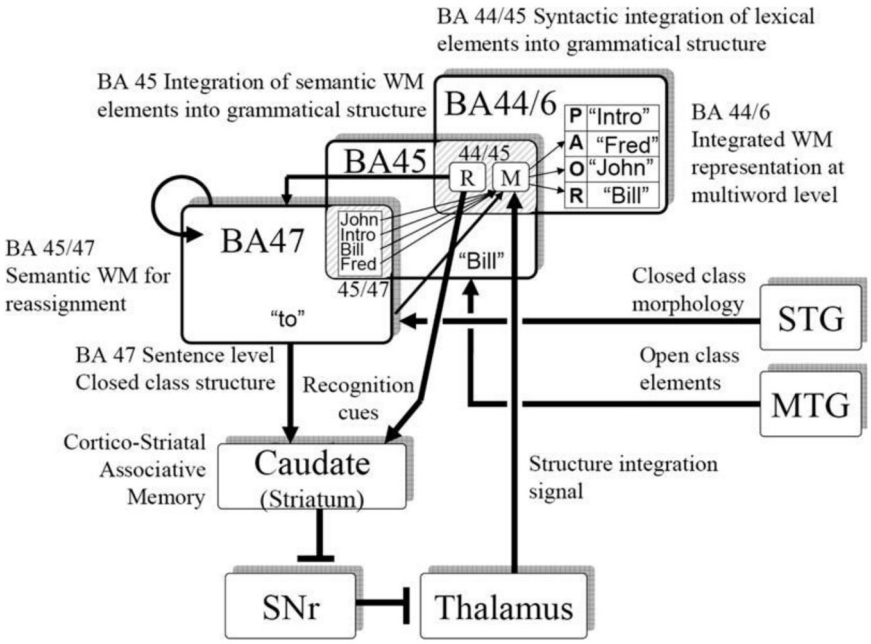



**FIGURE 5 | Neurophysiologically based model of sentence and artificial grammar processing.**

This perspective is consistent with an emerging view of a dorsal-ventral distinction in language processing. While there is indeed significant variability in the details of the functional significance of the dorsal vs. ventral streams in language, there is an emerging consensus that these streams indeed have distinct roles, with the ventral stream related to semantic and conceptual content, and the dorsal stream related to more structural aspects of language (Hickok and Poeppel, 2004; Friederici, 2012; Bornkessel-Schlesewsky and Schlesewsky, 2013). Hickok and Poeppel (2004) thus suggested that the ventral stream would account for the sound-meaning interface, and the dorsal stream would accommodate the auditory-motor interface. In Bornkessel-Schlesewsky and Schlesewsky's model, the ventral stream is more associated with conceptual representations, and the dorsal stream is related to syntactic structuring and the linkage to action. Friederici (2012) proposes a dorsal-ventral model with the ventral stream subserving semantic integration and dorsal stream subserving structural processing. Interaction in ventral circuits linking BA45 and STG/MTG mediates semantic processing, whereas assignment of grammatical relations is mediated by dorsal connections between BA44 and STG/STS. This is consistent with the dorsal-ventral distinction in our model illustrated in **Figure 5**. Indeed, we noted (Dominey and Hoen, 2006) that BA44/46 can be considered to represent the frontal terminus of the dorsal visual pathway, and BA45 the frontal terminus of the ventral pathway (Ungerleider et al., 1998). In this neurophysiological context, Friederici (2012) points out the need to better understand the role of subcortical structures, including the striatum, in language processing. We suggest that corticostriatal plasticity plays a role in implementing the structural mapping processes required for assignment of open-class elements to their appropriate thematic roles. Ullman notes that this is consistent with his declarative-procedural model of language processing, in which the cortico-striatal system contributes to the procedural learning of grammatical rules (Ullman, 2004).

## LARGER CORPORA AND GENERALIZATION IN THE RESERVOIR COMPUTING FRAMEWORK

One of the major limitations with the neural implementation of our model of corticostriatal function is related to the performance of the learning algorithm. A simple form of reward based learning is used to associate states of activity in the recurrent network with neurons in the striatum that correspond to the appropriate thematic role assignment. This requires repetitive training on the corpus with progressive adjustment of learning rates which is prohibitive for the investigation of large corpora. In order to resolve this problem, we apply more robust machine learning methods to our corticostriatal model, in the context of reservoir computing. In reservoir computing, a reservoir of neurons with fixed recurrent connections is stimulated by external inputs, and the desired output is produced by training connections from the excited reservoir units and readout neurons. As noted in Pascanu and Jaeger (2011) this reservoir principle was independently discovered in our own work in cognitive neuroscience with the TRN (Dominey et al., 1995), in computational neuroscience (Maass et al., 2002) with the liquid state machine of Maass, and in machine learning (Jaeger, 2001) with the echo state machine of Jaeger. In the machine learning domain, fast and efficient mechanisms for learning the reservoir-to-readout connections have been developed, and this provides a significant improvement in performance for sentence processing. Using these methods, rather than repeated training with multiple iterations through the

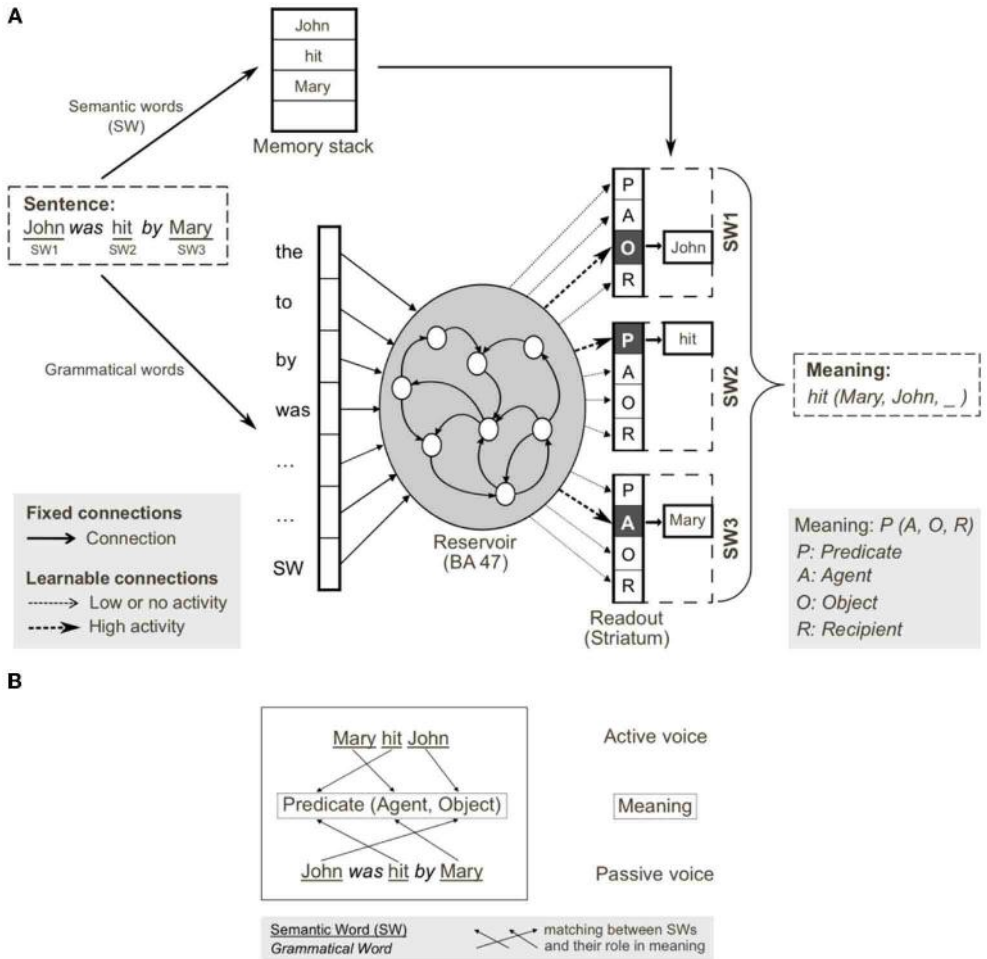



**FIGURE 6 | Reservoir computing implementation of the cortico-striatal sentence processing model. (A)** Semantic and grammatical words (i.e., open and closed class words, respectively) are separated on input. Semantic words (SW) are stored in a memory stack. Grammatical words and a single input for all SWs are inputs to the reservoir (analogous to prefrontal cortex). During training, input sentences are presented word-by-word, and readout units (corresponding to striatum) are forced to the corresponding coded meaning (i.e., SW1-Object, SW2-Predicate, SW3-Agent). In testing, readout units code the predicted role(s) of each semantic word, forming the coded meaning. The meaning [i.e., hit(Mary, John, _)] can be reconstructed from the coded meaning, as SWs in memory stack are reassigned to the thematic roles (predicate, agent, object, recipient) identified in the read-outs. **(B)** Active and passive grammatical constructions (i.e., mapping from sentence form to meaning), and their shared meaning. Coded meaning (indicated by the arrows) corresponds to specific mapping from open class words to meaning, which defines the grammatical construction. From Hinaut and Dominey (2013).

corpus, we could present the corpus to the reservoir only once, collect the reservoir activation and then use linear regression to learn the connections between reservoir units and readout units coding the meaning of the sentences.

Using the model in **Figure 6**, we provided input sentences one word at time, with grammatical words feeding into the recurrent reservoir. Starting at the outset of the sentence presentation, the corresponding readout neurons that coded the correct role for each semantic word, were activated. The model was trained to generate these activations starting at the outset of the sentence, thus providing for a potential predictive capability. This training protocol corresponds to the infant seeing and interpreting the scene before hearing the sentence. **Figure 7** illustrates the activation of a set of readout neurons during the presentation of four different sentence types. The individual traces represent activation of illustrative readout neurons coding for the role of the second noun. We observe that from the outset of the sentence presentation, the system predicts that Noun 2 is the object of verb 1[1]. This remains true in three of the four illustrated constructions, with the exception of the passive in panel **(D)**. Note that when the word "was" arrives, the system reconfigures its prediction. Later in these constructions (at the point indicated by the labeled arrow b) note the distinct responses respectively to "to," and "that," and then finally at the point indicated by the labeled arrow c, the responses to the arrival of "Verb" vs. "was." What we observe is that time locked with words that designate the possible

[1]Hinaut and Dominey also perform a more general treatment where verbs are included in the processing of semantic (or open class words) in constructions as illustrated in **Figure 6**.

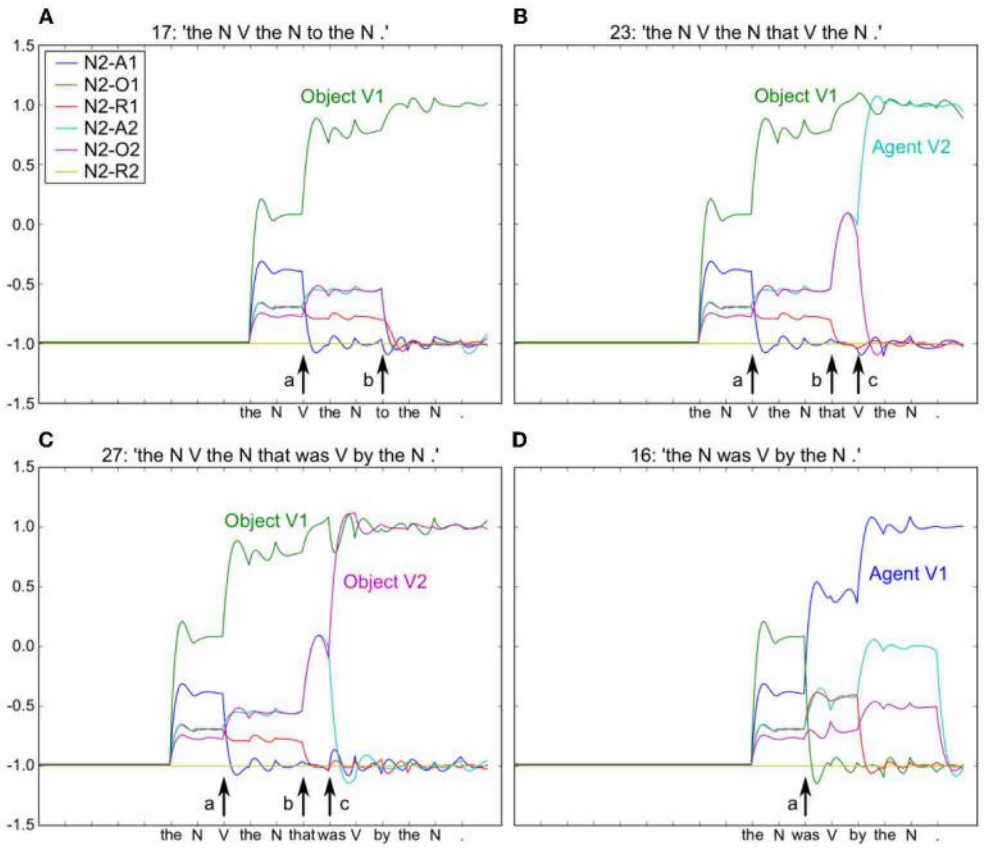



**FIGURE 7 | Neurons coding thematic roles indicated by colored traces (see embedded legend).** For all four sentences [see period before arrow (a)], the model initially predicts that Noun 2 is the Object of Action 1 (green trace). In **(B)** and **(C)** this remains true, but Noun 2 is also the Agent and Object of Action 2 in **(B)** and **(C)** respectively. At point (b), arrival of "to" confirms the correct prediction of N2-O1 (green trace) in **(A)**, and the arrival of "that" induces a change in activity in **(B)** and **(C)**, with increased prediction of both Agent and Object roles for V2, respectively. Note that this is resolved at the arrival of the "V" and "was" in **(B)** and **(C)** respectively [arrow (c)]. In **(D)** the arrival of "was" provokes a new analysis with Noun 2 as the Agent of Action 1. Embedded legend: N2-A1 – Noun 2 is the agent of Action 1. A, Agent; O, Object; R, Recipient.

licensing of a construction, the model neurons react. If we dissect in panels **(B,C)** what happens between the events labeled b and c, we can consider that the system is maintaining parallel parses, and the decision between these parses is determined when the appropriate disambiguating words arrive at point c. This graphically illustrates the ranked parallel parses. That is, each of the neurons in these panels corresponds to a possible role for Noun 2. Activation of these neurons corresponds to the choice, and the level of activation corresponds to the rank. Thus, multiple parses can be entertained in parallel. In panels **(B,C)**, between marked locations b and c, two possible parses are equally active, and at the arrival of the next word at c, the choice is made.

The changes in neural activation as observed at point *c* can be interpreted in the context of human brain activity, revealed by event related potentials (ERPs) recorded during sentence processing. We can consider that the summed relative changes in activity of the model neurons represent a form of ERP signal. In this context, a larger ERP response was observed for subject-object vs. subject subject relative sentences time locked with the disambiguating word in the sentence (Hinaut and Dominey, 2013), similar to the effect observed in human subjects (Friederici et al., 2001). In our corpus, similar to human language (Roland et al., 2007), constructions with subject-object structure are less frequent than subject-subject, and canonical types where the head noun is the agent. Thus, this change in neural activity is in a sense due to a form of expectation violation, based on the corpus statistics. MacDonald and Christiansen (2002) have provided detailed simulation evidence for such phenomena involving an interaction between complexity, frequency, and experience. They demonstrated that with an equal distribution of subject- and object-relatives, their recurrent network gave superior performance on the subject relatives due to the networks' abilities to generalize to rare structures as a function of experience with similar, more common simple sentences.

The performance of the model, as revealed by these readout activation profiles can potentially be linked to reading times, such that the time required for a neuron to reach a threshold could be plausibly interpreted as a reading time.

The model thus provides an implementation of a form of ranked parallel processing model, where the parallel maintenance of possible parses is an inherent aspect of the model (Gibson and Pearlmutter, 2000; Lewis, 2000). This behavior is a reflection of

the statistical structure of the training corpus. In effect, the activity of the readout neurons reflects the probability of their being activated in the training corpus.

Indeed, the behavior of the trained system is clearly influenced by the nature of the grammatical structure inherent in the training corpus. Working in the machine learning context of reservoir computing allowed us to perform experiments with corpora up to $9 \times 10^4$ different constructions. The advantage of performing these large corpora experiments is that it allows for a systematic analysis of the influence of the training corpus on the ability to generalize. Here we speak of compositional generalization, where the system is actually able to handle new constructions that were not used in the training corpus (as opposed to using learned constructions with new open class words). We performed a series of experiments with a small corpus of 45 constructions in which we examined very specific timing effects of the parallel processing, and two larger corpora of 462 and 90,582 distinct constructions, respectively. Training with the larger corpora revealed quite promising generalization in cross-validation studies, where different proportions of a corpus are removed from training, and then used in testing to evaluate generalization. We observed generalization of up to 90% for the 462 corpus, and over 95% in the 90 K corpus. Interestingly, when we scrambled the 462 corpus, generalization was reduced to 30%, indicating that the system learned the underlying grammatical structure encoded in the training corpus. Most interestingly, this generalization in the largest corpus could be achieved with exposure to only 12.5% of the corpus. Thus we see that the grammatical structure of language can be learned and generalized by such recurrent networks. The clear distinction in this work is that the learning is revealed by extracting thematic roles, rather than predicting the next word or lexical category (Tong et al., 2007). Indeed, the power of such recurrent models in now employed in concrete natural language processing applications of semantic role labeling (Barnickel et al., 2009).

## DISCUSSION

The study of recurrent networks for language processing has a rich history. A vast and productive line of research has characterized language processing in terms of predicting the next word in a sentence, initiated by the ground-breaking work of Elman (1990, 1991, 1993), and fruitfully pursued by others (e.g., Christiansen and Chater, 1999). Characterizing language in terms of thematic role assignment also has a rich history in connectionist modeling (e.g., McClelland et al., 1989; Miikkulainen, 1996). Miikkulainen provides an extensive review of this literature. His novel contribution is a modular distributed architecture based on the separation of parsing (an SRN), segmentation (a RAAM model), and a stack (for handling depth recursion). Communication between the modules includes the transfer of activation patterns, and control. The resulting system can learn case role mapping, as well as phrasal embedding structure, and then generalize these to sentences with new relative phrase embedding structure.

The research reviewed here presents a coherent framework in which a recurrent network encodes grammatical structure in the input, and modifiable connections from the recurrent network learn the mappings from that grammatical structure to the corresponding meaning representations for large corpora of grammatical constructions. With sufficiently large corpora the system displays a significant capability to generalize to new constructions, exploiting the regularities that define grammatical well formedness (Hinaut and Dominey, 2013). This argues that a system can display the ability to learn the grammatical structure implicit in a corpus without explicitly representing the grammar (Elman, 1991), and that it can generalize to accommodate new constructions that are consistent with that grammar (Voegtlin and Dominey, 2005). Part of the novelty in the position suggested here is that this recurrent network and readout system is implemented in the primate brain in the cortico-striatal system.

The computational properties of such recurrent systems is remarkable. Our initial work with recurrent networks with fixed connections and modifiable readouts demonstrated significant sequence learning capabilities (Dominey, 1995, 1998a,b), and accounted for neural coding of sequential structure in the PFC (Dominey et al., 1995; Dominey and Boussaoud, 1997). Subsequent work with such systems demonstrated their vast computational power (Maass et al., 2002; Jaeger and Haas, 2004). Projecting inputs into such reservoirs allows a mapping into a high dimensional space. This provides a dynamic compositionality that can represent an arbitrary class of non-linear functions. Recent studies of primate electrophysiology provide evidence that indeed, the PFC operates based on reservoir-like properties (Rigotti et al., 2010, 2013). The key point—the use of random connection weights in a structured network—is echoed in the principal property of cortex—the high predominance of local recurrent connectivity (Douglas et al., 1995; Binzegger et al., 2009), particularly in PFC (Goldman-Rakic, 1987). The use of fixed recurrent connections within these reservoirs means eliminates the need to artificially constrain the processing of time and temporal structure in these networks, thus allowing a realistic processing of temporal structure that is much more difficult in networks with learning in the recurrent connections. Of course, there is plasticity in the cortex, but simplifying this with fixed connections, the dynamic compositionality of reservoir computing yields significant processing capabilities. The reservoir framework is thus highly appropriate for the study of complex cognitive function including establishing structure-meaning relations in language, and it has already been successfully employed in the context of predicting the next word in the context of language processing (Tong et al., 2007).

It should be noted that dynamic does not correspond to "out of control." That is, while a recurrent network will evolve in a dynamic pattern of activity, this dynamic activity can be associated with a stable representation of the meaning. In the human, this dynamic activity is observed in EEG responses (e.g., the ELAN, LAN, N400, P600 cascade of responses, modulated by lexical category and sentence context) that are dynamic in time, yet reflect the coherent processing of the successive words in sentences.

We have postulated that recurrent cortical networks provide the mechanism for representing grammatical structure, and that plastic corticostriatal connections participate in learning this structure in the acquisition of a language (Dominey and Inui, 2009; Dominey et al., 2009). We thus take a strong stance on the role of the human corticostriatal system in language processing. This would predict that patients with dysfunction in the corticostriatal system should have deficits in syntactic

processing, and should show neurophysiological anomalies during language processing. Significant data from a number of sources are consistent with this stance. Several studies from Friederici and Kotz (2003; Friederici et al., 2003; Frisch et al., 2003; Kotz et al., 2003) in patients with striatal dysfunction due to lesion or Parkinson's disease demonstrate that the P600 ERP evoked by syntactic anomalies or complexity in normal controls subjects is reduced or eliminated in these patients, while other language related responses (including the early left anterior negativity or ELAN and N400) remain intact. Similarly, these patients are impaired in the processing of grammatical complexity (Hochstadt et al., 2006; Hochstadt, 2009). This suggests that the intact corticostriatal system is required in generating this normal brain response to grammatical complexity processing. Ullman argues that the corticostriatal system implements procedural rules for word level grammatical processing (Ullman, 2001). We take this suggestion even further, arguing that the corticostriatal system participates in the implementation of grammatical constructions at the sentence level, in the mapping of the structure of the surface form of the construction to the meaning representation (Dominey and Inui, 2009; Dominey et al., 2009).

It is now relatively accepted that meaning is encoded in distributed embodied representations that have an analogical component that is not symbolic (Bergen and Chang, 2005; Barsalou, 2009; Lallée et al., 2010b; Madden et al., 2010). Interestingly, such representations are difficult to manipulate, when compared with symbolic representations. In this context, there is an emerging perspective that the complete language system likely involves both symbolic and distributed-embodied representations (Bergen and Chang, 2005; Barsalou, 2009; Lallée et al., 2010b; Madden et al., 2010).

This poses the question of how the link is made between language and embodied simulations. A promising area where these issues can be investigated is in the development of cognitive systems for robots. This link between language and meaning in cognitive science is not new. At the height of the cognitive revolution, Feldman and colleagues proposed the problem of miniature language acquisition as a "touchstone" for cognitive science (Feldman et al., 1990). A machine should be trained on a set of <sentence, picture> pairs, and then in testing should be able to say whether a given novel sentence correctly described a novel picture. In order to address this we modified the problem such that the training data were <sentence, video-scene> pairs. Based on the notion that infants can parse visual scenes by detecting sequences of perceptual primitives [inspired by Mandler (1999)] we developed an event parser that could detect actions including take, take-from, give, push and touch (Dominey, 2003b). Naïve subjects performed actions that were parsed by this system, and simultaneously narrated their actions, thus providing a set of <sentence, meaning> data on which to train the neural network grammatical construction model (Dominey and Boucher, 2005). The model learned a set of simple constructions and could generalize to new <sentence, meaning> pairs. We subsequently demonstrated how the system could learn to recognize such actions (Lallée et al., 2010a), similar to Siskind (2001; Fern et al., 2002). Such language—action mappings are becoming increasingly powerful in the domain of human—robot cooperation (Petit et al., 2013). What we will find, is that as the cognitive systems of robots become increasingly sophisticated, they will naturally afford richer language. For example, as mental simulation capabilities develop, the need for verb aspect to control the flow of time in these simulations will naturally arise (Madden et al., 2010).

Arguments on the learnability of language have held that because the compositional generative complexity of language is so vast, and the input to the child so impoverished, the underlying language learning capability must rely on a form of pre-specified universal grammar so that language learning consists in setting the parameters of this system (Chomsky, 1995). Usage-based approaches to acquisition argue, in contrast, that the input is actually guided by joint attention mechanisms and specialized mechanisms for human socialization which focus the learners attention on the intended meaning (Tomasello, 2000, 2003; Dominey and Dodane, 2004). This suggests that language acquisition should be characterized not formally as a problem of grammar induction, but rather socially, as a problem of expressing and extracting meaning. This perspective emphasizes the potential contribution that the structure of meaning can contribute to the learning process.

In this context, Chang (2002) has demonstrated that under equivalent conditions, providing a language processing model with a message that contained semantic content provided additional structuring information, and increased the learning performance. It is likely that this contributes to generalization. We have demonstrated that with corpora of moderate size (between 450 and 90,000 constructions) the recurrent network model demonstrates quite robust generalization (Hinaut and Dominey, 2013). We believe that this is because the structural regularities that allow the system to generalize are inherent within the training data. Interestingly, the training data include both the surface forms of the constructions, and the corresponding meaning structure. This suggests that part of what allows the system to generalize is this additional source of learnable structural regularities—not only those present in the surface structure, but also those present in the mapping of that structure to the meaning structure. Thus the meaning structure can contribute to learnability and generalization (Dominey, 2000; Chang, 2002). In response to Dominey's commentary (Dominey, 2003a) on the précis of "Foundations of Language," Jackendoff (2003) states "In the parallel architecture it is natural to suppose that the hierarchical complexity of syntax is but a pale reflection of that in meaning, and it exists only insofar as it helps express thought more precisely. Moreover, Dominey says, access to the compositionality of meaning provides a scaffolding for the child's discovery of syntactic structure. I concur." Thus, in the study and modeling of language acquisition, significant work remains in characterizing the structure of the conceptual system.

## ACKNOWLEDGMENTS

Supported the European FP7 Cognitive Systems and Robotics project EFAA Grant 270490. I thank Mike Tomasello for helping identify the link between our structure mapping and the work on grammatical constructions during a visit to MPG EVA in Leipzig in 2003. I thank Juan Segui for comments on the audacious equivalence hypothesis.

## REFERENCES


Alexander, G. E., DeLong, M. R., and Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci.* 9, 357–381. doi: 10.1146/annurev.ne.09.030186.002041

Barnickel, T., Weston, J., Collobert, R., Mewes, H.-W., and Stümpflen, V. (2009). Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS ONE* 4:e6393. doi: 10.1371/journal.pone.0006393

Barone, P., and Joseph, J. P. (1989). Prefrontal cortex and spatial sequencing in macaque monkey. *Exp. Brain Res.* 78, 447–464. doi: 10.1007/BF00230234

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1281–1289. doi: 10.1098/rstb.2008.0319

Bates, E., McNew, S., MacWhinney, B., Devescovi, A., and Smith, S. (1982). Functional constraints on sentence processing: a cross-linguistic study. *Cognition* 11, 245–299. doi: 10.1016/0010-0277(82)90017-8

Bates, E., Wulfeck, B., and MacWhinney, B. (1991). Cross-linguistic research in aphasia: an overview. *Brain Lang.* 41, 123–148. doi: 10.1016/0093-934X(91)90149-U

Bergen, B., and Chang, N. (2005). "Embodied construction grammar in simulation-based language understanding," in *Construction Grammars: Cognitive Grounding and Theoretical Extensions*, eds J.-O. Östman and M. Fried (Amsterdam: John Benjamins B.V), 147–190.

Binzegger, T., Douglas, R., and Martin, K. (2009). Special issue: topology and dynamics of the canonical circuit of cat V1. *Neural Netw.* 22, 1071–1078. doi: 10.1016/j.neunet.2009.07.011

Blanc, J.-M., Dodane, C., and Dominey, P. F. (2003). "Temporal processing for syntax acquisition: a simulation study," in *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (Cambridge, MA: MIT Press).

Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2013). Reconciling time, space and function: a new dorsal–ventral stream model of sentence comprehension. *Brain Lang.* 125, 60–76. doi: 10.1016/j.bandl.2013.01.010

Brown, C. M., Hagoort, P., and ter Keurs, M. (1999). Electrophysiological signatures of visual lexical processing: open- and closed-class words. *J. Cogn. Neurosci.* 11, 261–281. doi: 10.1162/089892999563382

Calabresi, P., Maj, R., Mercuri, N. B., and Bernardi, G. (1992). Coactivation of D1 and D2 dopamine receptors is required for long-term synaptic depression in the striatum. *Neurosci. Lett.* 142, 95–99. doi: 10.1016/0304-3940(92)90628-K

Caplan, D., Baker, C., and Dehaut, F. (1985). Syntactic determinants of sentence comprehension in aphasia. *Cognition* 21, 117–175. doi: 10.1016/0010-0277(85)90048-4

Chang, F. (2002). Symbolically speaking: a connectionist model of sentence production. *Cogn. Sci.* 26, 609–651. doi: 10.1016/S0364-0213(02)00079-4

Chomsky, N. (1995). *The Minimalist Program.* Cambridge, MA: MIT Press.

Christiansen, M. H., and Chater, N. (1999). Connectionist natural language processing: the state of the art. *Cogn. Sci.* 23, 417–437. doi: 10.1207/s15516709cog2304_2

Dienes, Z., Altmann, G., and Gao, S.-J. (1999). Mapping across domains without feedback: a neural network model of transfer of implicit knowledge. *Cogn. Sci.* 23, 53–82. doi: 10.1207/s15516709cog2301_3

Dominey, P., and Boucher, J. (2005). Learning to talk about events from narrated video in a construction grammar framework. *Artif. Intell.* 167, 31–61. doi: 10.1016/j.artint.2005.06.007

Dominey, P. F. (1995). Complex sensory-motor sequence learning based on recurrent state representation and reinforcement learning. *Biol. Cybern.* 73, 265–274. doi: 10.1007/BF00201428

Dominey, P. F. (1998a). Influences of temporal organization on sequence learning, and transfer: comments on Stadler, (1995) and Curran and Keele (1993). *J. Exper. Psychol. Learn. Mem. Cogn.* 24, 14.

Dominey, P. F. (1998b). A shared system for learning serial and temporal structure of sensori-motor sequences. evidence from simulation and human experiments. *Brain Res. Cogn. Brain Res.* 6, 163–172.

Dominey, P. F. (2000). Conceptual grounding in simulation studies of language acquisition. *Evol. Commun.* 4, 57–85. doi: 10.1075/eoc.4.1.05dom

Dominey, P. F. (2003a). A conceptuocentric shift in the characterization of language: comment on Jackendoff. *Behav. Brain Sci.* 26, 674–674.

Dominey, P. F. (2003b). "Learning grammatical constructions in a miniature language from narrated video events," in *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (Boston, MA).

Dominey, P. F. (2005). From sensorimotor sequence to grammatical construction: evidence from simulation and neurophysiology. *Adapt. Behav.* 13, 347–361. doi: 10.1177/105971230501300401

Dominey, P. F., and Arbib, M. A. (1992). A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cereb. Cortex* 2, 153–175. doi: 10.1093/cercor/2.2.153

Dominey, P. F., Arbib, M. A., and Joseph, J. P. (1995). A model of corticostriatal plasticity for learning oculomotor associations and sequences *J. Cogn. Neurosci.* 7, 25.

Dominey, P. F., and Boussaoud, D. (1997). Encoding behavioral context in recurrent networks of the fronto-striatal system: a simulation study. *Brain Res. Cogn. Brain Res.* 6, 53–65. doi: 10.1016/S0926-6410(97)00015-3

Dominey, P. F., and Dodane, C. (2004). Indeterminacy in language acquisition: the role of child directed speech and joint attention. *J. Neurolinguistics* 17, 121–145. doi: 10.1016/S0911-6044(03)00056-3

Dominey, P. F., and Hoen, M. (2006). Structure mapping and semantic integration in a construction-based neurolinguistic model of sentence processing. *Cortex* 42, 476–479. doi: 10.1016/S0010-9452(08)70381-2

Dominey, P. F., Hoen, M., Blanc, J. M., and Lelekov-Boissard, T. (2003). Neurological basis of language and sequential cognition: evidence from simulation, aphasia, and ERP studies. *Brain Lang.* 86, 207–225. doi: 10.1016/S0093-934X(02)00529-1

Dominey, P. F., and Inui, T. (2009). Cortico-striatal function in sentence comprehension: insights from neurophysiology and modeling. *Cortex* 45, 1012–1018. doi: 10.1016/j.cortex.2009.03.007

Dominey, P. F., Inui, T., and Hoen, M. (2009). Neural network processing of natural language: II. Towards a unified model of corticostriatal function in learning sentence comprehension and non-linguistic sequencing. *Brain Lang.* 109, 80–92. doi: 10.1016/j.bandl.2008.08.002

Dominey, P. F., Lelekov, T., Ventre-Dominey, J., and Jeannerod, M. (1998). Dissociable processes for learning the surface structure and abstract structure of sensorimotor sequences. *J. Cogn. Neurosci.* 10, 734–751. doi: 10.1162/089892998563130

Dominey, P. F., and Ramus, F. (2000). Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant. *Lang. Cogn. Process.* 15, 40.

Douglas, R. J., Koch, C., Mahowald, M., Martin, K., and Suarez, H. H. (1995). Recurrent excitation in neocortical circuits. *Science* 269, 981–985. doi: 10.1126/science.7638624

Elman, J. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* 7, 30. doi: 10.1007/BF00114844

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition* 48, 71–99. doi: 10.1016/0010-0277(93)90058-4

Feldman, J. A., Lakoff, G., Stolcke, A., and Weber, S. H. (1990). "Miniature language acquisition: a touchstone for cognitive science," in *Proceedings of the 12th Annual Conference of the Cognitive Science Society* (Cambridge, MA: MIT Press), 686–693.

Fern, A., Givan, R., and Siskind, J. M. (2002). Specific-to-general learning for temporal events with application to learning event definitions from video. *Artif. Intell. Res.* 17, 379–449.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci.* 6, 78–84. doi: 10.1016/S1364-6613(00)01839-8

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends Cogn. Sci.* 16, 262–268. doi: 10.1016/j.tics.2012.04.001

Friederici, A. D., and Kotz, S. A. (2003). The brain basis of syntactic processes: functional imaging and lesion studies. *Neuroimage* 20(Suppl. 1), S8–S17. doi: 10.1016/j.neuroimage.2003.09.003

Friederici, A. D., Kotz, S. A., Werheid, K., Hein, G., and von Cramon, D. Y. (2003). Syntactic comprehension in Parkinson's disease:

investigating early automatic and late integrational processes using event-related brain potentials. *Neuropsychology* 17, 133–142. doi: 10.1037/0894-4105.17.1.133

Friederici, A. D., Mecklinger, A., Spencer, K. M., Steinhauer, K., and Donchin, E. (2001). Syntactic parsing preferences and their on-line revisions: a spatio-temporal analysis of event-related brain potentials. *Brain Res. Cogn. Brain Res.* 11, 305–323. doi: 10.1016/S0926-6410(00)00065-3

Frisch, S., Kotz, S. A., von Cramon, D. Y., and Friederici, A. D. (2003). Why the P600 is not just a P300: the role of the basal ganglia. *Clin. Neurophysiol.* 114, 336–340. doi: 10.1016/S1388-2457(02)00366-8

Gibson, E., and Pearlmutter, N. J. (2000). Distinguishing serial and parallel parsing. *J. Psycholinguist Res.* 29, 231–240. doi: 10.1023/A:1005153330168

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure.* Chicago, IL: University of Chicago Press.

Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. *Handb. Neurophysiol.* 5, 40.

Grahn, J. A., Parkinson, J. A., and Owen, A. M. (2009). The role of the basal ganglia in learning and memory: neuropsychological studies. *Behav. Brain Res.* 199, 53–60. doi: 10.1016/j.bbr.2008.11.020

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* 9, 416–423. doi: 10.1016/j.tics.2005.07.004

Hickok, G., and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92, 67–99. doi: 10.1016/j.cognition.2003.10.011

Hinaut, X., and Dominey, P. F. (2013). Real-time parallel processing of grammatical structure in the fronto-striatal system: a recurrent network simulation study using reservoir computing. *PLoS ONE* 8:e52946. doi: 10.1371/journal.pone.0052946

Hochstadt, J. (2009). Set-shifting and the on-line processing of relative clauses in Parkinson's disease: results from a novel eye-tracking method. *Cortex* 45, 991–1011. doi: 10.1016/j.cortex.2009.03.010

Hochstadt, J., Nakano, H., Lieberman, P., and Friedman, J. (2006). The roles of sequencing and verbal working memory in sentence comprehension deficits in Parkinson's disease. *Brain Lang.* 97, 243–257. doi: 10.1016/j.bandl.2005.10.011

Hoen, M., and Dominey, P. F. (2000). ERP analysis of cognitive sequencing: a left anterior negativity related to structural transformation processing. *Neuroreport* 11, 3187–3191. doi: 10.1097/00001756-200009280-00028

Hoen, M., Golembiowski, M., Guyot, E., Deprez, V., Caplan, D., and Dominey, P. F. (2003). Training with cognitive sequences improves syntactic comprehension in agrammatic aphasics. *Neuroreport* 14, 495–499. doi: 10.1097/00001756-200303030-00040

Hoen, M., Pachot-Clouard, M., Segebarth, C., and Dominey, P. F. (2006). When Broca experiences the Janus syndrome: an ER-fMRI study comparing sentence comprehension and cognitive sequence processing. *Cortex* 42, 605–623. doi: 10.1016/S0010-9452(08)70398-8

Ilinsky, I., Jouandet, M., and Goldman−Rakic, P. (1985). Organization of the nigrothalamocortical system in the rhesus monkey. *J. Comp. Neurol.* 236, 315–330. doi: 10.1002/cne.902360304

Inui, T., Ogawa, K., and Ohba, M. (2007). Role of left inferior frontal gyrus in the processing of particles in Japanese. *Neuroreport* 18, 431–434. doi: 10.1097/WNR.0b013e32805dfb7e

Jackendoff, R. (2003). Precis of foundations of language: brain, meaning, grammar, evolution. *Behav. Brain Sci.* 26, 651–65. discussion: 66–707. doi: 10.1017/S0140525X03000153

Jaeger, H. (2001). The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148.

Jaeger, H., and Haas, H. (2004). Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science* 304, 78–80. doi: 10.1126/science.1091277

Jaeger, H., Lukosevicius, M., Popovici, D., and Siewert, U. (2007). Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Netw.* 20, 335–352. doi: 10.1016/j.neunet.2007.04.016

Kotz, S. A., Frisch, S., von Cramon, D. Y., and Friederici, A. D. (2003). Syntactic language processing: ERP lesion data on the role of the basal ganglia. *J. Int. Neuropsychol. Soc.* 9, 1053–1060. doi: 10.1017/S1355617703970093

Lallée, S., Lemaignan, S., Lenz, A., Melhuish, C., Natale, L., Skachek, S., et al. (2010a). "Towards a platform-independent cooperative human-robot interaction system: I. Perception," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference* (Taipei), 4444–4451. doi: 10.1109/IROS.2010.5652697

Lallée, S., Madden, C., Hoen, M., and Dominey, P. (2010b). Linking language with embodied teleological representations of action for humanoid cognition. *Front. Neurobot.* 4:8. doi: 10.3389/fnbot.2010.00008

Lehericy, S., Ducros, M., Van de Moortele, P. F., Francois, C., Thivard, L., Poupon, C., et al. (2004). Diffusion tensor fiber tracking shows distinct corticostriatal circuits in humans. *Ann. Neurol.* 55, 522–529. doi: 10.1002/ana.20030

Lewis, R. L. (2000). Falsifying serial and parallel parsing models: empirical conundrums and an overlooked paradigm. *J. Psycholinguist. Res.* 29, 241–248. doi: 10.1023/A:1005105414238

Li, P., and Macwhinney, B. (2013). *Competition Model. The Encyclopedia of Applied Linguistics.* Hoboken, NJ: Blackwell Publishing Ltd.

Lukosevicius, M., and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Comput. Sci. Rev.* 3, 22.

Maass, W., Natschlager, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.* 14, 2531–2560. doi: 10.1162/089976602760407955

MacDonald, M. C., and Christiansen, M. H. (2002). Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychol. Rev.* 109, 35–54. doi: 10.1037/0033-295X.109.1.35

Madden, C., Hoen, M., and Dominey, P. F. (2010). A cognitive neuroscience perspective on embodied language for human-robot cooperation. *Brain Lang.* 112, 180–188. doi: 10.1016/j.bandl.2009.07.001

Mandler, J. (1999). "Preverbal Representations and Language," in *Language and Space*, eds P. Bloom, M. A. Peterson, L. Nadel, and M. F. Garrett (Cambridge, MA: MIT Press), 365–384.

Marcus, G. F., Vijayan, S., Bandi Rao, S., and Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science* 283, 77–80. doi: 10.1126/science.283.5398.77

McClelland, J. L., St. John, M., and Taraban, R. (1989). Sentence comprehension: a parallel distributed processing approach. *Lang. Cogn. Process.* 4, SI287–SI335.

Miikkulainen, R. (1996). Subsymbolic case-role analysis of sentences with embedded clauses. *Cogn. Sci.* 20, 47–73. doi: 10.1207/s15516709cog2001_2

Morgan, J. L., and Demuth, K. (1996). *Signal to Syntax: An overview. Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Nazzi, T., Bertoncini, J., and Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 756–766. doi: 10.1037/0096-1523.24.3.756

Pascanu, R., and Jaeger, H. (2011). A neurodynamical model for working memory. *Neural Netw.* 24, 199–207. doi: 10.1016/j.neunet.2010.10.003

Pearlmutter, B. A. (1995). Gradient calculations for dynamic recurrent neural networks: a survey. *IEEE Trans. Neural Netw.* 6, 1212–1228. doi: 10.1109/72.410363

Perruchet, P., and Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends Cogn. Sci.* 10, 233–238. doi: 10.1016/j.tics.2006.03.006

Petit, M., Lallee, S., Boucher, J.-D., Pointeau, G., Cheminade, P., Ognibene, D., et al. (2013). The coordinating role of language in real-time multi-modal learning of cooperative tasks. *IEEE Trans. Auton. Mental Dev.* 5, 3–17. doi: 10.1109/TAMD.2012.2209880

Reading, P. J., Dunnett, S. B., and Robbins, T. W. (1991). Dissociable roles of the ventral, medial and lateral striatum on the acquisition and performance of a complex visual stimulus-response habit. *Behav. Brain Res.* 45, 147–161. doi: 10.1016/S0166-4328(05)80080-4

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., et al. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature* 497, 585–590. doi: 10.1038/nature12160

Rigotti, M., Rubin, D. B. D., Wang, X.-J., and Fusi, S. (2010). Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Front. Comput. Neurosci.* 4:24. doi: 10.3389/fncom.2010.00024

Robbins, T. W., Giardini, V., Jones, G. H., Reading, P., and Sahakian, B. J. (1990). Effects of dopamine depletion from the caudate-putamen and nucleus accumbens septi on the acquisition and performance of a conditional discrimination task. *Behav. Brain Res.* 38, 243–261. doi: 10.1016/0166-4328(90)90179-I

Roland, D., Dick, F., and Elman, J. L. (2007). Frequency of basic english grammatical structures: a corpus analysis. *J. Mem. Lang.* 57, 348–379. doi: 10.1016/j.jml.2007.03.002

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928. doi: 10.1126/science.274.5294.1926

Selemon, L. D., and Goldman-Rakic, P. S. (1985). Longitudinal topography and interdigitation of corticostriatal projections in the rhesus monkey. *J. Neurosci.* 5, 776–794.

Siskind, J. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Intell. Res.* 15, 31–90.

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends Cogn. Sci.* 4, 156–163. doi: 10.1016/S1364-6613(00)01462-5

Tomasello, M. (2003). *Constructing a Language: A Usage Based Approach to Language Acquisition.* Boston, MA: MIT Press.

Tong, M. H., Bicket, A. D., Christiansen, E. M., and Cottrell, G. W. (2007). Learning grammatical structure with Echo State Networks. *Neural Netw.* 20, 9. doi: 10.1016/j.neunet.2007.04.013

Ullman, M. T. (2001). A neurocognitive perspective on language: the declarative/procedural model. *Nat. Rev. Neurosci.* 2, 717–726. doi: 10.1038/35094573

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition* 92, 231–270. doi: 10.1016/j.cognition.2003.10.008

Ungerleider, L. G., Courtney, S. M., and Haxby, J. V. (1998). A neural system for human visual working memory. *Proc. Natl. Acad. Sci. U.S.A.* 95, 883–890. doi: 10.1073/pnas.95.3.883

Voegtlin, T., and Dominey, P. F. (2005). Linear recursive distributed representations. *Neural Netw.* 18, 878–895. doi: 10.1016/j.neunet.2005.01.005

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 25 April 2013; accepted: 16 July 2013; published online: 05 August 2013.*
*Citation: Dominey PF (2013) Recurrent temporal networks and language acquisition—from corticostriatal neurophysiology to reservoir computing. Front. Psychol. **4**:500. doi: 10.3389/fpsyg.2013.00500*
*This article was submitted to Frontiers in Language Sciences, a specialty of Frontiers in Psychology.*

frontiers in
PSYCHOLOGY

OPINION ARTICLE
published: 05 August 2013
doi: 10.3389/fpsyg.2013.00504

# The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects

***Martial Mermillod*[1,2]*, *Aurélia Bugaiska*[3] *and Patrick Bonin*[2,3]**

[1] *Centre National de la Recherche Scientifique, LPNC UMR 5105, Université Grenoble Alpes, Grenoble, France*
[2] *Institut Universitaire de France, Paris, France*
[3] *LEAD-Centre National de la Recherche Scientifique, UMR 5022, University of Bourgogne, Dijon, France*
**Correspondence: martial.mermillod@upmf-grenoble.fr*

***Edited by:***
*Julien Mayor, University of Geneva, Switzerland*

***Reviewed by:***
*Michael Thomas, Birkbeck College, UK*

The stability-plasticity dilemma is a well-know constraint for artificial and biological neural systems. The basic idea is that learning in a parallel and distributed system requires plasticity for the integration of new knowledge, but also stability in order to prevent the forgetting of previous knowledge. Too much plasticity will result in previously encoded data being constantly forgotten, whereas too much stability will impede the efficient coding of this data at the level of the synapses. However, for the most part, neural computation has addressed the problems related to excessive plasticity or excessive stability as two different fields in the literature.

## THE PROBLEM OF CATASTROPHIC FORGETTING FOR DISTRIBUTED NEURAL NETWORKS

The problem of catastrophic forgetting has emerged as one of the main problems facing artificial neural networks. The problem can be stated as follow: a distributed neural system, for example any biological or artificial memory, has to learn new inputs from the environment but without being disrupted by them. Catastrophic forgetting is defined as a complete forgetting of previously learned information by a neural network exposed to new information (McCloskey and Cohen, 1989; Ratcliff, 1990). This problem is a general problem that exists in different types of neural networks from standard back-propagation neural networks to unsupervised neural networks like self-organizing maps (Richardson and Thomas, 2008) or for connectionist models of sequence acquisition (Ans et al., 2002). Concerning artificial connectionist neural networks (such as, for instance, standard backpropagation networks), they are highly sensitive to catastrophic forgetting because of their highly distributed internal representations (French, 1992). Therefore, it is possible to reduce the problem of catastrophic forgetting by reducing the overlap among the internal representations stored in the neural network, for example using larger systems, or for example sparse or interleaved learning (Hetherington and Seidenberg, 1989; McRae and Hetherington, 1993). For this reason, when learning input patterns, connectionist networks have to alternate between them and adjust the corresponding synaptic weights by small increments in order to appropriately associate each input vector with the related output vector. By contrast, sequential learning in a standard connectionist network would result in the complete forgetting of previously learned input-output patterns. This problem affecting artificial neural networks clearly distinguishes them from the cognitive abilities of biological neural systems that are able to learn new patterns in sequential order without catastrophic forgetting.

In order to prevent catastrophic forgetting, various researchers have suggested using a dual-memory system which, fundamentally, simulates the presence of a short-term and a long-term memory (Robins, 1995; Ans and Rousset, 1997; French, 1997; Mermillod et al., 2003). The principle is to consolidate information, initially present in a short-term memory, within a long-term memory in order to prevent catastrophic forgetting in connectionist systems. This principle, investigated within the perspective of neural computation in artificial systems, could also point the way to a more general principle that also applies to biological neural systems (French, 1999).

## THE ENTRENCHMENT EFFECT: THE OPPOSITE EXTREME OF THE PLASTICITY-STABILITY DILEMMA

At the opposite extreme of the stability-plasticity continuum lies the entrenchment effect, which may contribute to age-limited learning effects (Zevin and Seidenberg, 2002; Bonin et al., 2004, 2009; Mermillod et al., 2009a). In the cognitive sciences, this research field emerged as part of the attempt to determine whether items which are acquired early in life are better memorized in adults than those which are acquired later in life. Various studies working within this perspective have shown that words acquired early are processed faster and more accurately than words acquired later in life (see Juhasz, 2005; Johnston and Barry, 2006 for reviews). These so-called age-of-acquisition effects have been found in a large variety of tasks, for example picture naming tasks, as well as in different populations (e.g., children and adults).

While distributed neural networks have long been used to address various issues in word recognition and spoken word production studies, they have also recently been used to investigate the computational basis of these age-limited learning effects (e.g., Ellis and Lambon Ralph, 2000; Zevin and Seidenberg, 2002; Lambon Ralph and Ehsan, 2006). In these connectionist models, lexical frequency is encoded in the strength of the connections between the different types of representations which

are involved in recognizing and producing words (Seidenberg and McClelland, 1989; Plaut et al., 1996). As far as connectionist simulations of age-limited learning effects are concerned, Ellis and Lambon Ralph (2000) were the first to show that the order of introduction of the encounters determines the number of errors produced by the neural network at the end of training. More precisely, the items introduced first in their study produced fewer errors than the late-introduced items, even after cumulative frequency had been carefully controlled for. This effect of age-limited learning effects in connectionist networks is referred to as the entrenchment effect.

At a computational level, the question is to understand how this entrenchment effect emerges. According to Zevin and Seidenberg (2002), the loss of plasticity in connectionist networks such as Seidenberg and McClelland's (1989) was due to the adjustments of the weights that occur on the basis of the logistic function used by the backpropagation algorithm and permits adjustments to the weights (initially set to random values between 0 and 1). These adjustments are at their largest when the activations occur in the middle of the logistic function (around 0.5) and become smaller as the weights converge on values that cause unit activations to approximate more closely to the target values (for instance 1 or 0). Thus, there is a loss of plasticity (early trained patterns become entrenched in the weights) associated with the learning of the first patterns in the training regime. Therefore, according to Zevin and Seidenberg (2002), the loss of plasticity in connectionist systems should vary as a function of the transfer function and the error signal computed. For example, a root mean square vs. cross-entropy error should produce different sensitivity to the entrenchment effect, but also to catastrophic forgetting. Of course, other factors as competition effects, loss of resources, and assimilation effects are important to produce age limited learning effects (Thomas and Johnson, 2006) and are important to control as possible confounded variables. In the current article, we suggest that the Fahlman offset (Fahlman, 1988) could constitute a simple and efficient way to test the computational basis of the loss of plasticity assumed by Zevin and Seidenberg (2002).

## THE FAHLMAN OFFSET: A WAY TO INVESTIGATE BOTH ENDS OF THE CONTINUUM

It is interesting to note that the above-mentioned research fields investigate two extremes of the same continuum. In other words, the entrenchment effect is related to a lack of plasticity (and an excess of stability) in response to newly acquired items, whereas catastrophic interference is related to an excess of plasticity (and a lack of stability) in response to new items presented sequentially. There are a number of ways of overcoming this difficulty, for instance by manipulating the orthogonality or the sparseness of the input-output patterns (French, 1992; Robins, 1995). However, among the different possibilities proposed to modulate the plasticity of neural networks, the method proposed by Fahlman (1988) is both simple and efficient. The basic idea is to add a constant number to the derivative of the sigmoid function (synaptic weights are adjusted by multiplying the error produced by a neuron by the derivative of the transfer function, i.e., the sigmoid function). This method makes it possible to avoid the entrenchment effect in the flat part of the sigmoid function and is relevant because this entrenchment effect is due to the flat spots at which the derivative of the sigmoid function approaches zero. Once the output value of a trained neural network starts to become entrenched around this flat spot of the sigmoid function, it becomes very difficult for the standard backpropagation algorithm to modify the synaptic weights responsible for producing this error. Even if an output value represents the maximum possible error, a unit whose output is close to 0.0 or 1.0 will be able to backpropagate only a tiny fraction of this error to the incoming weights and to units in earlier layers. Although it is theoretically possible to recover from entrenchment, this takes a very long time. The method proposed by Fahlman (1988), which consists of adding a small constant number to the derivative of the sigmoid function so that it does not go to zero for any output value, is therefore, both very simple and efficient to improve the efficiency of connectionist networks to simulate human cognitive processes (Mermillod et al., 2009b, 2010). For example, adding a constant of 0.1 to the sigmoid function before using it to scale the error prevents neuron values from approaching 0 and avoids the flat spots in the sigmoid function where the synaptic weights can become entrenched.

## NEW FINDINGS AND PERSPECTIVE

In a recent article (Mermillod et al., 2012), we showed that age of acquisition can be considered, at a computational level, as an extreme case of frequency trajectory (i.e., the frequency with which a word is encountered during a certain period of life) and can help explain age-limited learning effects. Interestingly, no age-limited learning effects appeared when we used a Fahlman offset of 0.1 whereas it reappeared when we used a Fahlman offset of 0.0. This result was not consistent with Ellis and Lambon Ralph (2000) who reported an age-limited learning effect despite the improvement in the plasticity of the neural network brought about by modulating the Fahlman offset. This could be due to differences in the number or size of the training set between the two studies (Ellis and Lambon Ralph, 2000 or Mermillod et al., 2012). Therefore, the role of the training set in modulating the effects of learning parameters on age-limited learning and catastrophic interference remains a target of further investigation (since these factors could have a combined effect with neural plasticity). However, our results were not unambiguous: modifying the plasticity of an identical neural network by manipulating the Fahlman offset clearly modified the ability of the neural network to simulate (or not) age-limited learning effects. On the other side of the continuum, when the Fahlman constant was set to 0.0, we observed the age-limited learning effects reported in the literature (Ellis and Lambon Ralph, 2000; Zevin and Seidenberg, 2002; Lambon Ralph and Ehsan, 2006). Moreover, one result that will surprise researchers working in the field of catastrophic forgetting is that this catastrophic forgetting effect was largely reduced after the period of entrenchment of synaptic weights (early acquired patterns for "adult" networks having been learnt at an early stage, compared to the medium and late patterns being learnt sequentially in a later stage). To conclude, we suggest here that investigating the plasticity-stability continuum by modulating the Fahlman offset

should help us understand a wide range of cognitive phenomena from age-limited learning effects through to catastrophic forgetting, as well as various forms of memory disorders.

## ACKNOWLEDGMENTS

This work was supported by a Institut Universitaire de France grant to Patrick Bonin and Martial Mermillod.

## REFERENCES

Ans, B., and Rousset, S. (1997). Avoiding catastrophic forgetting by coupling two reverberating neural networks. *C. R. Acad. Sci. III Sci. Vie* 320, 989–997.

Ans, B., Rousset, S., French, R. M., and Musca, S. (2002). "Preventing catastrophic interference in multiple-sequence learning using coupled reverberating elman networks," in *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*, eds W. D. Gray and C. D. Shunn (Mahwah, NJ: Lawrence Erlbaum Associates).

Bonin, P., Barry, C., Méot, A., and Chalard, M. (2004). The influence of age of acquisition in word reading and other tasks: a never ending story. *J. Mem. Lang.* 50, 456–476.

Bonin, P., Méot, A., Mermillod, M., Ferrand, L., and Barry, C. (2009). The effects of age of acquisition and frequency trajectory on object naming. *Q. J. Exp. Psychol.* 62, 1–9.

Ellis, A. W., and Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 1103–1123.

Fahlman, S. E. (1988). "Faster-learning variations on back-propagation: an empirical study," in *Proceedings of the 1988 Connectionist Models Summer School*, eds D. S. Touretzky, G. E. Hinton, and T. J. Sejnowski (Los Altos, CA: Morgan Kaufmann), 38–51.

French, R. M. (1992). Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connect. Sci.* 4, 365–377. doi: 10.1080/09540099208946624

French, R. M. (1997). Pseudo-recurrent connectionist networks: an approach to the "sensitivity–stability" dilemma. *Connect. Sci.* 9, 353–379. doi: 10.1080/095400997116595

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends cogn. Sci.* 3, 128–135. doi: 10.1016/S1364-6613(99)01294-2

Hetherington, P. A., and Seidenberg, M. S. (1989). "Is there 'catastrophic interference' in connectionist networks?" in *Proceedings of the 11th Annual Conference of the Cognitive Science Society*, (Hillsdale, NJ: Erlbaum), 26–33.

Johnston, R. A., and Barry, C. (2006). Age of acquisition and lexical processing. *Vis. Cogn.* 13, 789–845.

Juhasz, B. (2005). Age-of-acquisition effects in word and picture identification. *Psychol. Bull.* 131, 684–712. doi: 10.1037/0033-2909.131.5.684

Lambon Ralph, M. A., and Ehsan, S. (2006). Age of acquisition effects depend on the mapping between representations and the frequency of occurrence: empirical and computational evidence. *Vis. Cogn.* 13, 884–910.

McCloskey, M., and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. *Psychol. Learn. Motiv.* 24, 109–165. doi: 10.1016/S0079-7421(08)60536-8

McRae, K., and Hetherington, P. (1993). "Catastrophic interference is eliminated in pretrained networks," in *Proceedings of the 15h Annual Conference of the Cognitive Science Society*, (Hillsdale, NJ: Erlbaum), 723–728.

Mermillod, M., Bonin, P., Méot, A., Ferrand, L., and Paindavoine, M. (2012). Computational evidence that frequency trajectory theory does not oppose but emerges from age of acquisition theory. *Cogn. Sci.* 36, 1499–1531. doi: 10.1111/j.1551-6709.2012.01266.x

Mermillod, M., Bonin, P., Mondillon, L., Alleysson, D., and Vermeulen, N. (2010). Coarse scales are sufficient for efficient categorization of emotional facial expressions: evidence from neural computation. *Neurocomputing* 73, 2522–2531. doi: 10.1016/j.neucom.2010.06.002

Mermillod, M., Bonin, P., Morisseau, T., Méot, A., and Ferrand, L. (2009a). "Frequency trajectory gives rise to an age-limited learning effect as a function of input-output mapping in connectionist networks," in *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, eds N. Taatgen and H. van Rijn (Mahwah, NJ: Lawrence Erlbaum Associates), 2322–2327.

Mermillod, M., Vermeulen, N., Lundqvist, D., and Niedenthal, P. M. (2009b). Neural computation as a tool to differentiate perceptual from emotional processes: the case of anger superiority effect. *Cognition* 110, 346–357. doi: 10.1016/j.cognition.2008.11.009

Mermillod, M., French, R. M., Quinn, P. C., and Mareschal, D. (2003). "The importance of long-term memory in infant perceptual categorization," in *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, eds R. Alterman and D. Kirsh (Mahwah, NJ: Lawrence Erlbaum Associates), 804–809.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding normal and impaired reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115. doi: 10.1037/0033-295X.103.1.56

Ratcliff, R. (1990). Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychol. Rev.* 97, 285–308. doi: 10.1037/0033-295X.97.2.285

Richardson, F., and Thomas, M. S. C. (2008). Critical periods and catastrophic interference in self-organising feature maps. *Dev. Sci.* 11, 371–389. doi: 10.1111/j.1467-7687.2008.00682.x

Robins, A. (1995). Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connect. Sci.* 7, 123–146. doi: 10.1080/09540099550039318

Seidenberg, M. S., and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychol. Rev.* 96, 523–568. doi: 10.1037/0033-295X.96.4.523

Thomas, M. S., and Johnson, M. H. (2006). The computational modeling of sensitive periods. *Dev. Psychobiol.* 48, 337–344. doi: 10.1002/dev.20134

Zevin, J. D., and Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *J. Mem. Lang.* 47, 1–29. doi: 10.1006/jmla.2001.2834

*Received: 29 April 2013; accepted: 17 July 2013; published online: 05 August 2013.*

*Citation: Mermillod M, Bugaiska A and Bonin P (2013) The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. Front. Psychol.* ***4****:504. doi: 10.3389/fpsyg.2013.00504*

*This article was submitted to Frontiers in Language Sciences, a specialty of Frontiers in Psychology.*

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 16 August 2013
doi: 10.3389/fpsyg.2013.00528

# An amodal shared resource model of language-mediated visual attention

***Alastair C. Smith*[1,2]*, *Padraic Monaghan*[3] *and Falk Huettig*[1,4]**

[1] *Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands*
[2] *International Max Planck Research School for Language Sciences, Nijmegen, Netherlands*
[3] *Department of Psychology, Lancaster University, Lancaster, UK*
[4] *Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, Netherlands*

***Edited by:***
*Julien Mayor, University of Geneva, Switzerland*

***Reviewed by:***
*Barbara C. Malt, Lehigh University, USA*
*Daniel Mirman, Moss Rehabilitation Research Institute, USA*

****Correspondence:***
*Alastair C. Smith, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, Netherlands*
*e-mail: alastair.smith@mpi.nl*

Language-mediated visual attention describes the interaction of two fundamental components of the human cognitive system, language and vision. Within this paper we present an amodal shared resource model of language-mediated visual attention that offers a description of the information and processes involved in this complex multimodal behavior and a potential explanation for how this ability is acquired. We demonstrate that the model is not only sufficient to account for the experimental effects of Visual World Paradigm studies but also that these effects are emergent properties of the architecture of the model itself, rather than requiring separate information processing channels or modular processing systems. The model provides an explicit description of the connection between the modality-specific input from language and vision and the distribution of eye gaze in language-mediated visual attention. The paper concludes by discussing future applications for the model, specifically its potential for investigating the factors driving observed individual differences in language-mediated eye gaze.

**Keywords: language, vision, computational modeling, attention, eye movements, semantics**

## INTEGRATIVE PROCESSING IN A MODEL OF LANGUAGE-MEDIATED VISUAL ATTENTION

### LANGUAGE-MEDIATED VISUAL ATTENTION

Within daily communicative interactions a vast array of information sources have to be integrated in order to understand language and relate it to the world around the interlocutors. Such multimodal interactions within the speaker and listener have been shown to be vital for language development (Markman, 1994; Bloom, 2000; Monaghan and Mattock, 2012; Mani et al., 2013) as well as for adult sentence and discourse processing (Anderson et al., 2011; Huettig et al., 2011b; Lupyan, 2012). Eye gaze has been used to demonstrate the nature of the processes supporting online integration of linguistic and visual information (Halberda, 2006; Huettig et al., 2011a). Such observations of eye gaze have opened up the possibility to investigate how multiple sources of information, within the environment and within the language signal, interact in the human cognitive system. We begin by describing the observed properties of eye gaze behavior that have informed our understanding of the representations and processes involved in language—vision interactions. We then present a computational model of language-mediated visual attention that implements the representations and processes identified within a parsimonious neural network architecture. Finally, we demonstrate that many of the characteristic features of language-mediated eye gaze can be captured by the emergent properties of this parsimonious architecture and therefore do not necessitate separate information processing channels or modular processing systems.

One influential paradigm for measuring language and vision interactions is the Visual World Paradigm (VWP; Cooper, 1974; Tanenhaus et al., 1995), in which participants are presented with a visual display comprising a set of objects and/or actors whilst hearing an auditory stimulus and during this period their eye gaze is recorded. Although eye gaze is a measure of overt attention and thus not a direct reflection of linguistic processing, the VWP has been utilized largely to investigate questions that explore how the cognitive system processes spoken language (see Huettig et al., 2011b, for review). A few studies, however, have investigated multimodal interactions. Such studies tend to focus on how eye gaze alters as the auditory stimulus unfolds and how varying the relationships between objects in the display can highlight which modalities of information are implicated at varying time points in language processing.

Many visual world studies have demonstrated that eye gaze can be modulated by phonological relationships between items presented in the visual display and spoken target words. Allopenna et al. (1998), for instance, showed that when hearing a target word (e.g., "beaker") participants looked more toward items in the display whose names overlapped phonologically with the target word either in initial (e.g., beetle) or final (e.g., speaker) positions, than items that were not related phonologically (e.g., carriage) to the spoken target word. They found that, relative to unrelated items, there was increased fixation of phonological competitors. Furthermore, fixations to onset competitors occurred earlier than those to rhyme competitors and the probability of fixating onset competitors was greater than the probability of fixating rhyme competitors.

Visual relationships between items have also been shown to influence fixation behavior (Dahan and Tanenhaus, 2005; Huettig and Altmann, 2007). Dahan and Tanenhaus (2005) presented scenes containing a target (e.g., a snake), a visual competitor (e.g., a rope) and two unrelated distractors (e.g., a couch and an umbrella), while Huettig and Altmann (2007) presented scenes without a visual depiction of the target but with a visual competitor and three unrelated distractors. Thus, items within the display that shared visual features associated with the spoken target word, yet whose names did not overlap phonologically with the target word, attracted greater fixation than unrelated items.

Another dimension in which relationships between visually displayed items and spoken target words has been shown to modulate eye gaze is that of semantics. Huettig and Altmann (2005) and Yee and Sedivy (2006) demonstrated that items that share semantic (but not visual or phonological) relationships with target words are fixated more than unrelated items. Yee and Sedivy (2006) presented displays containing a target item (e.g., lock), a semantically related item (e.g., key) and two unrelated distractors. Similarly, Huettig and Altmann (2005) presented scenes containing both a target (e.g., piano) and a semantic competitor (e.g., trumpet) or scenes containing only a semantic competitor (e.g., only the trumpet) and unrelated items. In both target present and target absent conditions increased fixations of semantically related items were observed. *Post-hoc* analyses revealed that the likelihood of fixation was proportional to the degree of semantic overlap as measured by feature production norms (cf. Cree and McRae, 2003) and corpus-based measures of word semantics (Huettig et al., 2006). Further evidence for a relationship between semantic overlap and eye gaze is provided by Mirman and Magnuson (2009) who directly tested the gradedness of semantic overlap. They presented scenes containing a target item (e.g., bus) paired with either a near semantic neighbor (e.g., van) or a distant semantic neighbor (e.g., bike) and two unrelated items (e.g., ball). The likelihood of fixating each item was predicted by the level of semantic overlap, with near semantic neighbors fixated with greater probability than far semantic neighbors, while both were fixated with lower probability than targets and greater probability than distractors (see **Figure 1**).

In order to probe the relationships between previously observed phonological, visual and semantic word level effects in the VWP, Huettig and McQueen (2007) presented scenes containing phonological onset, semantic and visual competitors in addition to an unrelated distractor. They observed distinct patterns of fixation for each competitor type, with participants initially looking more toward phonological onset competitors before later displaying greater fixation of visual and semantic competitors. From these results they concluded that language-mediated visual attention is determined by matches between information extracted from the visual display and speech signal at phonological, visual and semantic levels of processing.

Taken together, this significant body of evidence shows that visual, semantic and phonological information is co-activated and integrated during spoken word processing. However, the nature of the information and mechanisms involved in visual world and language processing interactions are as yet underspecified (Huettig et al., 2011a, 2012). How is information activated within one modality integrated with information activated within another, what form does this information take, how does such information interact and how is this interaction connected to eye gaze behavior? There are two principle possibilities for interactions to occur: They may be a consequence of modality specific systems interacting via direct connections; alternatively, interactions may occur as a consequence of amodal shared resources facilitating interaction between the various information modalities (Lambon Ralph and Patterson, 2008; Plaut, 2002). Computational implementation of theoretical models offers a means of testing their plausibility and often provides a means of probing aspects of theoretical models that may lie beyond the reach of behavioral studies. The VWP provides a high degree of experimental control that offers a well constrained environment in which models can operate. Models of the processes involved in performing VWP tasks force researchers to define explicitly how information carried in the visual and auditory stimuli is connected to distributions of eye gaze.

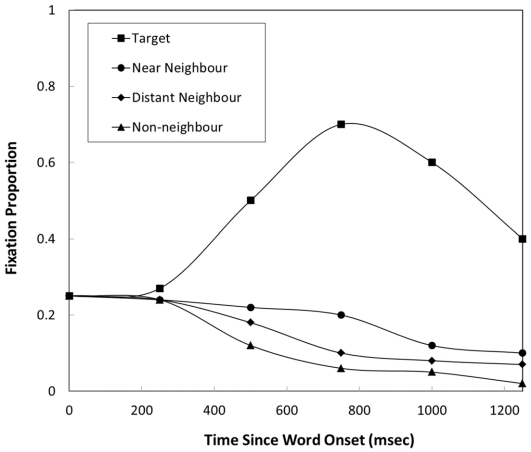



**FIGURE 1 | Figure adapted from Mirman and Magnuson (2009).** Figure displays approximate fixation proportions for targets, near semantic neighbors, distant semantic neighbors and unrelated items displayed by participants in Mirman and Magnuson (2009).

In this paper, we first present previous modeling approaches that have accounted for the various VWP results presented above before elaborating the modular vs. shared-resource computational approaches to multimodal information processing. We then present our model of the shared resource account of language-mediated visual attention and demonstrate that it is not only sufficient to account for the experimental effects of VWP studies but also that these effects are emergent properties of the architecture of the model itself.

## PREVIOUS MODELS OF LANGUAGE-MEDIATED VISUAL ATTENTION

Most previous models of the VWP have focused on explaining interactions between vision and a single feature of language processing. For instance, Allopenna et al. (1998) chose TRACE (McClelland and Elman, 1986) to simulate the mechanisms driving differences in the effect of phonological onset and rhyme overlap. TRACE is a continuous mapping model of speech perception,

implemented in an interactive activation network that hierarchically processes speech at the level of phonemic features, phonemes and words. The model successfully replicated the contrasting patterns of fixation displayed by participants toward onset and rhyme competitors and offered explanation for contrasts between the location of overlap and its influence on eye gaze. However, the model focuses purely on phonological processing and therefore as a model of language-mediated visual attention it provides no description of the role other information sources play in this process.

Magnuson et al. (2003) further examined the mechanisms underlying observed cohort and rhyme effects. They demonstrated that differences in sensitivity to both cohort and rhyme competitors displayed by adults over the course of word learning could be captured in the emergent behavior displayed by an SRN (Elman, 1990) trained to map between phonetic features and localist word level representations. Unlike TRACE, in which connection weights were fixed by the modeler, connection weights within the SRN were adjusted using an error based learning algorithm. This not only reduces the number of parameters directly manipulated by the modeler and therefore the number of assumptions underpinning the model but also allowed authors to chart model behavior over the course of word learning. Using this approach they were able to demonstrate that a fundamental difference between adult and child lexical representations was not required to explain differences in sensitivity to rhyme and cohort competitors. Instead such differences were captured by their model due to the strengthening of lexical representations over the course of word learning. Again, however, the focus of this work is on aspects of phonological processing in the VWP. Therefore, as a model of language-mediated visual attention it ignores the role of other knowledge types in this process.

Similarly, Mirman and Magnuson (2009) used the attractor network of Cree et al. (1999) to simulate the graded effect of semantic competitors influencing eye gaze. The network consisted of a word form input layer and semantic feature output layer. The model was trained to map 541 words onto their corresponding semantic features derived from feature norming studies. However, as in the case of Magnuson et al. (2003) and TRACE, such models offer representation of items from only a single information source (phonological or semantic similarity) and therefore are unable to account for the full range of intermodal effects demonstrated in the VWP. Also, none of these models offer a description of how information activated by distinct visual and auditory sources can be combined to influence fixation behavior. They therefore do not provide a comprehensive model of the word level effects observed in the VWP.

There have, however, been some notable models of multimodal processing in VWP (Spivey, 2008; Mayberry et al., 2009; Kukona and Tabor, 2011). Spivey (2008) extended TRACE to incorporate visual processing, by connecting lexical activations in TRACE to a normalized recurrent localist attractor network that represented the presence or absence of items within the visual environment. However, in using localist visual representations the model lacks depth of representation in the visual modality to capture subtle relationships between items known to influence fixation behavior in VWP, such as visual similarity effects.

Mayberry et al. (2009) also provided a model of the interaction between visual and linguistic information in the VWP. Their connectionist model (CIANET) displays emergent properties that capture sentence level effects such as case role interpretation. A potential weakness of the model is its use of the same form of representation to encode both visual and linguistic information, thereby masking potential distinct effects of visual vs. linguistic similarities. A further weakness of both CIANET and Spivey (2008) is that neither provide representation at the word level in a semantic dimension, although we know from previous VWP studies that items can differ in both visual and phonological dimensions yet still share semantic properties that influence eye gaze behavior.

Finally, Kukona and Tabor (2011) presents a dynamical systems model of eye gaze in VWP in which localist representations at phonological, lexical-semantic, cross-word and action-space layers interact in a hierarchically structured network. Visual information is modeled by the presence or absence of its corresponding representations within the network. By representing items at this level of abstraction their model is unable to capture complex relationships between representations in the same modality. It seems then that none of the current multimodal models that have been used to explicitly model VWP data offer sufficient depth of representation in the multiple modalities involved to capture the subtle relationships between items shown to influence eye gaze at the word level in VWP.

Yet, previous models and their success in replicating individual VWP data sets have provided valuable insight into the type of architecture capable of supporting language-mediated visual attention. The architecture must allow for competition at multiple levels of representation (Allopenna et al., 1998), allow both excitatory and inhibitory connections (Mirman and Magnuson, 2009), facilitate parallel activation of representations (Kukona and Tabor, 2011) and integrate information from multiple sources (Mayberry et al., 2009). Such integration could be accomplished by connectivity between individual representational modalities, or via processing interconnectivity through a shared resource.

## MODULAR vs. SHARED-RESOURCE MODELS

A framework able to capture the architectural features of language-mediated visual attention identified in the previous section is the Hub-and-spoke (H&S) framework. H&S models are defined by an amodal central resource (hub) that integrates and translates information between multiple modality specific sources (spokes). The framework arose as one side of a debate regarding the neural structures that support human conceptual and semantic knowledge. Lambon Ralph and Patterson (2008) compared two alternative theoretical models to account for visual and linguistic semantic processing in unimpaired and patient populations. One consisted purely of modality specific processing regions connected via direct connections, the second instead connected regions via a modality invariant central hub, the H&S model. The authors argue that although a web of direct connections may provide a simpler architectural solution, only a model that contains a central connecting hub offers a system capable of performing the multilevel non-linear computations required

for semantic generalization and inference based on conceptual structure rather than surface similarities. There is also converging empirical evidence for both the existence of a semantic hub and its implementation in specific neural populations in the anterior temporal lobe (ATL). This evidence includes neuropsychological studies of patients suffering from semantic dementia (SD) (Lambon Ralph et al., 2010) who possess lesions in the ATL and display deficits in performance on tasks requiring semantic generalization. Similarly, non-patient groups that experience artificial lesions in the ATL using rTMS (Pobric et al., 2007) have reported similar deficits in performance on such tasks. Finally, neuroimaging studies (Vandenberghe et al., 1996), have observed activity in the ATL on tasks that require semantic generalization. These data support the notion that a central resource that integrates modality specific information is a crucial component of the architecture supporting semantic processing.

Models that postulate integrative processing from multiple sources are embedded in a broader literature that has debated the inherence of sensory and motor systems to conceptual representations. Studies of "embodied cognition," for instance, have made the case for the importance of motor and sensory systems for cognitive processing (e.g., Barsalou et al., 2003, but see Mahon and Caramazza, 2008). An important debate concerns the format of mental representations with some proponents of the embodied cognition hypothesis suggesting that conceptual knowledge consists entirely of "representational codes that are specific to our perceptual systems" (Prinz, 2002, p. 119). This contrasts with representational theories which assume that sensory and motor knowledge is amodal and abstracted away from modality-specific systems (e.g., Kintsch, 2008). A third view posits the existence of both amodal and modal representations providing an explanation of how we are able to acquire knowledge which goes beyond sensory and motor experience (Goldstone and Barsalou, 1998; Dove, 2009). This view is supported by recent demonstrations that co-activation of multimodal systems can be effectively simulated by models with an amodal shared resource (Yoon et al., 2002; Monaghan and Nazir, 2009). Given that activation in a spoke of a H&S model represents modality specific processing of an item, and activation within the hub captures an items amodal properties, then the interaction of modal (spoke) and amodal (hub) representations is a natural consequence of the architecture of H&S models. A recent review of the mechanisms and representations involved in language-mediated visual attention (Huettig et al., 2012) concluded that the most promising theoretical model to date postulates that language-mediated visual attention is dependent on a system in which both linguistic, non-linguistic and attentional information are all instantiated within the same coding substrate, which is required in order for information to be bound across modalities. The H&S framework offers a parsimonious solution by connecting modalities through a central processing hub.

Research examining the plausibility of alternative theoretical models of multimodal cognition has profited from testing their predictions using explicit neural network implementations of the H&S framework. In the following sections we detail the nature of these studies and how they have contributed to our understanding of the mechanisms that support semantic processing. We also identify the features of the Hub and Spoke framework that make it a valuable tool for modeling various aspects of multimodal cognition. We then test the framework's scope by using it as a foundation for a model of language-mediated visual attention.

## INSIGHTS FROM HUB AND SPOKE MODELS

The H&S framework offers a parsimonious architecture in which single modality models can be drawn together to examine the consequences of multimodal interaction. Producing an explicit model of the mechanisms thought to underlie a given process allows one to test theoretical positions and probe deeper the mechanisms that may be involved in a controlled and tractable manner.

The framework provides a single system architecture with only minimal initial assumptions on connectivity. As the systems architecture imposes minimal constraints on the flow of information within the network, emergent behavior is largely driven by (1) representational structure and/or (2) the tasks or mappings performed by the system during the learning process. Therefore, within the framework the scope of such factors in driving emergent properties of complex multimodal systems can be examined largely independent of modality specific architectural constraints.

Two alternative means of exploring the role of representational structure are presented in previous H&S models. Plaut (2002) simulates impairments displayed by optic aphasics in an H&S model that mapped between distinct vision, action (gesturing), touch and phonological (naming) layers. The author takes a fundamentalist approach (see Kello and Plaut, 2000) ensuring he has total control over any relationships embedded in representations within or across modalities. This allows the study to isolate emergent properties driven by individual aspects of representational structure. In Plaut (2002) the variable manipulated was systematicity in representation between modalities. He embedded systematic mappings between tactile, vision and action representations while those between phonology and other modalities were arbitrary. This feature of representations allowed the model to capture key features of patient behavior with the lack of systematicity in phonological representations leading to poor performance on naming tasks post lesioning.

In contrast, Rogers et al. (2004) (approach replicated in Dilkina et al., 2008, 2010) employs a realist approach with representations derived from feature norming studies. Within the study deficits in semantic processing displayed by SD patients are modeled using an H&S framework. The model consisted of a visual layer connected via a central resource to a verbal descriptor layer comprising names, perceptual, functional, and encyclopaedic information about objects. Although a realist approach requires the modeler to relinquish control over the structure embedded within the corpus, the resulting structure aims to provide a closer representation of that available within the natural learning environment. Consequently, this reduces the extent to which emergent properties are determined by prior assumptions of the modeler and provides a means of examining the content of behavior determined by naturally occurring structure within the environment. The model presented in Rogers et al. (2004), generates the counterintuitive prediction that damaged semantic systems are more likely to perform better at specific relative to

general sorting in the case of fruits. This subtle aspect of behavior is captured as a result of the model implementing rich representations of the structure of information available within the environment.

With small corpora it is also possible to analyse the structure embedded within representations derived from natural data to identify features that may have an influence on emergent behavior. This is demonstrated in Dilkina et al. (2010), in which individual differences displayed by SD patients were modeled in an H&S framework that mapped between orthographic, action, vision and phonological layers. The behavior of a subset of SD patients whose performance on lexical and semantic tasks did not correlate by item had been argued to result from two functionally distinct systems (e.g., Coltheart, 2004). The study demonstrated the compatibility of a single system model with the empirical data and offered an alternative explanation based on the role of spelling and concept consistency. The authors argued that observed effects emerged due to the structure embedded within representations rather than modality specific architectural constraints.

Behavior is not only constrained by representational structure but also by the manner in which the system interacts with representations, for example the form and quantity of mappings demanded by the learning environment. H&S models have demonstrated how the framework is able to examine the consequences of such environmental factors. Dilkina et al. (2010) captures contrasts in mappings over the course of development. Training is split into two stages, with mapping from orthography to phonology only performed in the second stage. This aims to reflect the fact that learning to read only occurs at a later stage of development. The proportion and period in which certain mappings such as vision to action occur may remain relatively constant both over the course of development and populations. However, it is also true that in many cases there will be variation in the form and quantity of mapping between individuals and more broadly populations. Dilkina et al. (2008) uses this feature of the learning environment to explore one possible factor driving individual difference in SD, that being the level of prior reading experience. Within the study, prior reading experience is modeled by manipulating the amount of training on orthographic to phonological mapping. Demonstrating the influence of such factors, manipulation of this variable was able to account for four of the five SD patients behavior. Clearly, such variation in the type of mapping performed and stage at which it's performed can have dramatic consequences for emergent properties of the system. However, predicting the nature of such properties in complex multimodal systems is far from trivial. H&S offers a means of examining the consequences of variation in such environmental variables.

To conclude, behavioral data from the VWP suggests that language-mediated visual attention is driven by the interaction of information extracted from the visual environment and speech signal at semantic, visual and phonological levels of processing. The H&S framework provides a parsimonious architecture within which the emergent properties of this complex interaction can be modeled. Previous modeling of the VWP has identified further properties of the architecture involved. These include allowing competition at multiple levels of representation, parallel activation of representations, the integration of information from multiple sources and allowing inhibitory and excitatory associations. A neural network architecture such as those used in previous implementations of the H&S framework naturally captures these characteristics.

## INVESTIGATION GOALS

We next present a computational model of the various sources of information contributing to eye gaze in the VWP. Our aims were as follows. First, we tested whether a H&S model, with minimal computational architectural assumptions, was sufficient for replicating the effects of phonological and semantic influences on language processing in the VWP, or whether individual models combining the modal-specific features of the models of Allopenna et al. (1998) and Mirman and Magnuson (2009) would be required to effectively simulate the range of effects across these distinct modalities. Second, we tested whether the model could further generalize to simulate effects of visual information similarity in the VWP (Dahan and Tanenhaus, 2005; Huettig and Altmann, 2007). Third, we tested whether the model was further able to replicate sensitivity to the effects of presenting or not presenting the object corresponding to the target word in the various VWP experimental manipulations of visual, phonological, and semantic competitors. In each case, the model's performance is a consequence of the integrated processing of multimodal information, resulting from specified properties of the representations themselves and also the computational properties of the mappings between them.

The model we present connects visual, semantic and linguistic information to drive eye gaze behavior. Specifically, the model was tested on its ability to replicate the following features of language-mediated visual attention demonstrated in Visual World studies: (1) Fixation of onset and rhyme competitors above unrelated distractor levels in target present scenes (Allopenna et al., 1998); (2) Fixation of visual competitors above unrelated distractor levels in both target present (Dahan and Tanenhaus, 2005) and target absent (Huettig and Altmann, 2007) scenes; and (3) Fixation of semantic competitors above unrelated distractor levels and relative to semantic relatedness in both target present (Yee and Sedivy, 2006; Mirman and Magnuson, 2009) and absent (Huettig and Altmann, 2005) scenes. We present two simulations—one with no environmental noise, and one with background environmental noise. We later show that environmental noise is necessary for replicating all aspects of behavioral data.

## MODELING LANGUAGE-MEDIATED VISUAL ATTENTION IN A NOISELESS LEARNING ENVIRONMENT

### *Method*

***Architecture.*** The architecture of the H&S neural network used within this study is displayed in **Figure 2**. Akin to previous H&S models it was composed of a central resource (integrative layer) consisting of 400 units that integrated modality specific information from four "visible" layers, which encoded input and output representational information. The vision layer consisted of 80 units and modeled the extraction of visual information from four spatial locations within the environment. It contained four

slots each containing 20 units which extracted visual information from each of four distinct locations in the visual field. The phonological layer consisted of 60 units and encoded phonological information from the speech signal. This layer comprised six phoneme slots each represented by 10 units, such that words up to 6 phonemes in length could be represented unfolding across time. A semantic layer of 200 units represented semantic information of items, with units representing semantic features of the concept. The eye layer consisted of four units. Each unit within the eye layer was associated with one of the four locations within the model's visual field. The level of activation of an eye unit represented the probability of fixating the spatial location with which the unit was associated. All visible layers were fully connected to the central integrative layer, and the central integrative layer was in turn fully self-connected and fully connected to the eye and semantic layers.

At each time step of the model's processing, activation passed between all layers of units in the model. During training, there were 14 time steps to enable activation to cycle between representations in the model. During testing, the number of time steps was extended to enable insight into the time-course of representational information interacting between the modalities within the model.

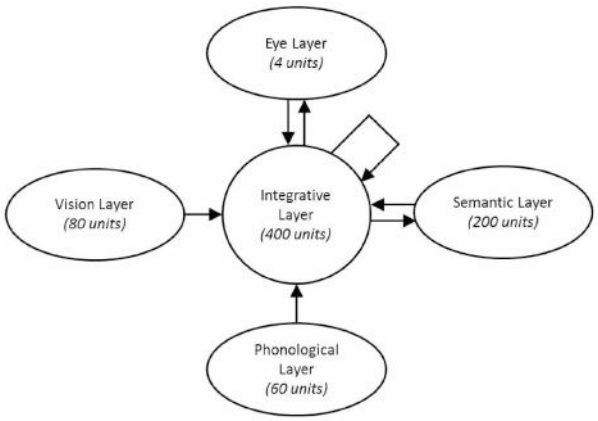



**FIGURE 2 | Network Architecture.**

***Artificial corpus.*** A fundamentalist approach (Kello and Plaut, 2000) was taken in construction of representations to ensure all aspects of the representations were controlled within simulations. Therefore, an artificial corpus composed of 200 items each with unique phonological, visual and semantic representations was constructed and used to train and test the model. Visual representations were generated to represent visual features in different spatial locations, with features representing both coarse (low frequency) and fine (high frequency) visual features. Phonological representations were encoded to create time-dependent slots for the unfolding speech, with categorical representations of phonemes shared across different words. Semantics in the model were rich, in that they were distributed feature based representations with structured relationships between items. They were also relatively sparse and discrete, reflecting behavioral studies of semantic feature-based representations (Harm and Seidenberg, 2004).

Visual representations were encoded as 20 unit binary feature vectors, with each unit representing the presence or absence of a given visual feature. Features were assigned to items randomly with $p(\text{active}) = 0.5$. Phonological representations consisted of a fixed sequence of six phonemes. Words were constructed by randomly sampling phonemes from a phoneme inventory containing a total of 20 possible phonemes. Each phoneme was encoded by a 10 unit binary feature vector, with $p(\text{active}) = 0.5$. For semantic representations, a unique subset of 8 semantic features was randomly assigned to each item from the set of 200 possible features.

The level of overlap between items in semantic, visual and phonological dimensions was controlled (see **Table 1**). Within the corpus were embedded 20 target items each with either visual, near semantic, far semantic, phonological onset or rhyme competitors. Competitors were defined by the increased number of features shared with their assigned target in either a semantic, visual or phonological dimension. A consistent level of representational overlap was implemented across all modalities (other than in the case of far semantic competitors) by ensuring that

**Table 1 | Controls used in the construction of artificial corpora and mean cosine distance calculated between targets, competitors and unrelated items all six randomly generated corpora used to train and test models.**

| Modality | Item | Artificial corpus | |
|---|---|---|---|
| | | **Constraint (*Features shared with target*)** | **Cosine distance ($\overline{x}$, $\sigma$)** |
| Phonological | Onset competitor | First 3 phonemes | 0.259 (0.026) |
| | Rhyme competitor | Final 3 phonemes | 0.260 (0.028) |
| | Unrelated | Max. 2 consecutive phonemes | 0.496 (0.052) |
| Semantic | Near neighbor | 4 of 8 functional properties | 0.500 (0) |
| | Far neighbor | 2 of 8 functional properties | 0.750 (0) |
| | Unrelated | Max. 1 functional property | 0.959 (0.072) |
| Visual | Competitor | Min. 10 of 20 visual features | 0.264 (0.040) |
| | Unrelated | Features shared with $p = (0.5)$ | 0.506 (0.068) |

the distance in terms of shared features between a target and a competitor was on average half the distance of that between a target and unrelated item in the modality that defined the competitor type. Six randomly generated corpora were generated using different initial random seeds, to ensure that no accidental correspondences between particular representations occurred systematically.

Onset competitors shared the initial three phonemes with their corresponding target word. No two words shared their initial four phonemes. Rhyme competitors shared the final three phonemes with their assigned target. No two words shared their final four phonemes. No item within the corpus contained more than two identical phonemes per word and no more than two consecutive phonemes overlapped between two unrelated items. These constraints resulted in a cosine distance between phonological representations of 0.259 between onset competitors and targets, 0.260 between rhyme competitors and targets and 0.496 between unrelated items and targets.

The length of vectors used to encode representations in both semantic and visual dimensions was determined by the constraints placed on relationships between items in these modalities. In the case of visual competitors, 10 of 20 visual features were shared between the target and competitor with $p(\text{shared}) = 1$, remaining features were shared with $p(\text{shared}) = 0.5$. For all visually unrelated items features were shared with $p(\text{shared}) = 0.5$. Such controls resulted in a smaller visual feature cosine distance between visual competitors and target items than between unrelated items and targets (see **Table 1**).

In the semantic dimension, near semantic competitors shared 4 of 8 semantic features with their corresponding target, while 2 of 8 were shared between far semantic competitors and targets. Controls ensured that unrelated items shared a maximum of one semantic feature. Semantic feature cosine distance was least between near neighbors and targets, medial between far neighbors and targets and most between unrelated items and targets (see **Table 1**).

***Training.*** Model training simulated learning experience in the natural environment that leads to the acquisition of associations between representations across modalities. We assume that individuals acquire semantic, visual and phonological knowledge of a given item through experience of repeated and simultaneous exposure to these multiple forms of representation within the natural learning environment. The model was trained on four cross modal tasks (see **Table 2**).

To simulate the events that lead to associations between an item's visual and semantic properties, the model was trained to map from visual to semantic representations using the following procedure. An example of such an event in the natural learning environment may be viewing an item while simultaneously experiencing some aspect of its function (e.g., seeing and eating from a fork). At trial onset the model was presented with four visual representations randomly selected from the corpus assigned to the four spatial locations within the visual field. One of the four items was then randomly selected as a target and the eye unit corresponding to its location fully activated. Throughout the entire test trial small levels of variable noise was provided as input to the phonological layer to simulate ambient background sound. Once sufficient time has allowed for activation to pass from eye and visual layers to the semantic layer (at time step 3) the item's semantic representation was provided as a target.

Models were also trained to map between phonological and semantic representations, simulating the learning that occurs through simultaneous exposure to the sound of a given word and its semantic properties (i.e., hearing and observing the function of "fork"). First, an item was randomly selected as a target from the corpus. The phonological representation of the target was then provided to the phonological input layer as a staggered input, with one additional phoneme provided at each time step. Once activation of the fourth phoneme (uniqueness point for phoneme competitors and corresponding targets) had had sufficient time to influence activation in the semantic layer (time step 5), the item's semantic representation was provided as a target.

Two further tasks trained the model to orientate toward a visual representation of an item in a spatial location according to given phonological or semantic information. As previously stated we assume that in the natural learning environment individuals are repeatedly exposed simultaneously to the visual and phonological or semantic form of an item. Consequently, the learner determines the association between these representations. Mapping from phonology to location was trained by randomly selecting four items from the corpus, randomly assigning them to four locations, and randomly selecting one as the target. The visual representations relating to each of these items was presented as input to the visual layer at trial onset. At the same point in time, input of the phonological representation of the target item was initiated in the phonological layer with one additional phoneme presented per time step. Once activation relating to the fourth phoneme had had time to influence activation in the eye layer (time step 5), the eye unit corresponding to the location of the target was provided as the target.

For mapping from semantics to location, the trial was similar to the phonology to location task, except that all the semantic features were simultaneously activated at time step 1 and time variant noise was presented to the phonological layer for the entire training trial. Once activation from the semantic and visual layer had been provided sufficient time to influence eye layer activation (time step 2), the training signal was provided.

Training tasks were randomly interleaved. Within the natural learning environment we assume that individuals orientate toward or select items based on their semantic features far more frequently than they orientate toward or select items in response to hearing their name. To reflect the assumption that phonologically driven orienting occurs less frequently than semantically driven orienting, training on phonologically driven orienting was four times less likely to occur than all other training tasks.

All connection weights within the network were initially randomized in a uniform distribution $[-0.1, 0.1]$. Weights were adjusted using recurrent back-propagation with learning rate $= 0.05$. In order to simulate participants' prior ability to orientate to items based on their phonological and semantic form and identify items' semantic properties based on their visual or phonological form, the models were required to perform with high accuracy on all four of these tasks prior to testing. To obtain this level of

**Table 2 | Temporal organization of events in model training.**

| Task | Vision | | Phonological | | Semantic | | Eye | |
|---|---|---|---|---|---|---|---|---|
| | **Description** | **Time step** | **Description** | **Time step** | **Description** | **Time step** | **Description** | **Time step** |
| Visual to Semantic | 4 visual representations randomly selected from corpus, 1 assigned as target | 0–14 | Random time invariant noise provided as input | 0–14 | Semantic representation of target provided post display onset | 3–14 | Location of target activated, all other locations inactive | 0–14 |
| Phonological to Semantic | Random time invariant noise provided as input across all 4 input slots | 0–14 | Speech signal of target provided as a staggered input | 0–14 | Semantic representation of target provided post disambiguation | 5–14 | No constraints on activation | |
| Phonological to Location | 4 visual representations randomly selected from corpus, 1 assigned as target | 0–14 | Speech signal of target provided as a staggered input | 0–14 | No constraints on activation | | Post disambiguation location of target activated, all other locations inactive | 5–14 |
| Semantic to Location | 4 visual representations randomly selected from corpus, 1 assigned as target | 0–14 | Random time invariant noise provided as input | 0–14 | Semantic representation of target provided | 0–14 | Location of target activated, all other locations inactive post functional onset | 2–14 |

performance training was terminated after 1 million trials. In total 6 simulation runs of the model were trained and tested, using each of the six artificial corpora.

### *Results*

***Pre-test.*** Following training all models were tested to assess performance on each of the four training tasks for all items within the training corpus. Noise was presented to visual and phonological slots that did not receive target related input. For tasks presenting the target in the visual input, performance was recorded with the target tested once in each of the four locations in the visual field.

For mapping from visual to semantic representations, activation in the semantic layer was closer in terms of cosine similarity to the target item's semantic representation for all items within the training corpus. When tested on mapping from phonological to semantic representations activation in the semantic layer was also most similar to that of the target's semantic representation for all items within the training corpus.

For the phonology to location mapping task, the location of the target was selected on at least 3 of 4 test trials for 99.83% of items in the training corpus. Averaging across all phonology to location test trials the proportion of trials in which the eye unit corresponding to the location of the target was most highly activated was 92%.

For the semantics to location mapping task, the location of the target was selected on at least 3 of 4 test trials for 99.92% of items within the corpus. The overall proportion of successful semantic to location test trials was 89%.

***Simulation of visual competitor effects in the VWP.*** To simulate the conditions under which participants were tested in Dahan and Tanenhaus (2005), the model was presented with a visual display containing a target item, a visual competitor and two unrelated distractors. Simulations of Huettig and Altmann (2007) were conducted using a similar approach yet with targets replaced by an additional distractor. In both cases, the visual input representing four items was presented at time step 0. Then onset of the phonology for the target item began at time step 5, to enable pre-processing of the visual information. There were 480 test trials, with each item ($n = 20$) occurring with competitors in all possible spatial configurations ($n = 24$). The model's "gaze" was computed as the Luce ratio of the eye layer units, for the target, competitor, and unrelated distractor item. **Figure 3** displays the performance of the model when presented with target present (**Figure 3A**: simulating Dahan and Tanenhaus, 2005) and target absent (**Figure 3B**: simulating Huettig and Altmann, 2007) scenes, averaged over each of the six simulation runs of the model. For analysis we calculated the mean fixation proportions [p(fix)] for each category of item (i.e., target, competitor or unrelated distractor) from word onset until the end of the test trial. The ratio was then calculated between the proportion of fixations toward item type A and the sum of the proportion of fixations toward item type A and B. A ratio above 0.5 would indicate greater fixation of item A. Although we would not anticipate substantial variation in model performance across instantiations for completeness this mean ratio (by instantiation and by item) was compared to 0.5 using one sample *t*-tests (cf. Dahan and Tanenhaus, 2005) to

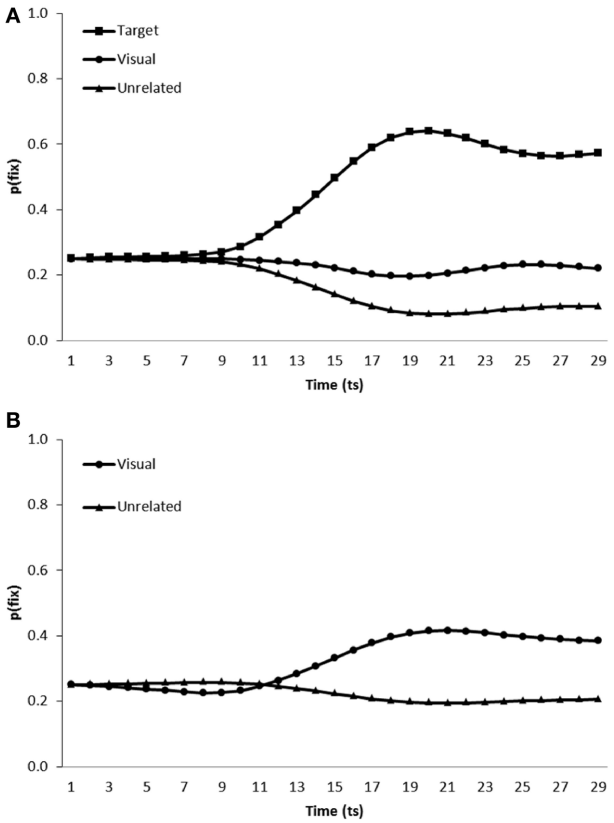



**FIGURE 3 | Proportion of fixations [p(fix)] directed toward items within scenes containing (A) a target, visual competitor and two unrelated distractors (B) a visual competitor and three unrelated distractors.**

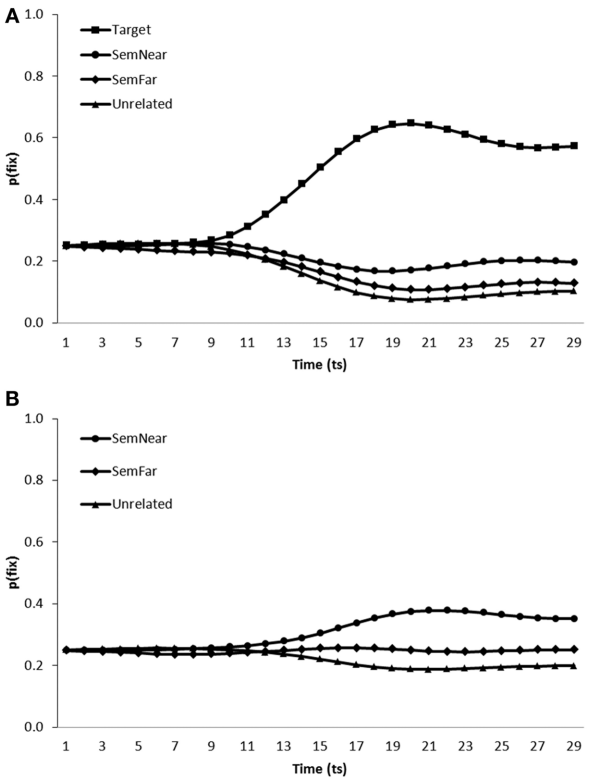



**FIGURE 4 | Proportion of fixations [p(fix)] directed toward items within scenes containing (A) a target, a near semantic neighbor (SemNear), a far semantic neighbor (SemFar) and an unrelated distractor, (B) a near semantic neighbor (SemNear), a far semantic neighbor (SemFar) and two unrelated distractors.**

test for differences in fixation behavior toward each category of item.

As can be observed from **Figure 3**, the model fixated target items [mean ratio = 0.75, $t_1(5) = 22.42$, $p < 0.001$; $t_2(19) = 78.50$, $p < 0.001$] and visual competitors [mean ratio = 0.60, $t_1(5) = 6.91$, $p = 0.001$; $t_2(19) = 18.18$, $p < 0.001$] more than unrelated distractors when scenes contained a target, visual competitor and two unrelated distractors (when by subjects and by items ratios are identical, only one ratio is presented). In target absent scenes, visual competitors were again fixated more than unrelated distractors [mean ratio = 0.58, $t_1(5) = 5.37$, $p < 0.01$; $t_2(19) = 15.290$, $p < 0.001$]. The model therefore replicates the increased fixation of visual competitors observed in Dahan and Tanenhaus (2005) and Huettig and Altmann (2007).

***Simulation of semantic competitor effects in the VWP.*** We simulated conditions similar to those under which participants were tested in Huettig and Altmann (2005), Yee and Sedivy (2006) and Mirman and Magnuson (2009) by testing model performance when presented with displays containing a near semantic neighbor and a far semantic neighbor in addition to either the target's visual representation and a single unrelated distractor (**Figure 4A**: simulating Mirman and Magnuson, 2009 and Yee and Sedivy, 2006) or two unrelated distractors (**Figure 4B**: Simulating Huettig and Altmann, 2005). As for the visual competitor effects, all items were presented in all combinations of positions in the visual input (480 trials in total), and again pre-processing of the visual features of the display were enabled by commencing word onset after a short delay (time step 5). **Figure 4** presents the average fixation proportions over time displayed by the model toward each category of item presented in both test conditions.

In target present trials, targets [mean ratio = 0.75, $t_1(5) = 25.89$, $p < 0.001$; $t_2(19) = 79.61$, $p < 0.001$], near semantic neighbors [mean ratio = 0.58, $t_1(5) = 5.37$, $p < 0.01$; mean ratio = 0.57, $t_2(19) = 9.89$, $p < 0.001$] and far semantic neighbors [mean ratio = 0.52, $t_1(5) = 2.82$, $p < 0.05$; mean ratio = 0.51, $t_2(19) = 4.07$, $p < 0.01$] were all fixated more than unrelated distractors. A similar pattern of behavior was observed when the model was tested on target absent trials, with both near [mean ratio = 0.58, $t_1(5) = 6.30$, $p < 0.01$; mean ratio = 0.57, $t_2(19) = 10.67$, $p < 0.001$] and far semantic neighbors [mean ratio = 0.53, $t_1(5) = 1.80$, $p > 0.1$; mean ratio = 0.52, $t_2(19) = 7.04$, $p < 0.001$] fixated more than unrelated items. Also in-line with behavioral findings far semantic neighbors were fixated less than near semantic neighbors, in both target absent [mean ratio = 0.44, $t_1(5) = -3.36$, $p < 0.05$; mean ratio = 0.45, $t_2(19) = -8.13$, $p < 0.01$] and target present [mean ratio = 0.44, $t_1(5) = -3.36$, $p < 0.05$; mean ratio = 0.44, $t_2(19) = -6.97$, $p < 0.001$] conditions.

The model therefore replicates the increased fixation of semantic competitors in both target absent and target present scenes as observed by Huettig and Altmann (2005) and Yee and Sedivy (2006) respectively, in addition to the graded effect of semantic similarity as reported in Mirman and Magnuson (2009).

***Simulation of phonological competitor effects in the VWP.*** To simulate the conditions under which participants were tested in Allopenna et al.'s (1998) study, the model was presented with scenes containing visual representations of a target item in addition to an onset competitor, a rhyme competitor and an unrelated distractor. For completeness we also tested model performance in a target absent condition (i.e., scenes containing onset competitor, rhyme competitor and two unrelated distractors). In every other way, simulations were conducted exactly as for the visual and semantic competitor simulations. **Figure 5** shows the average fixation proportions over time displayed by the model toward each category of item in test displays in both target present (**Figure 5A**) and target absent (**Figure 5B**) conditions.

In target present trials, target items [mean ratio = 0.75, $t_1(5) = 26.06$, $p < 0.001$, $t_2(19) = 66.45$, $p < 0.001$] and onset competitors [mean ratio =0.58, $t_1(5) = 6.20$, $p < 0.01$, $t_2(19) = 16.52$, $p < 0.001$] were fixated more than unrelated distractors. However, the model fixated rhyme competitors at levels similar to unrelated distractors [mean ratio = 0.51, $t_1(5) = 1.75$, $p > 0.1$, $t_2(19) = 1.69$, $p > 0.1$]. On target absent trials both onset [mean ratio = 0.59, $t_1(5) = 8.29$, $p < 0.001$; $t_2(19) = 15.62$, $p < 0.001$] and rhyme [mean ratio = 0.53, $t_1(5) = 5.62$, $p < 0.01$; mean ratio = 0.52, $t_2(19) = 3.05$, $p < 0.01$] competitors were fixated more than unrelated items. Allopenna et al. (1998) observed increased fixation of both onset and rhyme competitors in target present scenes. Model performance replicated the increased fixation of onset competitors displayed by participants. The model also displayed increased fixation of rhyme competitors although this effect was only clearly observable on target absent trials.

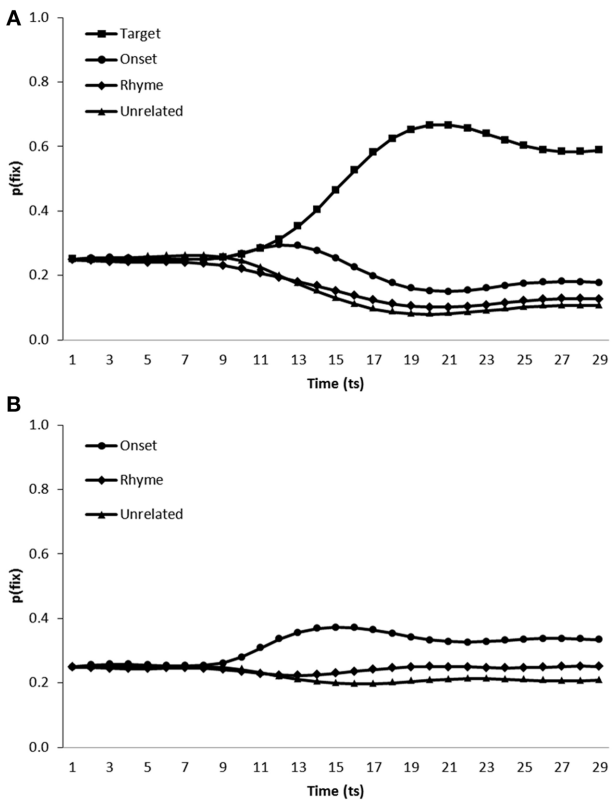



**FIGURE 5 | Proportion of fixations [p(fix)] directed toward items within scenes containing (A) a target, an onset competitor, a rhyme competitor and an unrelated distractor, (B) an onset competitor, a rhyme competitor and two unrelated distractors.**

### *Discussion*

The model was able to replicate a broad range of single modality word level effects described in the visual world literature, using a single architecture, and incorporating a single shared resource mapping between the modalities. The network replicates findings displaying a bias toward fixating items that overlap with spoken target words in either a visual, semantic or phonological dimension in both target present and absent scenes.

Importantly, the model captures differences in the effect of representational overlap between modalities. The model displays a graded effect of semantic overlap with the probability of fixating semantically related items proportional to the number of semantic features shared between the target and competitor. In a departure from the procedure used in Mirman and Magnuson (2009), within the above simulations both near and far semantic competitors were presented within the same display. Our simulations indicate that far semantic neighbor effects are robust to the additional competition that may result from the presence of closer semantic neighbors within the same scene.

For phonological overlap, the effect was dependent on the temporal location of overlapping features within the representation. Phonological overlap in onsets had a greater influence on fixation behavior than in rhymes, with the latter resulting in only marginal effects of overlap. Although many studies have demonstrated their existence (see Allopenna et al., 1998; Desroches et al., 2006; McQueen and Viebahn, 2007; McQueen and Huettig, 2012), rhyme effects tend to be weak and less robust than onset effects. However, a recent study by McQueen and Huettig (2012) provides evidence that the comparative onset effect is modulated by the level of noise present in the speech signal. They argue that the presence of noise influences the weight placed on initial phonemes as a predictor of the intended word. For example, in a noisy environment sounds heard may not necessarily relate to the identity of the target. Therefore, to make a judgement regarding an item's identity the system benefits from examining evidence from a larger portion of the auditory signal. This work highlights a weakness of current model training and testing, in that the model's learning environment always provided perfect perceptual input of an item in both visual and phonological representations. In the natural learning environment in which participants acquire their knowledge of items, the cognitive system is frequently receiving impoverished representations. This is particularly true in the case of speech, in which factors such as background noise or between speaker variation means that the speech signal received is likely to resemble only a very noisy version of the canonical form.

The following simulations extend the model by adding noise to the phonological representations to which the model is exposed during training.

## MODELING LANGUAGE-MEDIATED VISUAL ATTENTION IN A NOISY LEARNING ENVIRONMENT

### *Method*

To simulate exposure to noisy phonological input in the natural learning environment, the simulations were repeated but with noise applied to the phonological input during the training stage only. Noise was implemented by randomly switching the binary value of each unit within the phonological representation with $p = 0.2$. Noise was randomly generated for each training trial. To ensure comparable levels of performance between fully trained models on all four training tasks, the number of training trials performed was increased by 50%. In all other respects the procedure used to train and test the noisy model was identical to that applied to the previously detailed noiseless model.

### *Results*

***Pre-test.*** The noisy model displayed the same high level of performance on both visual to semantic mappings and phonological to semantic mappings as displayed by the noiseless model. In both cases, the noisy model produced activation in the semantic layer most similar (cosine similarity) to the target item's semantic representation for all items within the training corpus.

Performance on orientation tasks was also similar for models trained in both noise conditions. On phonological orienting test trials, the noisy model selected the location of the target on at least 3 of 4 test trials for 99.75% of items in the training corpus. The overall proportion of correct phonological orienting test trials (trials in which the eye unit corresponding to the location of the target was most highly activated) was 87% for the noisy model. When comparing the proportion of correct trials across instantiations between noise conditions, noiseless models performed significantly better than noisy models on this task ($p = 0.01$).

Noisy models correctly selected the location of the target as indicated by the presence of its semantic representation on at least 3 of 4 test trials for all items within the corpus. Overall accuracy on semantic orienting tasks for the noisy model was 90% ($\sigma = 0.02$). The difference between noisy and noiseless models was not significant on this task when comparing across instantiations.

***Simulation of visual competitor effects.*** **Figure 6** displays the performance of the noisy model when tested on scenes containing a visual competitor in addition to either the visual representation of the target and two unrelated distractors (**Figure 6A**) or no target and three unrelated distractors (**Figure 6B**).

On target present trials, both the targets [mean ratio = 0.77, $t_1(5) = 27.21$, $p < 0.001$; $t_2(19) = 89.97$, $p < 0.001$] and visual competitors [mean ratio = 0.62, $t_1(5) = 7.60$, $p < 0.01$; $t_2(19) = 22.22$, $p < 0.001$] were fixated more than unrelated distractors. Visual competitors were also fixated above distractor levels on target absent trials [mean ratio = 0.60, $t_1(5) = 14.52$, $p < 0.001$; mean ratio = 0.59, $t_2(19) = 18.75$, $p < 0.001$].

***Simulation of semantic competitor effects.*** The fixation behavior displayed by the noisy model on trials containing semantic competitors can be seen in **Figure 7**. The model was tested on scenes containing a near and far semantic neighbor in addition to either the target and a single unrelated distractor (**Figure 7A**) or no target and two unrelated distractors (**Figure 7B**).

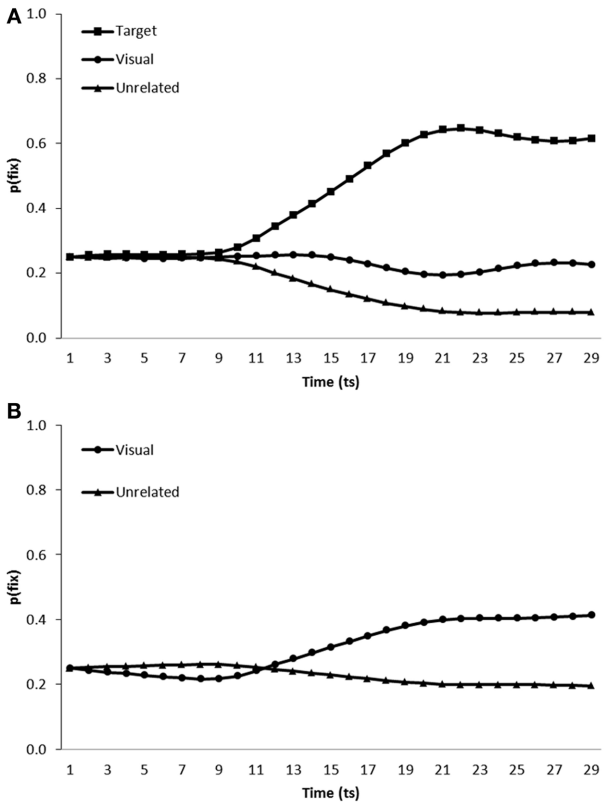



**FIGURE 6 | Proportion of fixations [p(fix)] directed toward items within scenes containing (A) a target, visual competitor and two unrelated distractors (B) a visual competitor and three unrelated distractors; by the model trained in a noisy learning environment.**

On target present trials, targets [mean ratio = 0.78, $t_1(5) = 29.48$, $p < 0.001$; mean ratio = 0.76, $t_2(19) = 102.21$, $p < 0.001$], near semantic neighbors [mean ratio = 0.62, $t_1(5) = 6.42$, $p < 0.01$; mean ratio = 0.60, $t_2(19) = 18.389$, $p < 0.001$] and far semantic neighbors [mean ratio = 0.54, $t_{1(5)} = 2.31$, $p < 0.1$; mean ratio = 0.52, $t_2(19) = 5.934$, $p < 0.001$] were all fixated more than unrelated distracters. On target absent trials, both near [mean ratio = 0.60, $t_1(5) = 13.78$, $p < 0.001$; mean ratio = 0.59, $t_2(19) = 22.51$, $p < 0.001$] and far [mean ratio = 0.53, $t_1(5) = 2.75$, $p < 0.05$; mean ratio = 0.52, $t_2(19) = 7.13$, $p < 0.001$] semantic neighbors were again more likely to be fixated than unrelated items. When comparing between near and far semantic competitors, far neighbors were fixated less than near neighbors both in target present [mean ratio = 0.42, $t_1(5) = -12.45$, $p < 0.001$; $t_2(19) = -12.81$, $p < 0.001$] and absent [mean ratio = 0.43, $t_1(5) = -11.81$, $p < 0.001$; $t_2(19) = -15.84$, $p < 0.001$] trials.

***Simulation of phonological competitor effects.*** Finally, the model was tested on scenes containing onset and rhyme

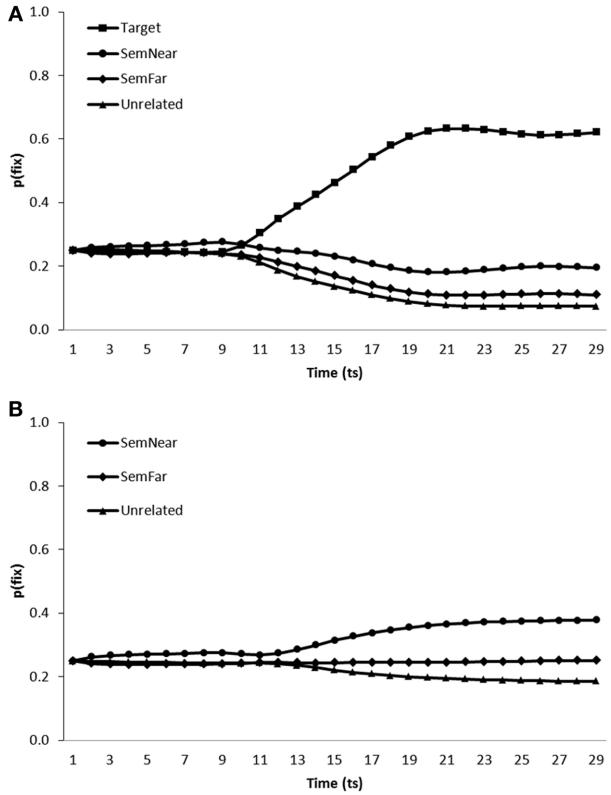



**FIGURE 7 | Proportion of fixations [p(fix)] directed toward items within scenes containing (A) a target, a near semantic neighbor (SemNear), a far semantic neighbor (SemFar) and an unrelated distractor, (B) a near semantic neighbor (SemNear), a far semantic neighbor (SemFar) and two unrelated distractors; by the model trained in a noisy learning environment.**

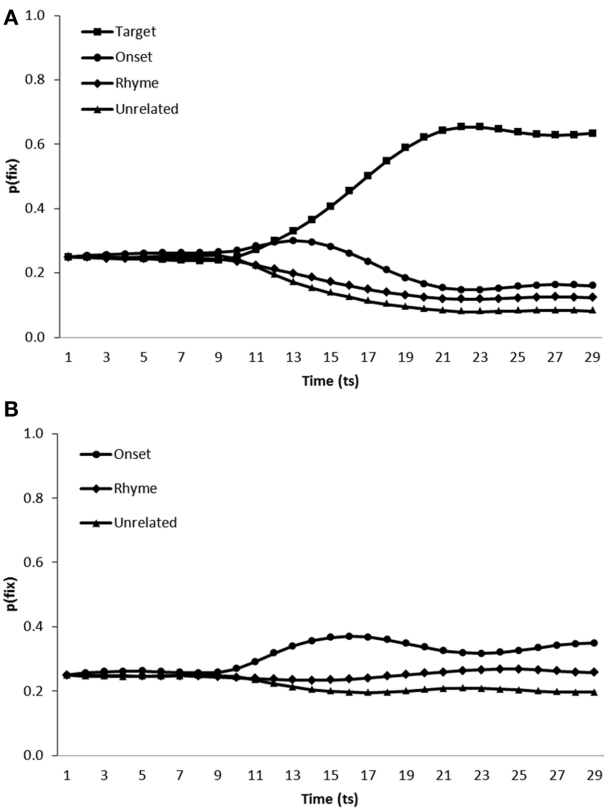



**FIGURE 8 | Proportion of fixations [p(fix)] directed toward items within scenes containing (A) a target, an onset competitor, a rhyme competitor and an unrelated distractor, (B) an onset competitor, a rhyme competitor and two unrelated distractors; by the model trained in a noisy learning environment.**

competitors in addition to either the target and a single unrelated distractor (**Figure 8A**) or two unrelated distractors (**Figure 8B**).

In target present scenes, the model displayed increased fixation of target items [mean ratio = 0.77, $t_1(5) = 36.71$, $p < 0.001$; $t_2(19) = 76.149$, $p < 0.001$], onset competitors [mean ratio = 0.60, $t_1(5) = 6.51$, $p < 0.01$; mean ratio = 0.61, $t_2(19) = 18.11$, $p < 0.001$] and rhyme competitors [mean ratio = 0.54, $t_1(5) = 3.13$, $p < 0.05$; $t_2(19) = 6.842$, $p < 0.001$] in comparison to unrelated distractors. Onset [mean ratio = 0.60, $t_1(5) = 11.09$, $p < 0.001$; $t_2(19) = 17.35$, $p < 0.001$] and rhyme competitors [mean ratio = 0.54, $t_1(5) = 3.13$, $p < 0.05$; $t_2(19) = 8.90$, $p < 0.001$] were also fixated more than distractors in target absent scenes.

### *Discussion*

The above results demonstrate that the model of language-mediated visual attention presented in this paper is still able to replicate a broad range of features of language-mediated visual attention when trained in a noisy learning environment. Further, and as predicted, by representing noise in the speech signal during training, we are able to replicate additional features of language-mediated visual attention, specifically sensitivity to rhyme competitors.

## GENERAL DISCUSSION

The amodal shared resource model presented here offers a description of the information and processes underlying language-mediated visual attention and a potential explanation for how it is acquired. The model accomplishes these effects with minimal imposed constraints on information processing modules or channels, and performance in the model is thus driven by representational structure and the different requirements of forming mappings between the distinct types of information. Language-mediated visual attention is simulated as a function of the integration of past and current exposure to visual, linguistic and semantic forms. The model thereby provides an explicit description of the connection between the modality-specific input from language and vision and the distribution of eye gaze in language-mediated visual attention.

The model replicated the following features of language-mediated visual attention demonstrated in VWP studies: Fixation of onset and rhyme competitors above unrelated distractor levels in target present scenes (Allopenna et al., 1998); (2) Fixation of visual competitors above unrelated distractor levels in target present (Dahan and Tanenhaus, 2005) and target absent (Huettig and Altmann, 2007) scenes; and (3) Fixation of semantic competitors above unrelated distractor levels and relative to

semantic relatedness in both target present (Yee and Sedivy, 2006; Mirman and Magnuson, 2009) and absent (Huettig and Altmann, 2005) scenes. A summary of the effects replicated by the model is presented in **Table 3**.

The results of the above simulations met the objectives of our study as follows. First, the model demonstrates that a H&S model, with minimal computational architectural assumptions, was sufficient for replicating the word level effects of phonological and semantic influences on language processing in the VWP. The simulation results replicate a broad range of the word level effects described within the VWP literature as features of this complex cognitive ability, without requiring separate resources or individually trained pathways between distinct representational information. Second, the model further generalized to replicate the effects of visual similarity in the VWP and sensitivity to the effects of presenting or not presenting the target object in various experimental manipulations of visual, phonological and semantic competitors.

Within our model language-mediated visual attention is described as an emergent property of the structure of representations present in the natural environment and the task demands imposed on the system by that environment. Knowledge of an item is acquired by repeated, simultaneous exposure to its multiple forms. For example, hearing the name of an object while looking at it, or experiencing the function of an item while hearing its name. Such experience leads to associations between the properties defining an object in separate modalities. With repeated and simultaneous exposure to their various forms inhibitory or excitatory connections between such properties are strengthened in order for the system to efficiently map between representations or carry out a given task. In this way, the model provides an explicit and detailed description of how multimodal knowledge of an item is acquired and stored, in addition to how complex multimodal behaviors such as selecting an item based on its function may be achieved and acquired. Thus, the model argues that many word level features of language-mediated visual attention are a necessary consequence of developing multimodal knowledge of items through such a mechanism.

Critically, the model captures contrasts in the effect of overlap in differing modalities. For example, for items that only overlap in a semantic dimension the probability of the model fixating an item is directly proportional to the number of semantic features the two items share. This replicates findings observed in the VWP in which the probability of fixating items has been predicted by semantic norming data (Mirman and Magnuson, 2009) and corpus-based measures of semantic similarity (Huettig et al., 2006). However, in the case of phonological overlap, the temporal location of the overlapping phonemes has a critical influence on the resulting effect. The model replicates the effects of phonological overlap observed in Allopenna et al. (1998) with items that share initial phonemes fixated earlier and with greater probability than items that share phonemes in final positions.

Within the model the level of overlap between target and competitor was strictly controlled both across modalities and between rhyme and onset competitors. Contrasts in fixation behavior toward differing categories of competitor therefore arise as an emergent property of differences in the structural characteristics of representations in each modality. For example, speech unfolds over time. Therefore, phonological representations have a temporal, sequential component not possessed by semantic or visual representations. As the speech signal gradually manifests, early phonemes provide a good, or in the case of a noiseless learning environment they provide a perfect, predictor of the intended word. Therefore, any item that shares the same initial sequence of phonemes with the target is more likely to be fixated by the model. By the time later phonemes are available, the system already has sufficient information, in the case of the noiseless simulations, to identify the target and therefore information provided by later phonemes does not have the opportunity to exert influence on target selection. It is for this reason increased sensitivity to rhyme competitors is displayed by a model trained in a noisy environment compared to one trained in a noiseless environment in which onset phonemes are perfect predictors of the unfolding word. Behavior of the noisy model demonstrates that introducing a low level of noise to speech in the learning environment is sufficient to allow the subtle influence of rhyme overlap to emerge.

This line of argument overlaps with the explanation provided in Magnuson et al. (2003) for the observed reduced sensitivity over the course of word learning to rhyme competitors. They argue that it takes time for the system to learn the value of early phonemes as predictors of the unfolding word. Therefore, at earlier stages of development other overlapping aspects of a word's phonology may exert equal or greater influence on target selection. In a noiseless environment an optimal model should display no influence of rhyme overlap, as sufficient information is

**Table 3 | Table comparing the results of both noiseless and noisy simulations with behavioral results reported in the VWP literature.**

| Study | Scene | | | | Effect A | | | Effect B | | |
|---|---|---|---|---|---|---|---|---|---|---|
| References | Item 1 | Item 2 | Item 3 | Item 4 | Behav. | Noiseless | Noisy | Behav. | Noiseless | Noisy |
| Allopenna et al., 1998 | Target | **Onset (A)** | **Rhyme (B)** | Dist | √ | √ (0.58) | √ (0.60) | √ | X(0.51) | √ (0.54) |
| Dahan and Tanenhaus, 2005 | Target | **Visual (A)** | Dist | Dist | √ (0.7) | √ (0.60) | √ (0.62) | | | |
| Huettig and Altmann, 2007 | **Visual (A)** | Dist | Dist | Dist | √ | √ (0.58) | √ (0.60) | | | |
| Yee and Sedivy, 2006 | Target | **Sem (A)** | Dist | Dist | √ | √ (0.58) | √ (0.62) | | | |
| Huettig and Altmann, 2005 | **Sem (A)** | Dist | Dist | Dist | √ | √ (0.58) | √ (0.60) | | | |
| Mirman and Magnuson, 2009* | Target | **Near Sem (A)** | **Far Sem (B)** | Dist | √ | √ (0.58) | √ (0.62) | √ | √ (0.52) | (0.54) |

*The items displayed within scenes in each empirical study are listed with observed competitor effects highlighted in bold. Competitor-Distractor ratios (by subject/instantiation) in parentheses if reported; √, behavioral effect replicated; X, failure to replicate behavioral effect; *, Study presented near and far semantic competitors on separate trials. Dist, distractor; Sem, semantic competitor; Onset, phonological onset competitor; Rhyme, phonological rhyme competitor.*

carried by initial phonemes to correctly identify the target item. However, in a noisy environment the optimal model would display sensitivity to rhyme overlap proportional to the level of noise in the environment, as this will dictate the probability that the rhyme competitor is the true target given the initially perceived input. Given this line of argument, it is not only external noise that would dictate a system's sensitivity to rhyme overlap but also the level of noise or error within the system itself. For example, noise simulated within the current model could equally reflect errors in phonological perception or fluctuations in attention, the contribution of which could possibly be examined through further combined modeling and VWP studies.

Similar to TRACE (McClelland and Elman, 1986), our model displays sensitivity to overlap in both phonological onsets and rhyme. However, there are differences between the models in their explanation for these effects. As in the model we present, TRACE is able to exploit similarity at all points within the phonological form of the word in terms of co-activating phonological competitors. However, unlike some previous models (Marslen-Wilson, 1987, 1993; Norris, 1994; Magnuson et al., 2003) and the model presented in this paper the disparity between sensitivity to cohort and rhyme competitors in TRACE is not driven by bottom-up mismatch but instead purely by onset competitors accumulating activation prior to rhyme competitors due to their inherent temporal advantage (Magnuson et al., 2003).

Many similarities are shared between our computational model and the theoretical model of language-mediated visual attention proposed in Huettig and McQueen (2007). Both models argue that behavior in the VWP is driven by matches between information extracted from visual and auditory input at phonological, semantic and visual processing levels. However, they differ subtly in how this is implemented. Huettig and McQueen suggest that contrasts in fixation dynamics displayed toward each category of competitor are driven by aspects of the systems architecture, specifically temporal contrasts in the nature of the cascade of information between modalities. For example, they argue that early fixation of phonological competitors reflects earlier activation of phonological representations in the speech-recognition system, with activation then later cascading to semantic and visual levels of processing, which in turn leads to the later increased fixation of visual and semantic competitors. In contrast, in the model proposed in the current paper, eye gaze is a continuous measure of the simultaneous integration of information activated across all three modalities. Therefore, activation of an item's phonological representation cannot influence gaze independent of currently activated visual and semantic representations.

Huettig and McQueen (2007) highlight the value of the VWP as a tool for probing finer aspects of the architecture of the cognitive system, as eye gaze offers a fine grained measure of the information activated over time. By combining this rich behavioral measure with the current model it may be possible to further examine more subtle aspects of the systems architecture that have so far proved difficult to isolate without implementation. We hope to test whether the parsimonious architecture presented in this paper is compatible with the data provided by Huettig and McQueen (2007). It remains to be seen whether such an architecture can also offer explanation for the complex time course dynamics that emerge when competitors from multiple modalities are presented simultaneously within the same display. The results of our simulations establish the applicability of the shared resource model to account for interactions between pairs of modalities. We demonstrate its ability to replicate a range of effects involving visual-semantic and visual-phonological interactions (see **Table 3**), a necessary precursor before extending to multiple interactive effects.

Within the model we present, noise is only applied to phonological input. However, in the human cognitive system, perceptual input from all modalities provides only a noisy representation of the true nature of objects in the environment. It may therefore be interesting to also extend the model to capture environmental noise in visual input. Unlike speech, visual descriptions of objects can often be improved by gathering additional information regarding its visual features over time. The literature indicates that certain groups of visual features are activated earlier than others, for example low spatial frequency information has been shown to be recruited early and rapidly by the visual system (Bar, 2003). A detailed implementation of such features of visual processing is yet to be implemented within the model. It is possible that such features may have interesting consequences for language-mediated visual attention. The model described in this paper potentially provides a means of exploring such questions.

Further applications of the model can be found in on-going experimental work that suggests that the relative influence of representational overlap in semantic, visual and phonological dimensions fluctuates over the course of child development (Mani and Huettig, in preparation). As previously discussed, model training simulates the interactions between the cognitive system and the learning environment through which the system acquires knowledge of objects in the world. Through sampling performance of the model as it moves through the training process it is possible to extract measures of its behavior on individual tasks across the course of development. It may therefore be possible, in this way, to explore the developmental story of language-mediated visual attention and provide an explicit description of the mechanism driving observed variation across development.

The model also provides scope for modeling individual differences in language-mediated visual attention observed between mature populations. In a recent study conducted by Huettig et al. (2011c), language-mediated visual attention varied as a consequence of literacy training. Their results showed that whereas a high-literate population demonstrated phonological competitor effects similar to those previously discovered (Allopenna et al., 1998), low-literates' eye gaze did not display sensitivity to phonological overlap between spoken target words and items presented in a visual display. Instead low-literates' gaze was strongly influenced by semantic relationships between items. One explanation for this difference that could be tested in the current model is whether observed differences in language-mediated visual attention between low and high literates emerge a consequence of finer grained processing of the speech signal that follows from increased literacy training (cf. Ziegler and Goswami, 2005). The modeling framework presented in this paper allows manipulation of environmental variables such as the form of representations processed and the tasks performed in the learning environment.

By manipulating such variables, it becomes possible to test theoretical explanations for these observed individual differences (see Smith et al., 2013).

As in previous H&S models, emergent properties of this style of model are dictated by multiple factors including environmental variables such as the structure of representations and the type and frequency of mappings performed, in addition to resource-related factors such as the number of units within the central resource. With so many degrees of freedom open to the modeler with which to fit H&S models to data sets, it is crucial that steps are taken to avoid simply data fitting and instead develop a model able to probe important theoretical questions (see Seidenberg and Plaut, 2006). Any assumptions made in the model development process should be justifiable with clear theoretical motivation. One effective method of model validation is to extract from a model testable non-trivial predictions. Our model of VWP effects was effective in simulating a broad range of behavior using a single set of parameters. When noise was present in the training environment, we effectively simulated processing of visual, phonological and semantic competitors and in differing situations—when targets were present or absent from the visual input to the model. Furthermore, subtle patterns of fixations over time were demonstrated by the model that were similar to behavioral data. **Figure 1** illustrated the effect of semantic competitors in behavioral data, with an emerging preference for the target, and a later, but smaller, diverging effect of near and distant semantic competitors. A similar pattern is illustrated in the model, as shown in **Figure 4**. Data-fitting to such nuanced patterns of behavior is likely to require many free parameters, and so our model's dynamics are effective in generalizing to a broad range of behavioral effects.

Connecting modalities via a central resource as in H&S does not provide the only solution for connecting the various modalities known to play a role in language-mediated visual attention. Other models are possible in which one builds in additional modalities separately. The advantage of the model presented in this paper is that the effects reported are emergent. A critical feature of the model's architecture is the amodal shared resource that intervenes between modal-specific representational systems. Such an architecture is characteristic of H&S models (Plaut, 2002; Rogers et al., 2004; Dilkina et al., 2010), and ensures that no unnecessary assumptions about how mappings are formed between distinct representations are included in the model. Furthermore, the amodal shared resource in the model appears to parsimoniously support the interactions between multiple representations that are so characteristic of complex language-processing behavior. Processing in each of the modalities described within the model is likely to involve complex hierarchical systems (see Simmons and Barsalou, 2003; McNorgan et al., 2011). However, the results of our study demonstrate that a parsimonious H&S architecture is able to capture a diverse range of effects reported in the language-mediated visual attention literature (see **Table 3**). We argue that the model presented operates at a suitable level of abstraction to act as a meaningful proxy for the cognitive system that supports language-mediated visual attention. In doing so the model provides a valuable contribution in describing the nature of the representations and processes involved in this complex multimodal behavior performed by individuals on a daily basis, and further it offers a tool through which the factors driving individual differences in language-mediated visual attention can be examined.

## REFERENCES


Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *J. Mem. Lang.* 38, 419–439. doi: 10.1006/jmla.1997.2558

Anderson, S. E., Chiu, E., Huette, S., and Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta Psychol.* 137, 181–189. doi: 10.1016/j.actpsy.2010.09.008

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *J. Cogn. Neurosci.* 15, 600–609. doi: 10.1162/089892903321662976

Barsalou, L. W., Kyle Simmons, W., Barbey, A. K., and Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends Cogn. Sci.* 7, 84–91. doi: 10.1016/S1364-6613(02)00029-3

Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge; London: MIT Press.

Coltheart, M. (2004). Are there lexicons? *Q. J. Exp. Psychol. Sec. A* 57, 1153–1171.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cogn. Psychol.* 6, 84–107. doi: 10.1016/0010-0285(74)90005-X

Cree, G. S., and McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *J. Exp. Psychol. Gen.* 132, 163–200. doi: 10.1037/0096-3445.132.2.163

Cree, G. S., McRae, K., and McNorgan, C. (1999). An attractor model of lexical conceptual processing: simulating semantic priming. *Cogn. Sci.* 23, 371–414. doi: 10.1207/s15516709cog2303_4

Dahan, D., and Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: conceptually mediated eye movements during spoken-word recognition. *Psychon. Bull. Rev.* 12, 453–459. doi: 10.3758/BF03193787

Desroches, A. S., Joanisse, M. F., and Robertson, E. K. (2006). Specific phonological impairments in dyslexia revealed by eyetracking. *Cognition* 100, B32–B42. doi: 10.1016/j.cognition.2005.09.001

Dilkina, K., McClelland, J. L., and Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cogn. Neuropsychol.* 25, 136–164. doi: 10.1080/02643290701723948

Dilkina, K., McClelland, J. L., and Plaut, D. C. (2010). Are there mental lexicons. The role of semantics in lexical decision. *Brain Res.* 1365, 66–81. doi: 10.1016/j.brainres.2010.09.057

Dove, G. (2009). Beyond perceptual symbols: a call for representational pluralism. *Cognition* 110, 412–431. doi: 10.1016/j.cognition.2008.11.016

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Goldstone, R. L., and Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition* 65, 231–262. doi: 10.1016/S0010-0277(97)00047-4

Halberda, J. (2006). Is this a dax which I see before me. Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cogn. Psychol.* 53, 310–344. doi: 10.1016/j.cogpsych.2006.04.003

Harm, M. W., and Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychol. Rev.* 111, 662–720. doi: 10.1037/0033-295X.111.3.662

Huettig, F., and Altmann, G. T. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition* 96, B23–B32. doi: 10.1016/j.cognition.2004.10.003

Huettig, F., and Altmann, G. T. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Vis. Cogn.* 15, 985–1018. doi: 10.1080/13506280601130875

Huettig, F., and McQueen, J. M. (2007). The tug of war between

phonological, semantic and shape information in language-mediated visual search. *J. Mem. Lang.* 57, 460–482. doi: 10.1016/j.jml.2007.02.001

Huettig, F., Mishra, R. K., and Olivers, C. N. (2012). Mechanisms and representations of language-mediated visual attention. *Front. Psychol.* 2:394. doi: 10.3389/fpsyg.2011.00394

Huettig, F., Olivers, C. N., and Hartsuiker, R. J. (2011a). Looking, language, and memory: bridging research from the visual world and visual search paradigms. *Acta Psychol.* 137, 138–150. doi: 10.1016/j.actpsy.2010.07.013

Huettig, F., Rommers, J., and Meyer, A. S. (2011b). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychol.* 137, 151–171. doi: 10.1016/j.actpsy.2010.11.003

Huettig, F., Singh, N., and Mishra, R. K. (2011c). Language-mediated visual orienting behavior in low and high literates. *Front. Psychol.* 2:285. doi: 10.3389/fpsyg.2011.00285

Huettig, F., Quinlan, P. T., McDonald, S. A., and Altmann, G. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychol.* 121, 65–80. doi: 10.1016/j.actpsy.2005.06.002

Kello, C. T., and Plaut, D. C. (2000). Strategic control in word reading: evidence from speeded responding in the tempo-naming task. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 719–750. doi: 10.1037/0278-7393.26.3.719

Kintsch, W. (2008). "Symbol systems and perceptual representations," in *Symbols and Embodiment: Debates on Meaning and Cognition*, eds M. de Vega, A. Glenberg, and A. Graesser (New York, NY: Oxford University Press), 145–163. doi: 10.1093/acprof:oso/9780199217274.003.0008

Kukona, A., and Tabor, W. (2011). Impulse processing: a dynamical systems model of incremental eye movements in the visual world paradigm. *Cogn. Sci.* 35, 1009–1051. doi: 10.1111/j.1551-6709.2011.01180.x

Lambon Ralph, M. A., and Patterson, K. (2008). Generalization and differentiation in semantic memory. *Ann. N.Y. Acad. Sci.* 1124, 61–76. doi: 10.1196/annals.1440.006

Lambon Ralph, M. A., Sage, K., Jones, R. W., and Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2717–2722. doi: 10.1073/pnas.0907307107

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Front. Psychol.* 3:54. doi: 10.3389/fpsyg.2012.00054

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., and Dahan, D. (2003). The time course of spoken word learning and recognition: studies with artificial lexicons. *J. Exp. Psychol. Gen.* 132, 202–227. doi: 10.1037/0096-3445.132.2.202

Mahon, B. Z., and Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *J. Physiol.* 102, 59–70.

Mani, N., Johnson, E., McQueen, J. M., and Huettig, F. (2013). How yellow is your banana. Toddlers' language-mediated visual search in referent-present tasks. *Dev. Psychol.* 49, 1036–1044. doi: 10.1037/a0029382

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua* 92, 199–227. doi: 10.1016/0024-3841(94)90342-5

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition* 25, 71–102. doi: 10.1016/0010-0277(87)90005-9

Marslen-Wilson, W. (1993). "Issues of process and representation in lexical access," in *Cognitive Models of Speech Processing: The Second Sperlonga Meeting*, (Hillsdale, NJ: Erlbaum), 187–210.

Mayberry, M. R., Crocker, M. W., and Knoeferle, P. (2009). Learning to attend: a connectionist model of situated language comprehension. *Cogn. Sci.* 33, 449–496. doi: 10.1111/j.1551-6709.2009.01019.x

McClelland, J. L., and Elman, J. L. (1986). The trace model of speech perception. *Cogn. Psychol.* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

McNorgan, C., Reid, J., and McRae, K. (2011). Integrating conceptual knowledge within and across representational modalities. *Cognition* 118, 211–233. doi: 10.1016/j.cognition.2010.10.017

McQueen, J. M., and Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *J. Acoust. Soc. Am.* 131, 509–517. doi: 10.1121/1.3664087

McQueen, J. M., and Viebahn, M. C. (2007). Tracking recognition of spoken words by tracking looks to printed words. *Q. J. Exp. Psychol.* 60, 661–671. doi: 10.1080/17470210601183890

Mirman, D., and Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Mem. Cogn.* 37, 1026–1039. doi: 10.3758/MC.37.7.1026

Monaghan, P., and Mattock, K. (2012). Integrating constraints for learning word-referent mappings. *Cognition* 123, 133–143. doi: 10.1016/j.cognition.2011.12.010

Monaghan, P., and Nazir, T. (2009). "Modelling sensory integration and embodied cognition in a model of word recognition," in *Connectionist Models of Behaviour and Cognition II*, eds J. Mayor, N. Ruh, and K. Plunkett (Singapore: World Scientific), 337–348.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition* 52, 189–234. doi: 10.1037/0033-295X.115.2.357.

Plaut, D. C. (2002). Graded modality-specific specialisation in semantics: a computational account of optic aphasia. *Cogn. Neuropsychol.* 19, 603–639. doi: 10.1080/02643290244000112

Pobric, G., Jefferies, E., and Ralph, M. A. L. (2007). Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rTMS in normal participants. *Proc. Natl. Acad. Sci. U.S.A.* 104, 20137–20141. doi: 10.1073/pnas.0707383104

Prinz, J. J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis.* Cambridge, MA: MIT Press.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., et al. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychol. Rev.* 111, 205–234. doi: 10.1037/0033-295X.111.1.205

Seidenberg, M. S., and Plaut, D. C. (2006). "Progress in understanding word reading: data fitting versus theory building," in *From Inkmarks to Ideas: Current Issues in Lexical Processing*, ed S. Andrews (New York, NY: Taylor & Francis), 25–49.

Simmons, W. K., and Barsalou, L. W. (2003). The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cogn. Neuropsychol.* 20, 451–486. doi: 10.1080/02643290342000032

Smith, A. C., Monaghan, P., and Huettig, F. (2013). "Modelling the effects of formal literacy training on language mediated visual attention," in *Proceedings of the 35th Annual Conference of the Cognitive Science Society (2013)*, eds M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Austin, TX: Cognitive Science Society).

Spivey, M. (2008). *The Continuity of Mind.* Vol. 40. New York, NY: Oxford University Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

Vandenberghe, R., Price, C., Wise, R., Josephs, O., and Frackowiak, R. S. (1996). Functional anatomy of a common semantic system for words and pictures. *Nature* 383, 254–256. doi: 10.1038/383254a0

Yee, E., and Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 1–14. doi: 10.1037/0278-7393.32.1.1

Yoon, E. Y., Heinke, D., and Humphreys, G. W. (2002). Modelling direct perceptual constraints on action selection: the naming and action model (NAM). *Vis. Cogn.* 9, 615–661. doi: 10.1080/13506280143000601

Ziegler, J. C., and Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychol. Bull.* 131, 3–29. doi: 10.1037/0033-2909.131.1.3

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 01 May 2013; accepted: 26 July 2013; published online: 16 August 2013.*
*Citation: Smith AC, Monaghan P and Huettig F (2013) An amodal shared resource model of language-mediated visual attention. Front. Psychol. **4**:528. doi: 10.3389/fpsyg.2013.00528*
*This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 20 August 2013
doi: 10.3389/fpsyg.2013.00503

# Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review

***James L. McClelland****

*Department of Psychology and Center for Mind, Brain, and Computation, Stanford University, Stanford, CA, USA*

***Edited by:***
*Julien Mayor, University of Geneva, Switzerland*

***Reviewed by:***
*Jonathan Grainger, Laboratoire de Psychologie Cognitive, CNRS, France*
*Naomi Feldman, University of Maryland, USA*

****Correspondence:***
*James L. McClelland, Department of Psychology and Center for Mind, Brain, and Computation, Stanford University, Jordan Hall, 450 Serra Mall, Stanford, CA 94305, USA*
*e-mail: mcclelland@stanford.edu*

This article seeks to establish a rapprochement between explicitly Bayesian models of contextual effects in perception and neural network models of such effects, particularly the connectionist interactive activation (IA) model of perception. The article is in part an historical review and in part a tutorial, reviewing the probabilistic Bayesian approach to understanding perception and how it may be shaped by context, and also reviewing ideas about how such probabilistic computations may be carried out in neural networks, focusing on the role of context in interactive neural networks, in which both bottom-up and top-down signals affect the interpretation of sensory inputs. It is pointed out that connectionist units that use the logistic or softmax activation functions can exactly compute Bayesian posterior probabilities when the bias terms and connection weights affecting such units are set to the logarithms of appropriate probabilistic quantities. Bayesian concepts such the prior, likelihood, (joint and marginal) posterior, probability matching and maximizing, and calculating vs. sampling from the posterior are all reviewed and linked to neural network computations. Probabilistic and neural network models are explicitly linked to the concept of a probabilistic generative model that describes the relationship between the underlying target of perception (e.g., the word intended by a speaker or other source of sensory stimuli) and the sensory input that reaches the perceiver for use in inferring the underlying target. It is shown how a new version of the IA model called the multinomial interactive activation (MIA) model can sample correctly from the joint posterior of a proposed generative model for perception of letters in words, indicating that interactive processing is fully consistent with principled probabilistic computation. Ways in which these computations might be realized in real neural systems are also considered.

**Keywords: interactive activation, context in perception, neural networks, probabilistic computation, generative models**

## INTRODUCTION

For well over a century (Huey, 1908), there has been an interest in understanding how context affects the perception of the spoken and written word. During the cognitive revolution of the 1950's and 60's, George Miller and others contributed important findings (e.g., Miller et al., 1951; Tulving et al., 1964) showing that context facilitated word recognition, and these findings were captured in the classical Logogen model (Morton, 1969). Reicher (1969) introduced the striking word superiority effect, demonstrating that letters are perceived more accurately in words than in isolation, and the phenomenon received extensive investigation in the early 1970's (e.g., Wheeler, 1970; Aderman and Smith, 1971; Johnston and McClelland, 1973, 1974; Thompson and Massaro, 1973). Rumelhart and Siple (1974) and Massaro (1979) offered models of context effects in letter perception, and Rumelhart (1977) laid out how such a model might be extended to address a broader range of contextual effects, including syntactic and semantic effects and effects of non-linguistic context on word identification and semantic interpretation.

The models mentioned above were all either explicitly probabilistic models or could be linked easily with probabilistic, Bayesian computations. But then a funny thing happened. On the one hand, Pearl (1982) offered a systematic Baysian framework that unified the earlier models into an general algorithm (subject to some limitations) for probabilistic Bayesian inference across multiple mutually interdependent levels of interpretation (feature, letter, word, syntactic/semantic interpretation). On the other hand, Rumelhart and I diverged from the path of probabilistic Bayesian models, proposing a model of context effects in letter perception (McClelland and Rumelhart, 1981) that did not refer explicitly to probabilistic Bayesian ideas, drawing inspiration, instead, from models of neural activation (Grossberg, 1978). In fact, as Massaro (1989) pointed out, our interactive activation (IA) model actually failed to account for aspects of data that were easily captured by the earlier models and by simple Bayesian considerations.

A considerable debate ensued, one in which it seemed for a while as though there might be an intrinsic conflict between probabilistic Bayesian models on the one hand and not just connectionist models but *any* model involving bi-directional propagation of influences on the other. Pearl's work clearly provided an interactive method of carrying out provably valid probabilistic

Bayesian computations, but Massaro (1989); Massaro and Cohen (1991) as well as Norris and co-authors (Norris et al., 2000) nevertheless argued that bi-directional propagation of information would lead to violations of correct probabilistic Bayesian inference. While I and my collaborators (McClelland, 1991; Movellan and McClelland, 2001; McClelland et al., 2006) were able to address many of the specific criticisms, the notion that distortion of valid inference is intrinsic to bi-directional propagation of information has persisted (Norris and McQueen, 2008).

In part, this debate reflects a simple failure on the part of psychologists (including myself!) to keep up with developments in computer science and related disciplines, and in part, it reflects an enthusiasm represented by early neural network models to draw inspiration from putative principles of brain function rather than principles of probabilistic inference. In any case, the purpose of the current article to establish a reconcilliation. Specifically, I seek to reassure those who stand firm for principled Bayesian models and those who seek inspiration from principles of brain-like processing that both sides can be happy at the same time.

The path I will take toward furthering this rapprochement will begin by introducing basic principles of probabilistic Bayesian inference and then indicating how these principles can be instantiated in models that also adopt principles of brain-like processing. The presentation is in part tutorial and in part historical, and is intended to help put experimentally oriented cognitive scientists, neural network modelers, and proponents of probabilistic Bayesian computation on the same page with respect to the relationship between models of perception, neural networks, and Bayesian inference.

Many of the concepts that will be reviewed are instantiated in a new version of the IA model of letter perception (McClelland and Rumelhart, 1981) called the multinomial interactive activation (MIA) model (Khaitan and McClelland, 2010; Mirman et al., in press), and that model will be used as a vehicle for discussion of these issues. The MIA model (like the IA model before it) can be viewed as a simplified model of the process of inferring the identities of objects in the external world (in this case, words and the letters of which these words are composed) from noisy visual input, and models based on the IA model and related interactive activation and competition networks (McClelland, 1981) are widespread in psychological research on topics ranging from written and spoken word perception (Elman and McClelland, 1988; Grainger and Jacobs, 1996), face perception (Burton et al., 1990), and memory retrieval (Kumaran and McClelland, 2012) to construal of personality (Freeman and Ambady, 2011). The development here will connect the intuitive principles of contextual influences on perceptual identification that were embodied in the original IA model with Bayesian ideas, showing how the new variant of the original model (the MIA model) provides a system for principled probabilistic inference similar to that envisioned in a precursor to the IA model by Rumelhart (1977) and systematized by Pearl (1982). The ideas draw heavily on the original framing of the Boltzmann Machine (Hinton and Sejnowski, 1983). They are related to ideas presented by Lee and Mumford (2003) and Dean (2005) that point out connections between Bayesian computational frameworks and real neural networks in the brain, and share several of the ideas underlying deep belief networks (Hinton and Salakhutdinov, 2006), which are, similarly, models of perceptual inference.

Taken together, the ideas we will develop provide a bridge between neurophysiological ideas and cognitive theories, and between probabilistic models of cognition and process-oriented connectionist or parallel-distributed processing models. Thus, this tutorial may prove useful as an introduction for those interested in understanding more about the relationship between a simple form of Bayesian computation and both real and artificial neural networks. While the specific examples are all drawn from perception of letters in words, the possible applications include many other perceptual problems as well as the more general problem of inferring underlying causes from observed evidence.

We begin by presenting Bayes' formula as a tool for inferring the posterior probability that some hypothesis is true, given prior knowledge of certain probabilistic quantities and some evidence[1]. This part of the presentation starts with the case of two mutually exclusive and exhaustive hypotheses and a single source of evidence, and shows how Bayes' formula follows from the definition of conditional probability. We then extend the formula to cover cases involving an arbitrary number of mutually exclusive and exhaustive hypotheses and to cases involving more than one element of evidence, introducing the concept of conditional independence. We then develop the idea of a generative model within which the quantities needed to infer posterior probabilities can be seen as representing parameters of a causal process that generates the inputs to a perceptual system.

We next consider how Bayesian inference can be carried out by a neural network. In particular, we observe how the softmax and logistic activation functions often used in neural networks can produce outputs corresponding to posterior probabilities, provided that the biases and connection weights used in producing these outputs represent the logarithms of appropriate probabilistic quantities.

With the above background, we then describe how bottom-up and top-down information can be combined in computing posterior probabilities of letters presented in context, in accordance with Bayes' formula and the generative model assumed to underlie the perceptual inputs to the MIA model. We describe three procedures by which such posteriors (or samples from them) can be computed—one that is completely non-interactive [appearing to accord with the proposals of Massaro (1989) and elsewhere, and of Norris and McQueen (2008)], and two that involve bi-directional propagation of information, as in the original IA model (McClelland and Rumelhart, 1981). One of these procedures computes these posteriors exactly, and relates to proposals in Rumelhart (1977) and Pearl (1982). The other samples from the posterior, using Gibbs sampling as in the Boltzmann machine (Hinton and Sejnowski, 1983); this is the approach taken in the MIA model. The connection to deep belief networks is considered briefly at the end of the article.

As can be seen from the citations above, the key ideas reviewed here have been in circulation for about 30 years. These ideas establish an intimate connection between the computations performed

[1]Often the word *data* is used instead of *evidence*. Some writers use evidence to refer to quite a different concept.

by neural networks and computations necessary to carry out correct probabilistic inference. Unfortunately, to my knowledge there has not been extensive recognition of these connections, at last among many researchers working in the psychological and cognitive science disciplines. The presentation draws on an earlier paper with similar goals (McClelland, 1998) and is intended to help provide an intuitive understanding of some of the relevant concepts involved, and of the reasons why certain things are true, without relying on formal proofs.

## USING BAYES' FORMULA TO INFER POSTERIOR PROBABILITIES

We begin by reviewing the canonical version of Bayes' formula, expressing the posterior probability that one of two mutually exclusive and exhaustive hypotheses is true given some evidence $e$ in terms of other quantities which we will shortly define:

$$p(h_i|e) = \frac{p(h_i)p(e|h_i)}{p(h_1)p(e|h_1) + p(h_2)p(e|h_2)} \quad (1)$$

In this expression, $p(h_i)$ corresponds to the prior probability that hypothesis $i$ is true, where $h_i$ could be hypothesis 1 or hypothesis 2. $p(e|h_i)$ corresponds to the probability of the evidence given that hypothesis $i$ is true, and $p(h_i|e)$ corresponds to the posterior probability of hypothesis $i$ given the evidence. The expression is often called "Bayes' law," or "Bayes' rule," although some use "Bayes' rule" for a formulation that expresses the ratio of the posterior probability of $h_1$ to $h_2$. Bayes' rule in that form is easily derived from Bayes' formula and *vice versa*. The formula is also sometimes described as "Bayes' Theorem," but we will use that phrase to refer to the proof of the validity of the formula, rather than the formula itself.

As an example [from the Wikipedia entry on (Bayes' theorem, n.d.)], suppose a friend of yours meets a person with long hair. What is the probability that this person is a woman? Our two possible hypotheses here are that the person is a woman or that the person is a man. We treat them as mutually exclusive and exhaustive—that is, a person must be either a man or a woman; there are no other possibilities, and the person cannot be both a man and a woman at the same time. The evidence $e$ is that the person has long hair.

Bayes' formula allows us to calculate the answer to this question, as long as some additional relevant facts are known. First, we need to know the overall probability that a person your friend might meet is a woman. We could call this probability $p(h_1)$, but to aid maintaining contact with the example, we will call it $p(W)$. Since we have assumed that the only other possibility is that the person is a man, the probability that the person is not a woman $p(\bar{W})$ is equal to the probability that the person is a man, $p(M)$. From this it follows that $p(W) + p(M) = 1$, and that $p(M) = p(\bar{W}) = 1 - p(W)$.

The quantity $p(W)$ represents to a given or assumed quantity corresponding to the overall probability that a person your friend might meet is a woman. This quantity is often called the *prior*, a usage that makes sense if our goal is to use evidence to update our beliefs about the probability that a person your friend might meet is a woman once we observe the particular person's gender. Here, we are just using this quantity as a premise in an inference process. Nevertheless, writers often use the term prior when describing such terms, and we will often do so here. Another phrase that is sometimes used is *base rate*. Humans often neglect base rates in carrying out probabilistic inference when given probabilistic information in explicit form. When the base rate is low, this can lead to an over-estimate of the posterior probability.

It might be noted that there could be uncertainty about the prior or base rate. This is certainly true, and indeed, the question that the Reverend Bayes was primarily interested in was how to use evidence to update one's beliefs about such probabilities. This is a rich and important topic, but it is not the one we are examining here. Instead we are considering the simpler problem of using a set of known probabilistic quantities to infer another probabilistic quantity, the probability that the hypothesis is true in a particular instance, given some evidence.

In addition to knowledge of the prior probability of the hypotheses, $p(h_1)$ and $p(h_2)$, we also must know the probability of observing the evidence when each hypothesis it true. In our example, we need to know the probability of long hair when the person is a woman (for our example, $p(L|W)$ or more generally $p(e|h_1)$), and also the probability of long hair when the person is a man ($p(L|M)$ or more generally, $p(e|h_2)$). Here, too, there could be considerable uncertainty. However, as with the prior, we will treat these as quantities that are known, and proceed from there.

Using these quantities, we can plug them into Equation 1 to calculate $p(W|L)$, the probability that the person your friend met is a woman given that the person had long hair. The expression below replaces the abstract variables $h_1$ and $h_2$ from Equation 1 with $W$ and $M$, and replaces the abstract variable $e$ with the $L$ for long hair, to connect the various quantities in the expression to the relevant conceptual quantities in the example:

$$p(W|L) = \frac{p(W)p(L|W)}{p(W)p(L|W) + p(M)p(L|M)}$$

Let's plug in some actual numbers. If the overall probability of your friend meeting a woman, $p(W)$, is 0.5; the probability of a woman having long hair $p(L|W)$ is 0.8; and the probability of a man having long hair, $p(L|M)$, is 0.3, then (relying on $p(M) = 1 - p(W) = 0.5$), we obtain:

$$p(W|L) = \frac{0.5 * 0.8}{0.5 * 0.8 + 0.5 * 0.3} = \frac{0.8}{0.8 + 0.3} = \frac{0.8}{1.1} = 0.727$$

As an exercise, the reader can explore what happens to the result when one of the relevant quantities changes. What if $p(L|M)$ goes down to 0.01? In a world where few men have long hair we get a much stronger conclusion. On the other hand, what if $p(L|M) = 0.8$? You should see that in this case we learn nothing about the person's gender from knowing the person has long hair. Now, what about the prior or base rate, $P(W)$? We have assumed that a person your friend might meet is equally likely to be a woman or a man, but what if instead $p(W)$ is only 0.1—this might happen, for example, if the people your friend meets are all computer science majors. Using our initial values for the likelihoods $p(L|W) = 0.8$ and $p(L|M) = 0.3$, you should find that the posterior probability that the person is a woman is less than 0.3. If you neglected the base rate, you might overestimate this probability.

As a second exercise, the reader should be able to calculate $p(W|S)$, the probability that a person your friend met is a woman given that the person had *short* hair, given specific values for $p(L|W)$, $p(L|M)$ and $p(W)$. Use 0.8, 0.3, and 0.5 for these quantities. What gender should we guess to maximize the probability of being correct if we were told that a person your friend met had short hair? Assume for this example that each person either has short hair or long hair—that is, that short and long are mutually exclusive and exhaustive alternatives. As before, also assume that male and female are mutually exclusive and exhaustive alternatives.

Bayes' formula can easily be applied to cases in which the two hypotheses under consideration are the hypothesis that some proposition is true and the hypothesis that the proposition is false. For example, we might want to determine whether a person is French or not. In this case, our hypotheses could be 'Person X is French' and 'Person X is not French,' where no specific alternative hypothesis is specified. Here it is natural to use $h$ for the positive case and $\bar{h}$ for the negative case, and to rewrite the formula as:

$$p(h|e) = \frac{p(h)p(e|h)}{p(h)p(e|h) + p(\bar{h})p(e|\bar{h})}$$

Given that $h$ and $\bar{h}$ are assumed to be mutually exclusive and exhaustive, $p(\bar{h}) = 1 - p(h)$, so we can also write our formula as:

$$p(h|e) = \frac{p(h)p(e|h)}{p(h)p(e|h) + (1 - p(h))p(e|\bar{h})} \quad (2)$$

It is also worth noting that the posterior probabilities sum to one: $p(h|e) + p(\bar{h}|e) = 1$, so $p(\bar{h}|e) = 1 - p(h|e)$. Thus, the evidence simultaneously informs us about the posterior probability that $h$ is true, and that $h$ is false.

***Remark*:** Clearly, Bayes' formula only gives valid results if the quantities that go into the calculation are accurate. It would likely be wrong to assume that human perception always relies on the correct values of these quantities. One could propose that human perceivers rely on estimates of such quantities, and that these may differ from their actual values. A further point is that an experimenter might generate inputs according to a protocol that is not fully consistent with the knowledge perceivers rely on to make perceptual inferences. In that case, if the estimates perceivers rely on are not altered to match the protocol used in the experiment, the inferences could be invalid, and therefore not optimal under the conditions of the experiment. For example, a perceiver in a word identification experiment might rely on estimates of each word's probability of occurrence based on its frequency of occurrence in past experience. However, an experimenter might choose words from a word list without regard to their frequency. Under these conditions, use of a word's frequency to represent its probability of occurrence would be invalid. Many perceptual "biases" or "illusions" can be explained as resulting from the use of estimates of probabilistic quantities that may be valid (or approximately valid) in the real world, but are not valid within the context of the experiment. If such knowledge were wired into the connections among neurons in a perceiver's perceptual system, as it is assumed to be in the IA model, it might not be easily discarded and replaced with other values.

### *Decision policies*

So far, we have shown how to calculate a posterior probability, but we have not discussed what one might actually do with it. In many situations, we may simply want to take note of the posterior probability—in the case of our first example above, we might not wish to reach a definite conclusion, since the evidence is far from conclusive. However, often a choice between the alternatives is required. There are two possibilities that are often considered: one policy tries to pick the best response, that is, the one that maximizes the probability of being correct, while the other generates responses probabilistically, according to the posterior probability.

The first policy is called *maximizing*. This policy amounts to choosing the alternative with the largest posterior probability. Formally, we could write:

$$Choice = \text{argmax}(p(h_1|e), p(h_2|e))$$

where the **argmax** function returns the index of the hypothesis with the largest posterior probability. In our example, with the priors $p(W) = 0.5$, $p(L|W) = 0.8$ and $p(L|M) = 0.3$, we calculated that $p(W|L) = 0.727$ and it follows that $p(M|L) = 0.273$. Following this policy, then, we would conclude that the person is a woman given that the person has long hair.

The second policy is called *probability matching* or just *matching*. Under this policy, decision makers' choices would vary from trial to trial with the same evidence, but would occur with a probability that matches the posterior probability. Formally, we would write this as:

$$p(Choice = i) = p(h_i|e)$$

One of these two policies is better than the other, in the sense that one maximizes the probability of choosing the correct answer. If you would win a dollar for guessing right and loose a dollar for guessing wrong, which of these policies should you chose? Surprisingly, in many cases, the behavior of humans and other animals appears closer to matching rather than maximizing, but there are situations in which people clearly do maximize (Green et al., 2010). There are worse policies than matching. One such policy sometimes used in explicit outcome guessing tasks by children around age five is to alternate choices from one trial to the next, regardless of the probability of each of the two outcomes, and even when the trial sequence is completely random (Derks and Paclisanu, 1967).

### *BAYES' theorem: Bayes' formula follows from the definition of conditional probability*

So far, we have used Bayes' formula without considering why it is true. Here, we will show that the validity of the formula follows from the definition of conditional probability. We have already used the concept of conditional probability. Here we will review its definition and then use it to derive Bayes' formula.

The conditional probability of some event $a$ given some other event $b$, written $p(a|b)$, is defined as the ratio of the probability of both $a$ and $b$, $p(a\&b)$ to the probability of $b$, $p(b)$:

$$p(a|b) = \frac{p(a\&b)}{p(b)}$$

The definition can be read as defining conditional probability $p(a|b)$ as the proportion of the times when $b$ occurs that $a$ also occurs. Let's relate this to our case, letting $e$ correspond to $a$ and $h$ correspond to $b$:

$$p(e|h) = \frac{p(e\&h)}{p(h)} \quad (3)$$

In our case, if 50% of the people your friend might meet are women, and 40% of the people your friend might meet are women with long hair, then the probability of long hair given that the person is a woman—or equivalently, the proportion of women who have long hair—would be $0.4/0.5 = 0.8$, the value we already used in our example.

Now we can also use the definition of conditional probability to express $p(h|e)$, letting $e$ correspond to $b$ and $h$ correspond to $a$:

$$p(h|e) = \frac{p(e\&h)}{p(e)} \quad (4)$$

Bayes' formula can now be derived from the fact that $p(e\&h)$ occurs in the definition of both $p(e|h)$ and $p(h|e)$. To derive it, we multiply both sides of Equation 3 by $p(h)$ to obtain:

$$p(e\&h) = p(h)p(e|h)$$

For our example, this corresponds to the fact that the proportion of people who have long hair and are women is equal to the proportion of all people who are women, times the proportion of women who have long hair.

We can now replace $p(e\&h)$ in Equation 4 with $p(h)p(e|h)$ to obtain:

$$p(h|e) = \frac{p(h)p(e|h)}{p(e)}$$

This can be stated: the probability of some hypothesis $h$ being true given some evidence $e$ is equal to the prior probability of the hypothesis, $p(h)$, times the probability of the evidence, given the hypothesis, divided by the overall probability of the evidence $p(e)$.

It remains only to note that the denominator, the probability of the evidence $p(e)$, is equal to the probability of the evidence occurring when the hypothesis is true plus the probability of the evidence occurring when the hypothesis is false, $p(e\&h) + p(e\&\bar{h})$. That is, the total probability of situations in which $e$ is true is the sum of the probabilities of two situations, one in which $e$ is true and the hypothesis $h$ is also true, and another in which $e$ is true and the hypothesis is false. This exhausts the cases in which $e$ is present, given that $h$ must either be true or not. Using the fact that $p(a\&b) = p(b)p(a|b)$ twice more, applying it to both $p(e\&h)$ and to $p(e\&\bar{h})$, we finally obtain:

$$p(h|e) = \frac{p(h)p(e|h)}{p(h)p(e|h) + p(\bar{h})p(e|\bar{h})}$$

and from $p(\bar{h}) = 1 - p(h)$, we can then obtain Equation 2. Of course the same all works out for cases in which we have two mutually exclusive and exhaustive hypotheses called $h_1$ and $h_2$ as in the version shown in Equation 1, as well.

**Figure 1** gives a graphical representation of the posterior probability of a hypothesis constructed by partitioning a square with sides of length 1. We use the horizontal dimension to partition the square into two parts by drawing a vertical line at $x = p(W)$, so that the area to the left of the line corresponds to the overall probability that a person your friend might meet would be a woman and the remaining area corresponds to the probability that the person your friend might meet would be a man. Restating, the areas of these two parts correspond to $p(W)$ and $p(M)$, respectively. Then, we partition the region corresponding to women into two parts along the vertical axis at the point $y = p(L|W)$. This divides the total probability that the person is a woman into two parts, one corresponding to the probability that the person is a woman and has long hair, and one corresponding to the probability that the person is a woman and does not have long hair. Likewise, we partition the region corresponding to men into two parts along the vertical axis at the point $y = p(L|M)$. This gives us two more rectangles, one whose area corresponds to the probability that the person is a man and has long hair, and the other corresponding to the probability that the person is a man and does not have long hair. The area of each resulting rectangle is a joint probability as well as the product of a prior and a conditional probability. The posterior probability $p(W|L)$ is the ratio of the area of the rectangle corresponding to women with long hair to the area corresponding to all persons with long hair, which in turn corresponds to the sum of the areas of the two shaded rectangles.

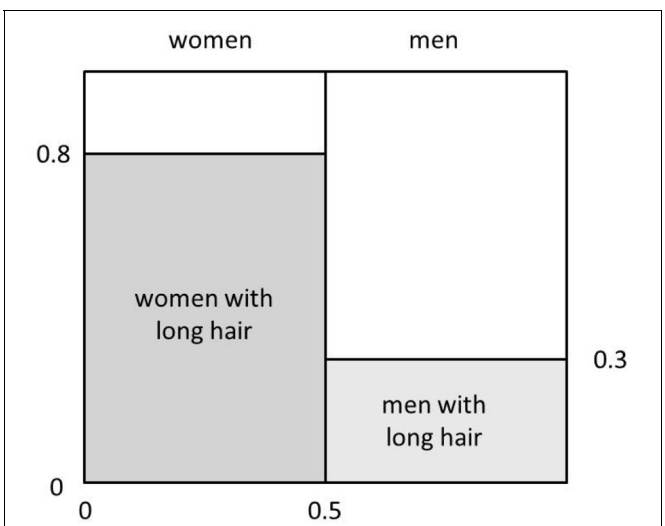



**FIGURE 1 | Graphical depiction of posterior probability based on relative area.**

To fix your understanding of these ideas, you could draw an approximate version of this figure for the case in which (i) the overall probability that a person your friend might meet is a woman is 0.25; (ii) the probability of a woman having long hair is 0.75; and (iii) and the probability of a man having long hair is 0.25. Inspecting the relevant subrectangles within the unit rectangle, you should be able to estimate the probability that the person your friend meets is a woman, given that the person has long hair. You would do this by noting the area corresponding to the probability of being a woman and having long hair, and comparing that to the area corresponding to the probability of being a man and having long hair. Given that these areas are about equal, what is the probability that a person with long hair is a woman in this case?

### *Multiple alternative hypotheses*

We have thus far considered cases in which there are only two possible hypotheses, for example, either the person my friend met was a woman or the person was a man. Now let us suppose we have many alternative hypotheses $\{h_i\}$, and we are trying to determine the posterior probability of each given some evidence $e$. One example arises if we are trying to determine the identity of a letter given one of its features. For example, in the font used by Rumelhart and Siple (1974), and in the MIA model, one of the features (which we will call $F_{ht}$) is a horizontal line segment at the top of a letter-feature block (See **Figure 2**). Some letters have this feature, and others do not. For example, the letter $T$ has it and the letter $U$ does not. Treating these statements as absolutes, we could state $p(F_{ht}|T) = 1$ and $p(F_{ht}|U) = 0$. However, let us allow for the possibility of error, so that with a small probability, say 0.05, feature values will be registered incorrectly. Then $p(F_{ht}|T) = 0.95$ and $p(F_{ht}|U) = 0.05$. Now, suppose we want to calculate $p(T|F_{ht})$. For each letter, $l_i$ we would need to know $p(F_{ht}|l_i)$ and we would also need to know the prior probability of occurrence of each letter as well. Given this information, the overall formula for the posterior probability now becomes:

$$p(T|F_{ht}) = \frac{p(T)p(F_{ht}|T)}{\sum_{i'} p(l_{i'})p(F_{ht}|l_{i'})}$$

Note that the summation[2] in the denominator runs over all possible letters, including $T$. In general, the probability that a particular hypothesis $h_i$ is correct given a specific element of evidence $e$ can be written:

$$p(h_i|e) = \frac{p(h_i)p(e|h_i)}{\sum_{i'} p(h_{i'})p(e|h_{i'})}$$

The indexing scheme is potentially confusing: Here and elsewhere, we use a bare single letter such as $i$ to index a specific item or hypothesis of interest and a primed version of the same letter such as $i'$ to index all of the items or hypotheses, including $i$.

It is useful at this point to introduce the notion of a *multinomial random variable*, defined as a random variable that can take any one of $n$ discrete values, such as letter identities. This generalizes the notion of a binary random variable, which is one that can take either of two possible values (such as *true* or *false* or *man* or *woman*). We can think of the identity of a given letter, for example, as a multinomial random variable having one of 26 possible values. The name of the multinomial interactive activation model is based on the idea that (in letter perception) the task of the perceiver is to infer the correct values of several such multinomial variables—one for the identity of each of the four letters in a letter string, and one for the identity of the visually presented word—from visual input. For now, we are working with the simpler case of attempting to set the value of a single multinomial variable corresponding to a single letter identity.

The prior associated with a multinomial random variable is the vector of prior probablities $p(h_i)$. Under the assumption that the hypotheses are mutually exclusive and exhaustive, the sum of the $p(h_i)$ should be equal to 1. In the specific case of possible letter identities, given that there are 26 letters, there are only 25 independent letter probabilities, since the probability of the last one must be equal to 1 minus the sum of the probabilities of all of the others. In general, if there are $N$ mutually exclusive possibilities, there are only $N - 1$ degrees of freedom in the values of their prior probabilities[3].

**FIGURE 2 | The line segments used in the Rumelhart and Siple font and the letters composed from these segments.** From Rumelhart and Siple (1974). Reprinted with permission.

[2]By convention, we use $\Sigma_s$ to refer to a sum of terms indexed by a subscript $s$ and we use $\Pi_s$ to refer to a product of terms indexed by $s$ (in the equation here the subscript is $i'$ for consistency with later usage as explained below). A summation or product applies to all of the factors multiplied together following the summation symbol. Thus $\Sigma_i a_i b_i c_i$ is equivalent to $\Sigma_i (a_i b_i c_i)$ and $\Sigma_i a_i \Pi_j b_{ij} c_{ij}$ is equivalent to $\Sigma_i (a_i \Pi_j (b_{ij} c_{ij}))$. Unless explicitly indicated with parentheses, summation does not extend across a plus or minus sign. Thus $\Sigma_i a_i + b$ is not the same as $\Sigma_i (a_i + b)$.

[3]It is worth noting that, if one tabulated counts of occurrences of letters in a randomly chosen book, the counts could be considered to be independent quantities. In this case, the total count of letters in the book would be the sum of the counts for all of the letters. If we were to be given this total count, we could infer the count for any given letter from the counts for the others letters and the total count. Either way, there are $N + 1$ quantities (each count and the sum) but only $N$ independent quantities.

Even when there are multiple possibilities, we note that if only one of these hypotheses is of interest—when, say, we are interested in knowing whether a given letter is a $T$ or not—all of the other possibilities can be lumped together and we have:

$$p(T|F_{ht}) = \frac{p(F_{ht}|T)p(T)}{p(F_{ht}|T)p(T) + \sum_{i' \neq T} p(F_{ht}|l_{i'})p(l_{i'})}$$

where the summation in the denominator runs over all possible letters other than T of terms corresponding to the product of the prior and the likelihood. This is a generalization of Equation 2, previously given, with $\sum_{i' \neq T} p(F_{ht}|l_{i'})p(l_{i'})$ playing the role of $p(F_{ht}|\bar{T})p(\bar{T})$[4].

It is also worth noting that in some situations, we may want to include the possibility that the observed input arose from some unknown cause, outside of a specifically enumerated set. For example, some feature arrays that appear in a letter perception experiment might have been generated from something other than one of the known letters. We can include this possibility as an additional hypothesis, if we also provide the probability that the feature value arises from this other cause. In this case the sum of the probabilities of the enumerated causes is less than one, with the other causes consuming the remainder of the total probability. Then we can write Bayes Formula as:

$$p(h_i|e) = \frac{p(h_i)p(e|h_i)}{\sum_{i'} p(h_{i'})p(e|h_{i'}) + p(o)p(e|o)}$$

where $p(o)$ is the prior probability for all other causes and $p(e|o)$ is the probability of the evidence arising from any of these other causes. In psychological models, e.g., the Logogen model of word recognition (Morton, 1969), or the generalized context model of categorization (Nosofsky, 1984), the elements of the expression $p(o)p(e|o)$ are not separately estimated, and are lumped together in a constant.

#### *Multiple elements of evidence and conditional independence*

In general, when we are attempting to recognize letters or other things, there may be more than one element of evidence (e.g., more than one feature) at a time. How can we deal with such situations? A first step is to generalize Bayes' formula by using a likelihood term that encompasses all of the evidence. For example, we might have evidence that there is a horizontal feature across the top of a feature array and a vertical segment down the middle. We could then make use of expressions such as $p(F_{ht}\&F_{vm}|T)$ to represent the probability of observing both of these features, given that the letter in question is $T$.

A problem that arises here is that the number of possible combinations of elements of evidence can grow large very quickly, and it becomes intractable to assume that a perceiver knows and represents all of these probabilities. Luckily, there is a condition under which the computation of the values of such expressions becomes very simple. This condition is known as *conditional independence*, which can be defined for two or more events with respect to some other, conditioning event. For two events, conditional independence is defined as follows:

***Definition of Conditional Independence.*** Elements of evidence $e_1$ and $e_2$ are conditionally independent given condition $c$ if the probability of both pieces of evidence given $c$, $p(e_1\&e_2|c)$, is equal to the product of the separate conditional probabilities $p(e_1|c)$ and $p(e_2|c)$ for each element of the evidence separately.

We can generalize this to an ensemble of any number of elements of evidence $e_i$ and express the relationship succinctly: Conditional independence of an ensemble of $n$ elements of evidence $e_i$ given some condition $c$ holds when:

$$p(e_1\&e_2\&\ldots\&e_n|c) = \prod_j p(e_j|c).$$

Considering our example, we can consider the presence of a horizontal across the top, $F_{ht}$, and the presence of a vertical down the middle, $F_{vm}$. These would be conditionally independent given that the underlying letter was in fact intended to be a $T$ if it were true of the world that error entered into the registration of each of these two features of the letter T independently.

We can now write a version of our formula for inferring posterior probabilities under the assumption that conditional independence holds for all elements of evidence $e_j$ conditioned on all of the hypotheses $h_i$:

$$p(h_i|e_1\&e_2\&\ldots\&e_n) = \frac{p(h_i)\prod_j p(e_j|h_i)}{\sum_{i'} p(h_{i'})\prod_j p(e_j|h_{i'})}$$

We are still relying on many probabilistic quantities, but not as many as we would have to rely on if we separately represented the probability of each feature combination conditional on each hypothesis.

***Remark*:** Clearly, the assumption of conditional independence is unlikely to be exactly correct. However, it is hard to imagine proceeding without it. One way of alleviating the concern that relying on this assumption will lead us astray is to note that in cases where the occurrence of elements of evidence is highly correlated (even after conditioning on hypotheses), we might treat these elements as a single element, instead of as separate elements. Maybe that is what features are: clusters of elements that have a strong tendency to co-occur with each other. Another response to this situation would be to note that any explicit probability model involving sets of explicit hypotheses and elements of evidence is unlikely to be exactly correct for naturalistic stimuli. Words spelled using letters and their features as in the Rumelhart font are not really natural stimuli, since these items actually do consist of discrete units (letters) and these in turn consist of independent sub-units (letter features). This allows for the possibility of validly characterizing displays of such features in terms of a process in which conditional independence of features holds exactly. A learned, implicit probability model of the kind embodied in a Deep Belief Network (Hinton and Salakhutdinov, 2006) is likely to be a better model for naturalistic stimuli.

[4]Note that in general, $\sum_{i' \neq T} p(F_{ht}|l_{i'})p(l_{i'})$ may not be equivalent to $\sum_{i' \neq T} p(F_{ht}|l_{i'}) \sum_{i' \neq T} p(l_{i'})$.

### A GENERATIVE MODEL OF FEATURE ARRAYS

Consider the following description of how displays of letter features registered by a perceiver might be generated. An experimenter selects a letter to display from the alphabet with probability $p(l_i)$, which for now we will take to be simply 1/26 for each letter, and then generates a feature array as follows. Each letter has a set of correct feature values. For example, for *T*, the feature $F_{ht}$ is present, the feature $F_{vm}$ is present, and the feature $F_{hb}$, a horizontal line across the bottom, is absent (for simplicity, we will just consider these three features for now). However, when the actual feature array is generated, there is some small probability that each feature will not be generated correctly. The correctness of each feature is separately determined by an independent random process, e.g., by rolling a 20-sided die with a spot on just one side. If the spot comes up, the incorrect value of the feature is displayed. If it does not, the feature is generated correctly. The die is rolled once for each feature, and we are expressly assuming that the outcome of each roll is independent of the outcomes of all other rolls.

The above is a simple example of a generative model. If features were generated according to this process, then the probabilities of features are conditionally independent, given the letter identities. Note that if the generative process usually works perfectly and correctly generates all the correct features, but occasionally hiccups and gets all the features wrong at the same time, the elements of the evidence would not be conditionally independent. Note also that conditional independence can hold if the probability of feature perturbation is different for different features; this is likely if we think of the perturbation as occurring within the visual system, so that some features are more likely to be mis-registered than others, due to differences in their size, retinal position, or other factors.

Now, the true process generating feature arrays may not be exactly as described, just as the prior and likelihood values used may not be exactly accurate. However, a generative model in which feature values are perturbed independently can be treated as an assumption about the actual generative process, or alternatively it can be treated as an assumption about the model of the generative process that is utilized by the perceiver in a letter perception experiment. Such a model could be false, or only approximately true, and still be used by a perceiver. A further possibility is that the true model used by the perceiver is more complex, but that the assumption that the perceiver uses such a model provides a good approximation to the true model being used by the perceiver.

### THE SUPPORT FOR AN HYPOTHESIS AND THE LUCE CHOICE RULE

It will be helpful in our later development to write an expression we will call the Support ($S_i$) for a given alternative hypothesis $h_i$, given a set of elements of evidence $\{e\} = \{e_1, e_2, \ldots\}$ as follows:

$$S_i = p(h_i) \prod_j p(e_j|h_i)$$

For our example, the $h_i$ correspond to the different possible letter hypotheses and the $e_j$ correspond to the elements of the evidence. We will describe this overall support as consisting of the product of two terms, the prior $p(h_i)$ and the likelihood $p(e|h_i)$, which under the generative model described above is equal to the product of terms that might be called the element-wise likelihoods of each element of the evidence.

With this expression for the support of hypothesis *i* in the presence of evidence $\{e\}$, we can write Bayes' formula as:

$$p(h_i|\{e\}) = \frac{S_i}{\sum_{i'} S_{i'}}$$

As before, $i'$ is an index running over all of the alternative hypotheses, including hypothesis *i*. Readers familiar with the Luce (1959) choice rule will notice that this expression corresponds to Luce's rule, with the $S_i$ corresponding to the response strengths associated with the different choice alternatives.

As an exercise, consider a letter microworld with just the three features we have considered so far and just the letters *T*, *U* and *I*. Assume that according to the generative model, each letter is equally likely $p(T) = p(U) = p(I) = 1/3$. Regarding the features, we follow a policy used in the original IA model and carried over in the multinomial IA model: we explicitly represent the absence of a feature as an element of evidence, just like the presence of a feature. Thus, there are six possible elements of evidence or feature values relevant to identifying letters: a feature can be present or absent, for each of the three possible features.

To proceed with our exercise, the probability of each possible feature value (present or absent) is given for each of the three possible feature dimensions of each letter in **Table 1**. Here *h* stands for a high probability (let's say 0.95) and *l* for a low probability (0.05). Features cannot be both present and absent, so $l = 1 - h$. Assuming actual features are generated in a conditionally independent manner, we can then ask, what is the probability that the underlying letter was a T given that the following evidence $\{e\}$ is available: Horizontal at top *present*, Vertical at middle *absent*, Horizontal at bottom *absent*. Although these features do perfectly match the high-probability values for the letter *T*, the letter is more likely to be a *T* than a *U* or an *I*. See if you can verify this. Using the two equations above, along with **Table 1** and the specific numerical values given in this paragraph, you should be able to obtain an explicit probability for $p(T|\{e\})$. You should also be able to express simply why T is more probable that U or I given the available evidence.

**Table 1 | Probability that features take given values in the Letters T, U, and I.**

| Letter | Feature | | | | | |
|---|---|---|---|---|---|---|
| | Horiz. at Top | | Vert. thru Middle | | Horiz. at Bottom | |
| | Present | Absent | Present | Absent | Present | Absent |
| T | h | l | h | l | l | h |
| U | l | h | l | h | h | l |
| I | h | l | h | l | h | l |

*h in the table corresponds to a high probability, such as 0.95, and l corresponds to a low probability, such as 0.05.*

One additional issue may now be considered. We may ask, what happens if we are not told about one of the elements of the evidence? For example, we are told that the horizontal bar across the top is present and the vertical bar down the center is present but we simply are not told about the horizontal bar across the bottom (perhaps something is blocking our view of that feature in a perceptual display, for example). We would simply use those elements of evidence that we do have, and exclude the elements that are unspecified. Our existing expression already captures this policy implicitly, since when an element of evidence is missing it simply does not show up in the ensemble of elements $e_j$. However, it will prove useful to capture this case by elaborating the expression for S above to include explicit information specifying whether particular items of evidence are or are not present. A nice way to do this is to have a binary vector indicating whether the element of evidence is present or not. We have six possible elements of evidence in our example, as enumerated above. If we are given Horizontal at Top *present*, Vertical thru Middle *absent*, this vector would become: $v = 100100$. Then we would obtain the same results as before by writing $S_i$ as follows:

$$S_i = p(h_i) \prod_j p(e_j|h_i)^{v_j}$$

Where $\prod_j$ represents the product over all possible elements, and $v_j$ is equal to 1 for elements of evidence that are present, or 0 otherwise. Note that elements that are absent have no effect since for any non-zero $x$, $x^0 = 1$, and for all $p$, $p \cdot 1 = p$.[5] Note that the model we use here distinguishes between evidence of absence ("No horizontal bar is present at the bottom of the feature array") and the absence of evidence ("We do not know whether or not a bar is present at the bottom of the feature array"). In many cases, it is useful to distinguish between these two situations.

***Remark: Using p(e|h) to infer p(h|e)*** It is worth noting that we use knowledge of the probability of evidence given a hypothesis to infer the probability of a hypothesis given evidence. At first, this may seem counter-intuitive. Why don't we just store the value of $p(h|e)$, rather than always having to compute it? A similar counter-intuition arises in thinking about the "bottom-up" support for letter hypotheses by feature evidence. One might think that the effect of a feature's presence on the probability of a letter should depend on the probability of the letter given the feature, and not the other way around. The resolution of this counter-intuition depends on noticing that the posterior probabilities are not directly defined in the generative model, while the prior and the $p(e|h)$ terms are. Indeed, the posterior probability that a hypothesis is true depends on the entire ensemble of quantities in the generative model and the particular ensemble of elements of evidence that may be present, while the $p(h)$ and $p(e|h)$ values can be stable and independent. To contemplate this in a specific context, let us return to the question of the probability that a person is a woman, given that she has long hair. This quantity depends on three other quantities: the overall probability that a person is a woman; the probability that a woman has long hair; and the probability that a man has long hair. Each of these quantities can be changed independently, without affecting the others, while the probability that a person with long hair is a woman depends on all three. In short, in many contexts at least, it makes sense that we use $p(h)$ and $p(e|h)$ to compute $p(h|e)$.

[5]Strictly speaking, this formula is valid if the causes that could lead to missing evidence—e.g., no information about whether a particular feature is present or absent—are independent of the process that generates the feature values. This would be true if, for example, the probability that an occluder would partially restrict the visibility of an object were independent of the identity of an object.

#### SUMMARY: GENERALIZED VERSION OF BAYES FORMULA

To summarize the above development, the generalized version of Bayes formula for the posterior probability of hypothesis $h_i$, for $i = 1, \ldots, n$ mutually exclusive hypotheses and $j = 1, \ldots, m$ possible conditionally independent elements of evidence is:

$$p(h_i|e) = \frac{S_i}{\sum_{i'} S_{i'}}, \tag{5}$$

where $S_i$ stands for the support for hypothesis $h_i$, defined as:

$$S_i = p(h_i) \prod_j p(e_j|h_i)^{v_j} \tag{6}$$

## CALCULATING POSTERIOR PROBABILITIES WITH CONNECTIONIST UNITS USING THE SOFTMAX AND LOGISTIC FUNCTIONS

We now develop the idea that the posterior probability calculation just presented can be computed by a group of connectionist processing units, using a function called the **softmax** function. The neural network is illustrated in **Figure 3**. In this network, each

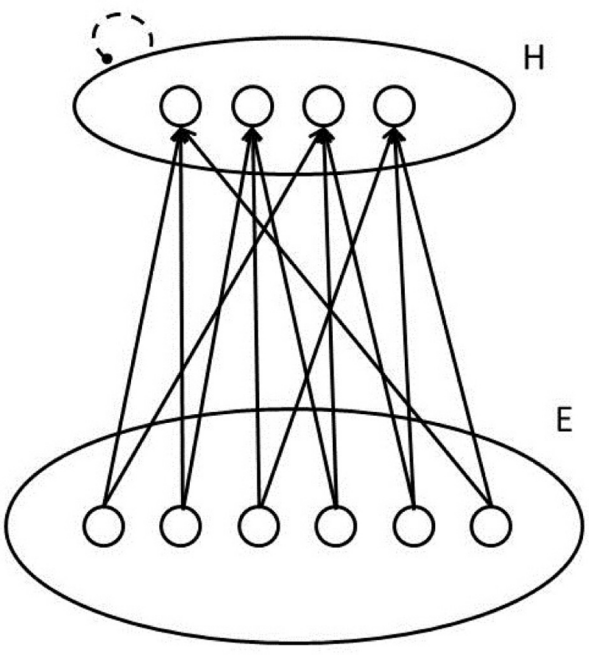



**FIGURE 3 | Sketch of a pool of units that can calculate posterior probabilities of patterns represented on its inputs using the softmax function.** Dashed line signifies lateral inhibition to normalize the activations of units in the pool.

unit corresponds to an hypothesis $h_i$, and has a bias term $b_i$, as well as incoming connections from units outside the ensemble. Each of these outside units indexed by $j$ stands for a possible element of evidence. When the element of evidence is present, the unit will have an activation value $a_j$ equal to 1; when it is absent, its activation will be 0. Each connection to a hypothesis unit from an evidence unit will have a strength or weight represented by the variable $w_{ij}$.

Concretely, pursuing our example, the units in the pool could correspond to possible letters, each unit's bias term could reflect a perceiver's bias to think the input contains the given letter, and the connection weights could reflect the perceiver's tendency to think the hypothesis is more (or less) likely, when the corresponding element of evidence is present. The pool described corresponds to one of the pools of letter level units in the MIA model, although we are considering just one such pool in isolation for now, without additional input from the word level.

In our network, as in most neural networks, each unit computes a summed or net input that reflects both its bias and the weighted sum of activations of other units:

$$net_i = b_i + \sum_j w_{ij} a_j$$

We will now see that if we set the weights and biases to appropriate values, then apply the **softmax** function defined below, the output of the function, represented here as $\rho_i$, will be equal to the posterior probability of the letter the unit stands for, as expressed by the generalized Bayes formula.

The **softmax** function is:

$$\rho_i = \frac{e^{net_i}}{\sum_{i'} e^{net_{i'}}}$$

The reader should already be able to see that the softmax has some relationship to the generalized Bayes formula. Indeed, as we shall discuss, the expressions $e^{net_i}$ and $e^{net_{i'}}$ correspond to the expressions for $S_i$ and $S_{i'}$ in that equation.

The essential idea is that the bias term and the weights will be chosen to correspond to the logarithms of the quantities that are multiplied together to determine the $S_i$ terms. Using the logs of these quantities, we add rather than multiply to combine the influences of the prior and the evidence. The resulting net input term corresponds to the log of the $S_i$ terms defined above. We then reverse the logarithmic transformation at the end of the calculation, using the exponential function.

The analysis relies on several facts about the log and exponential functions that we now review. First, the function $y = \log(x)$ is defined as the function that produces, when applied to its argument $x$, a number $y$ such that $e^y = x$. Note that log is used here to correspond to the natural logarithm, sometimes written $\log_e$ or ln. The exponential function of $y$, $e^y$ corresponds to the number $e$ taken to the power $y$, and is sometimes written $\exp(y)$. Given these definitions, it follows that $\log(e^y) = y$ and $e^{\log(x)} = x$. The graphs of the log and exp functions are shown in **Figure 4**.

The second important fact is that the log of the product of any number of quantities is the sum of the logs of the quantities:

$$\log(a \cdot b \cdot c \cdot \ldots) = \log(a) + \log(b) + \log(c) + \ldots$$

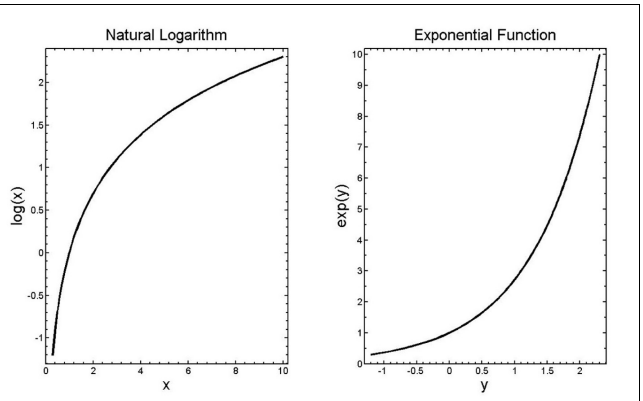



**FIGURE 4 | The log and exponential functions.**

Similarly, the log of the ratio of two quantities is equal to the difference between the logs of the quantities:

$$\log(a/b) = \log(a) - \log(b)$$

Finally, the log of a quantity to a power is that power times the log of the quantity:

$$\log(a^b) = b \log(a)$$

There are also useful related facts about exponentials, namely $e^{(a+b+c+\ldots)} = e^a \cdot e^b \cdot e^c \cdot \ldots$; $e^{(a-b)} = \frac{e^a}{e^b}$; and $e^{(a \cdot b)} = (e^a)^b$[6].

With this information in hand, we consider the expression we previously presented for $S_i$, the support for hypothesis $i$:

$$S_i = p(h_i) \prod_j p(e_j|h_i)^{v_j}$$

Taking logs, we see that:

$$\log(S_i) = \log(p(h_i)) + \sum_j v_j \log(p(e_j|h_i))$$

It should now be apparent that the net input as described above would correspond to the log of the support for the hypothesis represented by the unit if: (a) the value of the bias term were set to correspond to the log of the prior probability of the hypothesis; (b) each incoming weight were set to correspond to the log of the probability of the corresponding element of the evidence given the hypothesis; and (c) the activation of the external unit sending activation through the weight were to be equal to 1 when the evidence is present, and 0 otherwise. Stated succinctly in terms of defined quantities:

$$net_i = \log(S_i) \text{ if } b_i = \log(p(h_i)), w_{ij} = \log(p(e_j|h_i)), \text{ and } a_j = v_j.$$

[6]Those wanting to gain familiarity with these functions can obtain values by reading off of these functions, and check that the above relationships all hold up. For example $\log(2) + \log(3) = \log(2 \cdot 3) = \log(6)$ and $\log(8) - \log(4) = \log(8/4) = \log(2)$, and not $\log(4)$.

Now it should be clear that applying the softmax function:

$$\rho_i = \frac{e^{net_i}}{\sum_{i'} e^{net_{i'}}}$$

should set the value of the variable $\rho_i$ to be equal to the posterior probability of hypothesis $i$ given the set of elements of the evidence $e_j$ as long as $net_i$ corresponds to $\log(S_i)$ for all $i$, since $e^{\log(x)} = x$, as noted above. Substituting $\log(S_i)$ and $\log(S_{i'})$ into the softmax function where we find $net_i$ and $net_{i'}$ we will clearly obtain our generalized Bayes formula.

Thus, the neural network in **Figure 3**, employing the softmax function, calculates posterior probabilities by relying on a non-linear but monotonic function (the exponential function) of the sum of a set of terms, one for the prior probability of the hypothesis and one for each of the elements of the evidence.

***Why sums rather than products?*** One might be inclined to ask at this point, why should neural network modelers even bother computing net inputs as additive quantities? Why not compute the posterior probabilities more directly, without ever taking logs? The answer may in part be historical: the original model neuron introduced by McCulloch and Pitts (1943) summed weighted inputs, and if they exceeded a threshold the neuron's output was set to 1; otherwise the output was 0. This was intended to mimic both real neurons (which fire action potentials if their state of depolarization reaches a critical level) and logic gates (devices then send out a 1 or a 0 based on some logical function of their inputs). The logistic function discussed below, a close relative of the softmax function, was adopted for use in neural network models because it produced a graded rather than a discrete response, and could be differentiated. Only later did the connection to probability become apparent [first reflected, to my knowledge, in Hinton and Sejnowski (1983)]. But what about the brain itself? It is common to treat synaptic currents as being summed to determine the neuron's potential, which in turn determines its firing rate according to a non-linear function. It is possible that addition may be more robust and easier to implement in neurons than multiplication, especially when small probabilities are involved, since noise affecting such quantities can drastically distort the results of multiplying products, and in any case the computations are just as valid when conducted using addition of logarithms rather than multiplication, as long as we have a non-linear activation function like softmax to convert the influences back. Some further relevant observations are provided below.

### *Maximizing and matching using the neural network*

We can imagine a number of policies we might employ in using the $\rho_i$ values as a basis for overt responding. One policy would be to choose the alternative with the largest value of $\rho_i$; this corresponds to maximizing. Matching would occur if we were to choose alternatives with probability equal to the value of $\rho_i$. A gradient of possibilities between these extremes can be obtained by introducing a parameter usually called temperature, following the analogy to statistical physics introduced into neural networks research by Hinton and Sejnowski (1983). This usage corresponds to the analogy from physics, in which the temperature determines the degree of randomness in the behavior of elements of the system. In this version of the formula, our expression now becomes:

$$\rho_i(T) = \frac{e^{net_i/T}}{\sum_{i'} e^{net_{i'}/T}}$$

Our previous case corresponds to the situation in which $T = 1$. We can now imagine a policy in which we choose each alternative with probability $\rho_i(T)$, for different values of the $T$ parameter. As $T$ becomes small, the largest net input term strongly dominates, and in the limit as $T \rightarrow 0$ our policy converges on maximizing, since $\rho_i(T)$ will approach 1 for the unit with the largest net input and will approach 0 for all other units. As $T$ becomes large, the $\rho_i(T)$ will all approach $1/N$ where $N$ is the number of alternatives, corresponding to random guessing.

*Example*. The softmax function can be used to model response choice probabilities in many situations, under a matching assumption, where the $\rho_i$ correspond to choice probabilities. One case where the model provided an excellent fit arose in an experiment by Salzman and Newsome (1994). Here a monkey received a visual motion stimulus, corresponding to evidence favoring a particular alternative direction out of eight alternative motion directions. On some trials, the monkey also received direct electrical stimulation of neurons representing motion in a particular direction (treated in the model as another source of conditionally independent evidence). The monkey's choice behavior when both sources of evidence were presented together corresponded well to the predictions of the model. The experimenters estimated quantities corresponding to the bias terms and weights used in the softmax formulation. Although they did not mention Bayesian ideas, these terms could be treated as corresponding to logarithms of the corresponding Bayesian quantities.

***Lateral inhibition and effects of noise in the net input.*** The denominator of the softmax function can be seen as expressing a particular form of lateral inhibition, in that strong support for one alternative will reduce the value of $\rho_i$ for another. Some readers may notice that the inhibitory influence a unit exerts on others depends on its net input term (specifically, $e^{net_i/T}$), whereas it is natural to think of the $\rho_i$ as corresponding to the activations of the units for different alternatives. In most neural network models, units are usually thought to transmit their activation value, not their net input, both to exert excitatory and inhibitory influences. Do units use one variable for mutual inhibition and another to influence outside units? It is certainly a possibility. A computation of this kind could certainly be carried out, say, if the units in our networks corresponded to columns of neurons, in which some engaged in lateral inhibitory interactions while others sent excitatory signals to neurons in other pools. Also, it may be worth noticing that in practice, an iterative computational procedure in which the net input terms build up gradually and the denominator relies on the $\rho_i$ terms instead of the $e^{net_i}$ terms should converge to the same result, as in the REMERGE model of memory trace activation (Kumaran and McClelland, 2012).

It is also possible to view the softmax function as describing the outcome of a simple winner-take-all process. Suppose we simply allow each unit to compute its net input, subject to noise, and adopt the policy of choosing as our response the unit with the largest net input. If the noise is very small, and the weights and biases correspond to the probabilistic quantities above, then by choosing the unit with the largest net input we will always be maximizing the posterior probability. On the other hand if the noise is sufficiently large, the net input will be effectively swamped by the noise, and choosing the unit with the largest net input will correspond to random responding. With an intermediate amount of noise, the process just described approximates choosing alternatives with probability $\rho_i(T)$ as calculated by the softmax function, for some value of the parameter $T$ that depends on the amount of noise. In fact, if the noise affecting each unit is identically distributed according to a distribution called the extreme value distribution, then the choice probabilities will match those described by the softmax function exactly (Train, 1993). For those not familiar with the extreme value distribution, it is somewhat different from the Gaussian distribution, in that it is slightly skewed, but the shape is not drastically different from Gaussian, and simulations using Gaussian noise yield similar results to those expected using the extreme value distribution. The best characterization of noise in real neural populations is a matter of ongoing investigation, and it may not exactly match either the Gaussian or the extreme value distribution. In the absence of complete consensus, it seems reasonable to treat the noise in neural population activity as reasonably well approximated by the extreme value distribution, and thus to conclude that a simple winner-take-all procedure that could be implemented in real neural circuits can approximate probability matching, if the weights and biases have the right values, and can also approximate all policies between maximizing and pure guessing depending on the level of the noise[7].

### *The logistic function*

We now consider a variant of the scenario described above, in which we have just two mutually exclusive hypotheses. In this case it is possible to use bias terms and weights that allow us to calculate the posterior probability of one of the two hypotheses more directly, using the logistic function—the function we mentioned above that is very frequently used in setting the activations of units in neural network models. The approach is very natural when $h_1$ corresponds to the hypothesis that some proposition is true, and $h_2$ corresponds to the proposition that it is false, but can be applied to any situation in which there are two mutually exclusive and exhaustive alternatives. We will present the logistic function by deriving it from the softmax function for the special case of two alternative hypotheses.

[7]It may be useful to note that what determines how likely it is that the unit with the strongest net input is the most active unit is not the absolute magnitude of the noise but the ratio of the strength of the noise to the size of the difference in the net inputs to different units. Given this, the noise might remain constant, while the net inputs (and therefore differences among them) might build up gradually over time. In this way, as time progressed, we could go from chance performance to matching, and, if signals continue to grow stronger, to maximizing as a function of time spent processing.

We consider the calculation of the posterior probability of $h_1$, noting that the posterior probability of $h_2$ must be 1 minus this quantity. Specializing the softmax function of this case, we can write:

$$\rho_1 = \frac{e^{net_1}}{e^{net_1} + e^{net_2}}$$

where $net_1$ and $net_2$ are based on the values of the biases and weights as described above. Dividing the numerator by $e^{net_2}$, recalling that $e^a/e^b = e^{a-b}$ and noting that $e^{net_2}/e^{net_2} = 1$ we obtain:

$$\rho_1 = \frac{e^{net_1 - net_2}}{e^{net_1 - net_2} + 1}$$

Rather than compute each net input term separately and then subtract them, we can instead compute a single net input using biases and weights corresponding to the difference between the corresponding terms in each of these two expressions. That is, we define the combined net input as:

$$net = b + \sum_j a_j w_j$$

where $b = b_1 - b_2$ and $w_j = w_{1j} - w_{2j}$. Replacing the bias and weight terms with their probabilistic values we have $b = \log(p(h_1)) - \log(p(h_2))$ and $w_j = \log(p(e_j|h_1)) - \log(p(e_j|h_2))$, and recalling that $\log(a) - \log(b) = \log(a/b)$, we see that if the old biases and weights corresponded to the appropriate Bayesian quantities, the new combined bias term will be equal to $\log(p(h_1)/p(h_2))$ and each new combined weight $w_j$ will be equal to $\log(p(e_j|h_1)/p(e_j|h_2))$.

In terms of a single hypothesis $h$ that is either true or false, the bias term becomes $\log(p(h)/p(\bar{h}))$ or $\log(p(h)/(1-p(h)))$ and the $w_j$ becomes $\log(p(e_j|h)/p(e_j|\bar{h}))$. These are quantities often used in discussions of probabilities. The first is called the *log-odds*. The second is the log of the likelihood ratio, although in this case it is the element-specific likelihood ratio, specifying the log of the ratio of the likelihood of a specific element of the evidence when $h$ is true to the likelihood of that same element of the evidence when $h$ is false. The overall log likelihood ratio given $n$ conditionally independent elements of evidence is the sum of these quantities over all of the conditionally independent elements of the evidence.

From this we now can see that the posterior probability that some hypothesis $h$ is true can be expressed as:

$$\rho = \frac{e^{net}}{e^{net} + 1}$$

where the net input is the sum of a bias term equal to the log of the prior odds and each weight in the contribution from each element of the evidence is equal to the element-specific log likelihood ratio. This expression does not look exactly like the logistic function as usually written, but it is equivalent to it. We can produce the usual form of the logistic function by dividing the

numerator and the denominator by $e^{net}$, relying on the fact that $1/e^x = e^{-x}$:

$$\rho = \frac{1}{1 + e^{-net}}$$

This form of the function is used in simulators since it involves calling the `exp()` function only once, but they are both essentially the same function.

To summarize this section: The softmax function can compute according to Bayes' formula using biases and weights corresponding to the logs of key Bayesian quantities, while the logistic function computes according to Bayes' formula using biases and weights corresponding to logs of ratios of these quantities. The minus sign in the exponentiation in the logistic function reflects a simplification of the formula that slightly obscures the relationship to Bayes' formula but makes calculation quicker. It is also worth reiterating that the softmax and logistic functions could be used to describe the outcome of a process in which one simply chooses the alternative with the largest net input, subject to Gaussian noise. In such a case we might think of the system as attempting to maximize, but appearing to be doing something more like probability matching, because the noise sometimes makes the wrong alternative come out ahead.

### *Logistic additivity*

Here we discuss a characteristic of patterns of data we will call *logistic additivity*. This is a condition on the relationship between the posterior probability that some binary hypothesis $h$ is true, as we manipulate two independent sources of evidence, under the assumption that the sources of evidence are conditionally independent given $h$ and given $\bar{h}$. It is also, at the same time, a condition on the expected output of the logistic function, given that each source of evidence has an additive effect on the net input variable that is the input to this function. Logistic additivity is of special interest for us because [as pointed out by Massaro (1989)], the original IA model failed to exhibit this pattern, thereby failing to correspond to a proper Bayesian computation and to patterns often seen in behavioral data at the same time.

We will say that logistic additivity holds for the effects of two independent sources of evidence on the probability of some outcome when they have additive influences on the **logit** of the probability of the outcome given the two sources of evidence. The logit of a probability $p$ is defined as follows:

$$\text{logit}(p) = \log(p/(1-p))$$

With this expression defined, we can write the statement of the condition under which logistic additivity holds as:

$$\text{logit}(p(h_1|e_1, e_2)) = b + f_1(e_1) + f_2(e_2)$$

This result is nice for visualization purposes since it says that for a factorial combination of different levels of $e_1$ and $e_2$, we should obtain parallel curves. While we will not develop this point further here, these parallel curves can be turned into parallel straight lines by appropriate spacing of points along the x axis. In his excellent early analysis of context effects in word recognition (Morton, 1969) used this approach. Further details are presented in **Figure 5** and the corresponding caption.

We now show how logistic additivity follows from Bayes formula for the case of two sources of evidence $e_1$ and $e_2$ for hypotheses $h$ and $\bar{h}$. We work from Bayes formula, using $S = p(h)p(e_1|h)p(e_2|h)$ to represent the support for $h$ and $\bar{S} = p(\bar{h})p(e_1|\bar{h})p(e_2|\bar{h})$ to represent the support for $\bar{h}$, so that:

$$p(h|e_1, e_2) = \frac{S}{S + \bar{S}}$$

Dividing the numerator and denominator of this expression by $\bar{S}$:

$$p(h|e_1, e_2) = \frac{(S/\bar{S})}{1 + (S/\bar{S})}$$

It follows from this that:

$$1 - p(h|e_1, e_2) = \frac{1}{1 + (S/\bar{S})}.$$

If you do not see this immediately, add the two quantities together—clearly they sum to 1. Dividing the first expression by the second, we obtain:

$$p(h|e_1, e_2)/[1 - p(h|e_1, e_2)] = S/\bar{S}$$

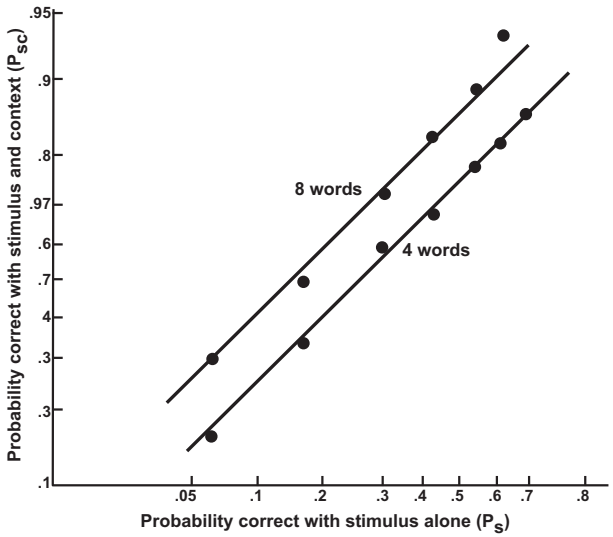



**FIGURE 5 | The joint effect of context and stimulus information on probability of identifying a word correctly, displayed on axes where points are spaced according to the logit of the indicated probabilities.** The *x* axis corresponds to the logit of the probability of identifying a target word when presented without context; in the experiment (Tulving et al., 1964), this probability was manipulated by using different exposure durations ranging from 0 to 120 ms. Two curves are plotted, one for cases in which an eight-word context was provided (e.g., for the target *raspberries*: "We all like jam made from strawberries and"), and one for the case in which only the last four words of the context was provided. The curves show that the context and stimulus information have additive effects on the logit of the probability of identifying the stimulus correctly. From Morton (1969). Reprinted with permission.

Replacing $S$ and $\bar{S}$ with the products they each stand for, and taking logs of both sides, we obtain:

$$\text{logit}(p(h|e_1, e_2)) = \log\frac{p(h)}{p(\bar{h})} + \log\frac{p(e_1|h)}{p(e_1|\bar{h})} + \log\frac{p(e_2|h)}{p(e_2|\bar{h})}$$

The right-hand side of this equation exhibits logistic additivity, with $\log(p(h)/p(\bar{h}))$ corresponding to $b$, $\log(p(e_1|h)/p(e_1|\bar{h}))$ corresponding to $f_1(e_1)$, and $\log(p(e_2|h)/p(e_2|\bar{h}))$ corresponding to $f_2(e_2)$.

Working directly from the logistic function we can proceed in a similar vein to arrive at the formula expressing logistic additivity. Given that $\rho = \frac{e^{net}}{e^{net}+1}$ it follows that $1-\rho = \frac{1}{e^{net}+1}$. From these observations, it follows that $\rho/(1-\rho) = e^{net}$, since the denominators cancel. Taking logs of both sides and replacing *net* with its definition we have:

$$\text{logit}(\rho) = b + a_1 w_1 + a_2 w_2$$

The idea that different sources of evidence—and in particular stimulus and context information—should exhibit logistic additivity was referred to as the *Morton–Massaro Law* by Movellan and McClelland (2001), and is a consequence of the assumptions of both Morton's and Massaro's (e.g., Massaro, 1989) models of how different sources of information are combined. Though neither model was explicitly formulated in Bayesian terms, it should be clear that these models follow from Bayes' formula and from the assumption that context and stimulus information are conditionally independent sources of evidence about the identity of an item in context.

Given the above analysis we can think of the logit transform of a probability (a number between 0 and 1) as converting the probability into an unbounded real number whose value exhibits additive influences arising from logs of prior odds and logs of the ratios of likelihoods of conditionally independent elements of evidence. The transform is the inverse of the logistic function, uncovering the underlying additivity of the contributions of the inputs to the function.

## PROBABILISTIC COMPUTATIONS IN THE MULTINOMIAL INTERACTIVE ACTIVATION MODEL

With the above background, we are finally ready to apply the ideas we have explored so far to the MIA model (Khaitan and McClelland, 2010; Mirman et al., in press). The goal of perception, according to this model, is to infer the underlying state of the world that gave rise to observed features. In this case, the goal is to infer the identity of the word and of the four letters that generated the features that reach the input to the model in a trial of a perception experiment using displays containing features in four letter positions.

A diagram of the model is presented in **Figure 6**. The diagram shows some of the units and a small subset of the connections in the neural network model, or equivalently, it depicts the set of multinomial random variables used in the model, and some of the constraints that influence the probability that these variables will take on particular values. The identity of the word is treated as the value of a multinomial random variable that can take on one of

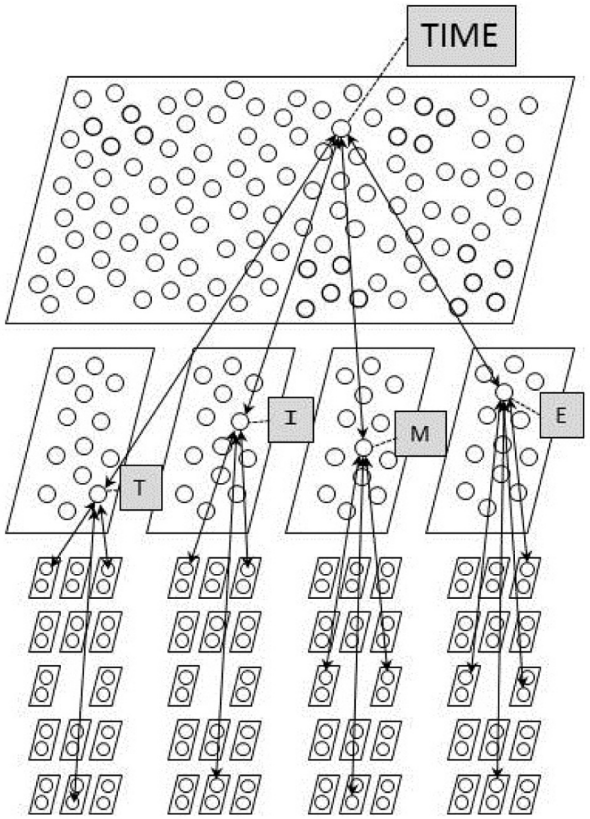



**FIGURE 6 | The architecture of the multinomial interactive activation model.** Each parallelogram in the figure corresponds to a pool of mutually exclusive units, corresponding to a multinomial random variable in the probabilistic conception of the model. The softmax function is used to calculate estimates of posterior probabilities for the word units and for each pool of letter units.

$n_w$ values where $n_w$ corresponds to the number of known words, and each word unit in the neural network model corresponds to one of the possible values this multinomial random value might take. Similarly, the identity of the letter in each position is treated as the value of one of four additional multinomial random variables each of which can take on one of 26 values corresponding to the letters of the alphabet, and each letter unit in each position corresponds to one of the values the variable for that position might take. Finally, the observed value of each feature in a given position is treated as the value of one of 14 multinomial random variables, each of which can take on either of two values (*present*, *absent*); in the neural network model, there is a separate unit for each of these two possible values within each of these 14 variables. There is a separate set of 14 multinomial variables for each position, or equivalently, a separate set of $14 \times 2$ feature units for each position.

Restating the goal of perception in terms of these variables, it is to infer values of the word and letter variables based on inputs specifying values of the feature variables. Note that the correct values of these variables cannot be determined with certainty, since the generative process that produces the observed features

is assumed to be probabilistic. The MIA model assumes that perception produces as its outcome a sample of a possible underlying state of the world that could have generated the observed features. This sample takes the form of a set of specific values for the multinomial word variable and the four letter variables (e.g., [WORD = TIME; LETTERS = {T,I,M,E}]), and corresponds to one word unit being active and one letter unit being active in each of the four letter positions. Alternative possible underlying states are sampled probabilistically, such that the probability of sampling each possible underlying state corresponds to the actual posterior probability that this was the underlying state that generated the observed features, according to the generative model embodied in its architecture and its connection weights. The model also provides a mechanism for doing so, based on a procedure we will describe below. Before we turn to the model, however, we must establish what the posterior probabilities of different underlying states of the world are, given that we have observed a set of feature values in each position as our evidence. To do this, we must first describe the generative model assumed to give rise to observed feature arrays.

## THE GENERATIVE MODEL OF THE MULTINOMIAL INTERACTIVE ACTIVATION MODEL

The generative model of the MIA model is a characterization of the process that produces the set of input features received by a participant in a letter and word perception experiment. The generative process can be envisioned with the help of **Figure 6**, with the proviso that the generative process runs strictly top down, whereas constraints among units run in both directions.

The first step in the generative process is to select a word at random from a list of words that are four letters long, in accordance with a base rate for each word (represented $p(w_i)$). We then generate letters independently in each position with probability $p(l_{j_p}|w_i)$, where we use $l_{j_p}$ to represent letter $j$ in position $p$[8]. Given the above procedure, the probability of a particular word $w_i$ and four letters $\{l_{j_p}\}$ is:

$$p(w_i, \{l_{j_p}\}) = p(w_i) \prod_p p(l_{j_p}|w_i).$$

Now using the letter sampled in each position independently, we sample values for features for each letter. As noted above, we treat the set of features as consisting of 14 separate feature dimensions, for each of which there are two explicitly represented possibilities, one that the feature is present and one that it is absent. Independently for each dimension, we select the value for a given feature dimension with probability $p(f_{v_{dp}}|l_{j_p})$[9].

The generative process has produced a word, a letter in each position, and a value for each feature of each letter. We will call this set of elements a *path* $P_{i,\{j_p\},\{v_{dp}\}}$ of the generative process, and subscript it with the indices of all of the selected elements, one for the word ($i$), a set of four indices $\{j_p\}$ for the letters, where $p$ runs over the four positions, and the set of $4 \times 14$ indices $\{v_{dp}\}$ each specifying the value $v$ (*present*, *absent*) of each feature dimension $d$ of each position $p$. The probability of a given path is:

$$p(P_{i,\{j_p\},\{v_{dp}\}}) = p(w_i) \prod_p p(l_{j_p}|w_i) \prod_d p(f_{v_{dp}}|l_{j_p}).$$

Simplify the notation slightly, using $p(\{v_{d_p}\}|l_{j_p})$[10] to represent $\prod_d p(f_{v_{dp}}|l_{j_p})$, this becomes:

$$p(P_{i,\{j_p\},\{v_{dp}\}}) = p(w_i) \prod_p p(l_{j_p}|w_i) p(\{v_{d_p}\}|l_{j_p}). \quad (7)$$

We will refer to this equation later as the *path probability equation*.

We can now consider the posterior probability of a particular combination of unobserved word and letter variables, given an observed set of features, representing this with the expression $p(w_i, \{l_{j_p}\}|\{v_{dp}\})$. This is just the path probability of the full path involving the given word, letters, and observed features, divided by the sum of the path probabilities of all of the paths that could have generated the observed features:

$$p(w_i, \{l_{j_p}\}|\{v_{dp}\}) = \frac{p(P_{i,\{j_p\},\{v_{dp}\}})}{Z_{\{v_{dp}\}}}.$$

The denominator represents a quantity called the *partition function*. It stands for the sum over all $n_w \times 26^4$ path probabilities. The above equation is nothing more than an application of Bayes formula, but in a situation where the alternative hypotheses are the alternative *combinations* of possible word and letter identities that could have produced the given evidence, or ensemble of features.

Let us now consider how we could calculate the posterior probability that the word responsible for a given path was word $i$, given that we observed the set of features $\{v_{dp}\}$. This will be the sum, over all paths that can generate these features starting from the word $i$, of the probabilities of these paths, divided by the sum over all of the paths that could have generated the observed features:

$$p(w_i|\{v_{dp}\}) = \frac{\sum_{j_1,j_2,j_3,j_4} p(w_i) \prod_p p(l_{j_p}|w_i) p(\{v_{d_p}\}|l_{j_p})}{Z_{\{v_{dp}\}}}$$

[8]In principle, for each letter in each word we could have a full table of 26 times 4 entries, $p(l_{j_p}|w_i)$, but for simplicity we will assume (as in Mirman et al., in press), that $p(l_{j_p}|w_i)$ is the same, and is equal to a fixed parameter value $c_{L|W}$ if $l_j$ is the correct letter in position $p$ of word $i$ and that the remaining probability, $1 - c_{L|W}$, is split evenly among the remaining 25 letters. Thus, if $c_{L|W} = 0.95$, the value of $p(l_{j_p}|w_i)$ for the incorrect letters will be $0.05/25 = 0.002$. Using these numbers, with probability $0.95^4 \simeq 0.81$ all four letters generated from the chosen word will be the correct letters, but with probability $1 - 0.95^4 \simeq 0.19$ there will be one or more incorrect letters.

[9]Again for simplicity, we use a single parameter for correct features, $c_{F|L}$; for incorrect features, the corresponding probability is $1 - c_{F|L}$.

[10]Note the distinction between $\{v_{dp}\}$, the full set of indices of the active feature value units across all dimensions and all positions, and $\{v_{d_p}\}$, the set of indices of the active feature values in position $p$.

The summation in the numerator is the sum over the $26^4$ possible combinations of the 26 possible letters, one in each of the four letter positions, and $Z_{\{v_{dp}\}}$ is the partition function as above.

It is useful at this point to introduce the conceptual and terminological distinction between the *joint* posterior probability of a *combination* of variables and the *marginal* posterior probability of a *single* variable. The quantity $p(w_i, \{l_{j_p}\}|\{v_{dp}\})$ is an example of a joint posterior probability (in this case, of the combination of the indexed word and the four indexed letters), whereas $p(w_i|\{v_{dp}\})$ is an example of a marginal posterior probability (in this case, of just the indexed word). There are also marginal posterior probabilities associated with each of the indexed letters, e.g., for the first position $p(l_{j_1}|\{v_{vp}\})$. The marginal posterior probability that a single variable has a given value is the sum of the joint posterior over all of the combinations of variables in which the variable has the given value. For example, the marginal posterior probability of word $i$ is the sum over all of the combinations involving word $i$ of the joint posterior probability of the combination. As we will see, some procedures naturally calculate marginal posterior probabilities, while other procedures naturally sample from joint posterior probabilities. We will consider these concepts further as we proceed.

It will simplify further analysis to note that $p(w_i)$ is a constant that can be pulled out of the summation in the numerator above, and that we can use the distributive law[11] to rewrite $\sum_{j_1 j_2 j_3 j_4} \prod_p x_{j_p}$ as $\prod_p \sum_j x_{j_p}$. Using these two facts, the above reduces to:[12]

$$p(w_i|\{v_{dp}\}) = \frac{p(w_i) \prod_p \sum_{j_p} p(l_{j_p}|w_i) p(\{v_{d_p}\}|l_{j_p})}{Z_{\{v_{dp}\}}}$$

The value we obtain for each word $i$ corresponds to the marginal posterior probability of the word given the observed features.

Now, let us turn to the problem of calculating the marginal posterior probability that the letter in some arbitrary letter position is letter $j$, given the full set of feature values $\{v_{dp}\}$ over the four positions. This probability is just the sum of probabilities of all of the paths that involve letter $j$ in position $p$ and the given feature values in all four positions, divided by the sum of the probabilities of all of the paths that could have generated the given features. The expression below represents this summation. We focus on position 1 to simplify the notation—analogous expressions can be written replacing the index 1 with the index of any of the other letter positions.

$$p(l_{j_1}|\{v_{dp}\}) = \frac{\sum_i \sum_{\{j_2 j_3 j_4\}} p(w_i) p(l_{j_1}|w_i) p(\{v_{d_1}\}|l_{j_1}) \prod_{p \neq 1} p(l_{j_p}|w_i) p(\{v_{d_p}\}|l_{j_p})}{\sum_{j'_1} \sum_i \sum_{\{j_2 j_3 j_4\}} p(w_i) p(l_{j'_1}|w_i) p(\{v_{d_1}\}|l_{j'_1}) \prod_{p \neq 1} p(l_{j_p}|w_i) p(\{v_{d_p}\}|l_{j_p})}$$

The expression looks complex[13], but if we approach it slowly it should make sense. Starting with the numerator, we start with the notation for the summation over all of the $n_w \times n_l^3$ possible paths that could have generated the given features and that involve letter $j$ in position 1. The probability of each of these paths is then the product of the prior probability for the word involved in the specific path, $p(w_i)$, times a corresponding expression for each of the letters involved in the path.

Once again we can simplify. Looking first at the numerator, we can pull out the expression $p(\{v_{d_1}\}|l_{j_1})$ from the summation over words and letter combinations, since this expression is constant with respect to these. Likewise, we can pull $p(w_i)p(l_{j_1}|w_i)$ out of the summation over letter combinations, since it too is constant in all of these combinations. We can then use the distributive law to replace $\sum_{\{j_2 j_3 j_4\}} \prod_{p \neq 1} p(l_{j_p}|w_i) p(\{v_{d_p}\}|l_{j_p})$ with $\prod_{p \neq 1} \sum_{j_p} p(l_{j_p}|w_i) p(\{v_{d_p}\}|l_{j_p})$. In the denominator, we have partitioned the sum of the full set of path probabilities into subsets, one for each set of paths involving a different letter in position 1. We can apply the simplifications just described for the numerator to each such term in the denominator to obtain:

$$p(l_{j_1}|\{v_{dp}\}) = \frac{p(\{v_{d_1}\}|l_{j_1}) \sum_i p(w_i) p(l_{j_1}|w_i) \prod_{p \neq 1} \sum_{j_p} p(l_{j_p}|w_i) p(\{v_{d_p}\}|l_{j_p})}{\sum_{j'_1} p(\{v_{d_1}\}|l_{j'_1}) \sum_i p(w_i) p(l_{j'_1}|w_i) \prod_{p \neq 1} \sum_{j_p} p(l_{j_p}|w_i) p(\{v_{d_p}\}|l_{j_p})}$$

The leftmost factor in the numerator $p(\{v_{d_1}\}|l_{j_1})$ now corresponds to the standard Bayesian quantity $p(\{e\}|h)$, where $\{e\}$ is the bottom-up evidence for the ensemble of features in position 1 and $h$ is the hypothesis that the letter in position 1 is letter $j$. Everything else in the numerator specifies what we will call $p(l_{j_1}|c)$, the probability of letter $j$ in position 1, given the context $c$, where the context is the set of features in all of the other letter positions. Thus, we could rewrite the numerator as $p(\{v_{d_1}\}|l_{j_1})p(l_{j_1}|c)$. The denominator consists of a sum over all of the letters of corresponding quantities, so we can rewrite the above to express the posterior letter probability:

$$p(l_{j_1}|\{v_{dp}\}) = \frac{p(\{v_{d_1}\}|l_{j_1}) p(l_{j_1}|c)}{\sum_{j'_1} p(\{v_{d_1}\}|l_{j'_1}) p(l_{j'_1}|c)} \quad (8)$$

This equation once again looks like Bayes' formula, but this time, we use the probability of the item given the context in place of the prior or base rate. This should make sense: We can think of what we are doing here as using the context to set the "prior" probabilities of different letters to context-specific values, combining these context specific prior probabilities with the contribution of

[11] This law states that for all $a$, $b$, $c$, $d$: $(a + b)(c + d) = ac + ad + bc + bd$.

[12] It is worth noting that the simplification here dramatically speeds calculation. Instead of computing $26^4$ separate products of four quantities and then adding these all up, we compute the product of four sums of 26 quantities, producing a speed up of 17,000:1. It is also easier to implement the add-then-multiply simplification as a parallel computation in a neural network.

[13] In both the numerator and denominator of this equation and the next one, there is a line break before the product symbol $\prod$.

the evidence, to calculate the total support for each of the possible alternative letters.

### ALTERNATIVE PROCEDURES FOR CALCULATING AND SAMPLING POSTERIOR PROBABILITIES GIVEN OBSERVED FEATURE ARRAYS

The above can be thought of as a mathematical characterization of the true posterior joint and marginal probabilities of each word and of each letter in each position, conditional on observing some set of features $\{v_{dp}\}$, under the generative model. How might we calculate, or sample from, these quantities during perception?

We now describe two different ways to calculate the *marginal* posterior probabilities of words and letters—a *undirectional* method and an *interactive* method. After that we will describe how the updating procedure used in the MIA model allows us to sample from the *joint* posterior distribution, and (as a byproduct) also the marginal posterior distribution over words and over letters in each position.

#### *A unidirectional calculation method*

Our first calculation method is completely non-interactive—information flows in a single direction between each pair of pools, as shown in **Figure 7A**[14]. Both **Figure 7A** and the text below apply to the particular case of calculating the posterior probability of possible letters in position 1, given the full set of features $\{v_{dp}\}$.

1. For each letter in each position, including position 1, we first calculate $p(\{v_{d_p}\}|l_{j_p})$. This corresponds to the *upward* arrows from each feature array to each letter array in **Figure 7A**.
2. For each word, we then calculate $p(w_i)\prod_{p\neq 1}\sum_{j_p} p(l_{j_p}|w_i)p(\{v_{d_p}\}|l_{j_p})$. This is the support for each word, given the feature information in all positions other than position 1, and we will thus call it $S_{i/1}$[15]. This corresponds to the three *upward* arrows from the position 2, 3, and 4 letter arrays to the word array in **Figure 7A**.
3. For each letter $j$ in position 1, multiply each of the above word-specific terms by $p(l_{j_1}|w_i)$ and sum over words to obtain: $\sum_i p(l_{j_1}|w_i)S_{i/1}$. These quantities are the $p(l_{j_1}|c)$ terms we need, and the computation corresponds to the downward arrow in **Figure 7A** from the word level to the letters in position 1.
4. Finally, calculate $p(l_{j_1}|\{v_{dp}\})$ using the posterior letter probability equation (Equation 8), taking $p(\{v_{d_1}\}|l_{j_1})$ from step 1 and the $p(l_{j_1}|c)$ from step 3.

The procedure can, of course, be applied to any letter position, just exchanging the roles of position 1 and any other position.

***A drawback of the unidirectional method.*** The method just reviewed is basically consistent with the ideas of Massaro (1989) and Norris and McQueen (2008) and elsewhere. Both argue that when identifying the letter (or phoneme) in a particular string position, we must separately calculate context and stimulus support for each alternative, then combine these quantities only as the final step of our computation. The idea seems sensible when we think about using preceding context to help recognize the next phoneme in a spoken word. We can imagine generating expectations based on the input received so far for the phoneme next to come, then combining these expectations with bottom-up information about this next phoneme to compute its posterior, iterating this procedure for each successive phoneme. However, experimental evidence (e.g., Rumelhart and McClelland, 1982) supports the view that perception of letters in every position of a briefly-presented word benefits from contextual influences from all other positions. The data indicates that the perception of each letter should benefit from the context provided by all of the other letters, and that these computations should be carried out in parallel, so that these influences can occur while the input is available for identification. Subsequent context also affects phoneme perception from spoken input, even though the context does not arrive until after the target phoneme (Warren and Sherman, 1974; Ganong, 1980), a key finding motivating the interactive architecture of the TRACE model of speech perception (McClelland and Elman, 1986).

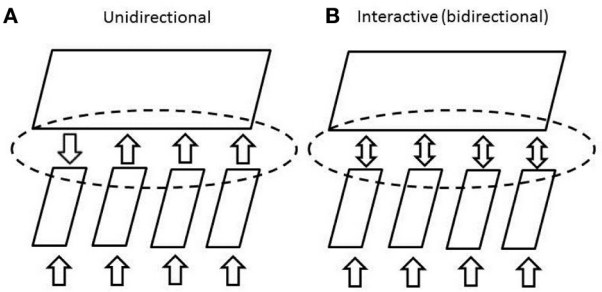



**FIGURE 7 | Schematic diagram of flow of computation for: (A) unidirectional procedure for calculating the posterior probabilities of letters in the first position and (B) bi-directional procedure for calculating the posterior probabilities of all four letters.** The dashed ovals highlight the differences between the two procedures. In the unidirectional procedure, calculation proceeds upwards to the word level from positions 2, 3, and 4, and downward only for position 1. In the interactive procedure, calculation proceeds upwards and downwards in all four positions.

In general, to maximize the use of context, it seems desirable to calculate posterior probabilities for each letter using the context provided by all the other letters, and it could be useful to calculate posterior probabilities for words, based on all of the letters, as well. The original IA model achieved something close to this, but not exactly—many of the complaints by Massaro and later by Norris et al. (2000; Norris and McQueen, 2008) focused on the fact that the model did not get the posterior probabilities exactly right; indeed, as documented by McClelland (1991) and Movellan and McClelland (2001), the original IA model failed to exhibit logistic additivity. Here we consider how the correct posterior probabilities can be calculated by an interactive procedure.

To calculate the posterior probabilities over words, we should of course include input from all four positions in the

[14]In a sense the flow is still both bottom up and top-down, but there is no back-and-forth communication, which is the essence of interactive activation.

[15]It could be worth noting that we have the option of normalizing the above quantities for each word by dividing them all by their sum. This step will not make any difference, since ratios of these quantities will be used later in calculating posterior probabilities at the letter level.

corresponding calculation of the $S_i$ for each word. To calculate the context terms for a given position—say position 1—we have to exclude its contribution to the word level to obtain the appropriate $S_{i/1}$ values. It would be possible to calculate $S_i$ along with $S_{i/1}$, $S_{i/2}$, *etc.*, separately in parallel, but it seems redundant and computationally wasteful. Fortunately, there is a simple way to avoid the redundancy.

### *A parallel, interactive method*

The approach we now consider is a specific instance of the approach proposed by Pearl (1982)—it is not the procedure we use in the MIA model, but it is useful to understand both procedures, and the differences between them. Pearl's approach allows processing to occur in parallel for all four letter positions, relying on the bi-directional propagation of information, as shown in **Figure 7B**, minimizing the redundancy just noted. The key observation (specialized for our specific circumstances) is that the posterior probability for each word contains a product of terms, one from each letter position. The term from a given letter position $p$ to each word unit $i$ is $\sum_j p(l_{j_p}|w_i)p(\{v_{d_p}\}|l_{j_p})$. Suppose we call each of these quantities $r_{i_p}$. Then we can calculate the full bottom-up support for each word combining the $r_{i_p}$ across all four positions, saving the $r_{i_p}$ values so that we can divide them back out in calculating the $p(l_{j_p}|c)$ factors for each position. In more detail, here is the procedure:

1. For each letter in each position, calculate $p(\{v_{d_p}\}|l_{j_p})$.
2. For each word, then calculate $S_i = p(w_i)\prod_p r_{i_p}$ where $r_{i_p}$ is as defined above, using the values calculated in step 1. $S_i$ represents the total support for each word, given the feature information in all positions and the prior word probability, and can be used to calculate the posterior probability of each word by dividing through by the sum of all of the $S_i$.
3. For each letter position, we now calculate the appropriate top-down term by dividing $S_i$ by $r_{i_p}$ to obtain $S_{i/p}$. We then proceed to calculate, for each letter $j$, $p(l_{j_p}|c) = \sum_i p(l_{j_p}|w_i)S_{i/p}$ as in the unidirectional procedure.
4. For each position, we finally calculate $p(l_{j_p}|\{v_{dp}\})$, using the $p(\{v_{d_p}\}|l_{j_p})$ from step 1 and the $p(l_{j_p}|c)$ from step 3.

This procedure is a specific instance of the one proposed by Pearl (1982) for unidirectional causal graphical models (models in which causality propagates only in one direction, as in our generative model), subject to the constraint that each multinomial variable (i.e., each a set of mutually exclusive hypotheses) in the graph has at most one parent, i.e., one variable that it is conditioned on in the generative model. The generative model underlying the MIA model is an example of such a graph: In the generative model, the multinomial word variable has no parents; each of the four multinomial letter position variables depends only on the word variable; and each of the 14 binomial feature dimension variables in each letter position depends only on the letter variable for that position. The method allows for iterative, i.e., interactive updating; as new information arrives at any of the variables, it can be propagated through to update all of the other variables. There is an inherent sequentiality moving upward and then downward, but information can flow back and forth in both directions. If feature information built up gradually over time, the process could be iterated repeatedly, updating all of the variables as new evidence arises.

Pearl's method is an elegant and general method, and is now a long established part of the fabric of probabilistic computation. Interestingly, the idea did not come up in the development of the IA model, even though the key idea of dividing back out one's contribution to a parent when receiving top-down input from the parent was proposed by Rumelhart (1977). Perhaps one reason why Rumelhart did not suggest we explore this idea when we were developing the original IA model may be that Pearl's method requires each multinomial variable to keep separate track of its bottom-up and top-down values. What gets passed up in Pearl's algorithm is strictly feed-forward information; what gets passed down to a given multinomial variable carefully excludes the information that came up through it, and must be kept distinct[16]. A feature Rumelhart found pleasing in the original IA model was that the computation was entirely homogeneous. In an early talk on the model, he had on one of his transparencies: "activation is the only currency" transmitted between units in the network.

An important characteristic of Pearl's approach is that the posterior probabilities calculated for each variable are marginalized over all possible values of all of the other variables. To see what this means concretely, consider the set of features shown in **Figure 8**. The features in each position are consistent with two letters (H or F in the first position, E or O in the second position, and W or X in the third)[17]. The features are also consistent with four words: FEW, FOX, HEX, and HOW[18]. Pearl's algorithm will allow us to calculate that these words and letters are the most likely. Ignoring differences in word frequency, the words would all be equally likely, and so would the letters. If we selected one

[16]The importance of keeping these messages distinct becomes even more clear if the top-down signals need to be passed down more than one level, a possibility that arises in Pearl's general formulation.

[17]To the human eye, the features in position 1 seem consistent with the letter B as well as the letters F and H. However, in the Rumelhart and Siple font, B does not have a vertical on the left, so that letter is ruled out by the given features. Also, humans appear not to entertain the possibility of W in the third position, perhaps because the segments appear to terminate at the bottom, but again, the given features are equally consistent with W and X in the Rumelhart and Siple font.

[18]The word HEW is also consistent with one letter in each position, but it is very low in frequency and for the sake of the example we assume it is not known to the perceiver.

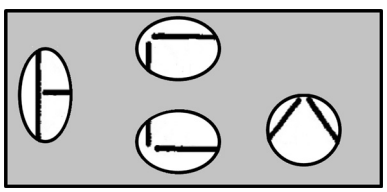


**FIGURE 8 | A few features from each position of a three-letter word.** Based on the Rumelhart and Siple font, there are two consistent letters in each position, and four consistent words.

word from those that are equally probable, and one letter from each position from those that are equally probable, we could end up with the conclusion that the word is FEW but that the letters are H, O, and X (letters that, together, don't even form a word).

The approach used in the MIA model samples from the *joint posterior* of the generative model. That is, it samples from the space of composite hypotheses consisting of one word and one letter in each position. For example, the joint hypothesis [WORD = FEW; LETTERS = {F,E,W}] is one such hypothesis, while the hypothesis [WORD = FEW; LETTERS = {H,O,X}] is another. The first of these joint hypotheses is far more likely than the second. The MIA model assumes that perceivers select among alternative joint hypothesis weighted by their overall probability. We now turn to the procedure used for doing this.

#### *Sampling from the joint posterior in the MIA model*

The final approach we will consider, the one used in the MIA model, is based on the Bayesian procedure known as *Gibbs sampling*, and was used in the Boltzman machine by Hinton and Sejnowski (1983). Gibbs sampling is discussed in more detail below; for now, we note that Gibbs sampling is a procedure used to sample from the joint posterior distribution of a probabilistic model by iteratively updating the states of unobserved variables probabilistically, based on current values of observed and other unobserved variables. We present the specific version of these ideas used in the MIA model, which have been adapted and specialized for our case.

In the model, there are word, letter, and feature units as illustrated in **Figure 6**, and weights are considered to be bidirectional, as in the figure, but their values are defined only in one direction. At the word level, each word unit has a bias term, corresponding to the log of its prior probability, $\log(p(w_i))$. The connection weight linking each word to each letter is set equal to $\log(p(l_{j_p}|w_i))$, and the weight linking each feature to each letter is set to $\log(p(f_{v_{d_p}}|l_{j_p}))$.

We specify the input to the model by setting the values of the feature units to correspond to a specific input feature array. With all units at the letter and word levels initially off, we proceed as follows:

1. For each position, we calculate each letter's net input. Since the word units are all off, there will be no contribution from the word level, and each letter unit's net input will correspond to $\log p(\{v_{d_p}\}|l_{j_p})$.
2. Within each letter position we then select one unit to be on, using the softmax function to calculate the probability of selecting each letter, given the net inputs to all of the letter units.
3. We then calculate the net input to each word unit, based on the single active letter unit in each position.
4. We then select one word unit to be on, again using the softmax function and the net inputs to all of the word units.
5. We calculate each letter's net input again, noting that now, one unit at the word level is active on each iteration, providing top-down input that affects the net input to each letter unit, in addition to the bottom-up input coming in from the feature level.
6. We then select one letter unit to be on in each letter position, using softmax.
7. We repeat steps 3–6 several times, then stop with one word unit active and one letter unit active in each position.

This iterative process in steps 3–6, which corresponds to Gibbs sampling, is called "settling." The initial bottom-up pass in steps 1–2 helps the network to start the settling process from an initial state usefully constrained by the featural input.

The state of activation in the network after settling for many iterations will be a sample from the joint posterior of the generative model (we will consider why this is true below). That is, if we ran this whole procedure a very large number of times, and counted the number of times the pattern at the end of settling corresponded to each possible joint hypothesis (one word and one letter in each position), the proportion of times the network settled to each such pattern would correspond to the posterior probability of the corresponding joint hypothesis.

Running the above procedure hundreds of thousands of times would not be very efficient, but we do not propose that perception involves such a process. Instead, we propose that each trial of a perceptual experiment involves a single instance of running the above procedure. Each such instance generates a single sample from the above process, capturing the probabilistic nature of performance in behavioral experiments. In a perceptual experiment where the task is to identify the letter in a specified position (as in most of the experiments modeled using the original IA model), we can imagine that the participant simply reads out the identity of the letter corresponding to the active unit in the appropriate position. Note that this procedure is a way to use a sample from the joint posterior as a sample from the marginal posterior for a particular multinomial variable (e.g., the letter in the specified position).

***Relationship to Gibbs sampling and Boltzmann machines.*** The above procedure is related to the updating procedure proposed for Boltzmann machines by Hinton and Sejnowski (1983, 1986). One difference is that in the original Boltzmann machine, units are not organized into pools corresponding to multinominal random variables. Rather, each unit is treated as a separate (binary) random variable. Units are updated one at a time, selecting the next unit to update sequentially and at random, using the logistic function. Our network is similar, but our units are organized into pools, each corresponding to a single multinomial variable, such that only one unit per pool/variable is allowed to be active at one time. In both cases, after an initial burn-in period (corresponding to what we called settling above), networks visit global states with probability proportional to $e^{G_s/T}$, where $G_s$ is the goodness of the state and $T$ corresponds to the temperature. The goodness of a state is defined as:

$$G(s) = \sum_{i<j} a_i a_j w_{ij} + \sum_i a_i b_i,$$

where the summation runs over all pairs of units (with each pair of units counted only once) and $w_{ij}$ corresponds to the

bi-directional weight between the two units[19]. Additional terms can be included to represent external inputs to units, but these can be captured using weighted inputs from units whose activations are treated as clamped, as the set of input feature units are in our procedure.

For the MIA model, the goodness of a state in which one word is active and one letter in each position is active, given a set of input feature values clamped onto the input units, is given by what we will call the MIA goodness equation:

$$G(s|\{v_{dp}\}) = b_i + \sum_p (w_{i j_p} + \sum_d w_{j_p, v_{d_p}}) \quad (9)$$

Here $i$ indexes the single word unit that is active at the word level, the four values of $\{j_p\}$ (one for each position $p$) index the active letter units in each of the four positions $p$, and the set of 4 times 14 values of $\{v_{d_p}\}$ represent the indices of the active feature-value units on each feature dimension in each feature position. The activations of the units involved are not expressed as variables because they are all equal to 1; no other terms occur in the goodness because all other units' activation values are 0.

As an exercise, the reader can check that the goodness of a state as represented by the MIA goodness equation is equal to the log of the probability of the corresponding path under the generative model, by taking the log of the path probability equation (Equation 7). Given that this is so, if we run our network at some temperature $T$, then the network will visit this state with probability proportional to $e^{\log(p(P_s))/T}$, where $p(P_s)$ is the probability of the path corresponding to state $s$. The posterior probability of visiting this particular state given the particular feature values can be calculated by dividing through by the sum of the exponentiated and $T$-scaled goodnesses of all of the states that can be visited given the feature values:

$$p(S|\{v_{dp}\}/T) = \frac{e^{G_{s|\{v_{dp}\}}/T}}{\sum_{s'} e^{G_{s'|\{v_{dp}\}}/T}}$$

For the case where $T$ is equal to 1, we obtain:

$$p(S|\{v_{dp}\}) = \frac{e^{G_{s|\{v_{dp}\}}}}{\sum_{s'} e^{G_{s'|\{v_{dp}\}}}}$$

The probability that the network is in state $S_i$ after settling is thus equal to the posterior probability of the state, given the evidence.

Hinton and Sejnowski (1983, 1986) focused on the task of finding the single best joint hypothesis using a process they called *simulated annealing*. In this process, one engages in a similar sequential update process to that described above, but with gradually reducing temperature. The procedure we have described operates at a fixed temperature. At lower temperatures, the preference for units with stronger net inputs is amplified, and as $T$ goes to zero, the procedure will allocate all of the probability to the alternative with the largest net input. Gradually lowering the temperature corresponds to gradually increasing the relative probability of visiting the alternatives with the largest posterior probability. It may be worth noting that a gradual change in the clarity of evidence can have a similar effect as a gradual change in temperature, or that running the procedure when the evidence is very weak can be similar to running the procedure at very high temperature. Thus, perception with very weak stimuli may correspond approximately to running the model at very high temperature, and gradual buildup of information over time may correspond to simulated annealing. These ideas may be worth developing further in extensions of the MIA model.

***Why does the MIA model sample correctly from the posterior?*** So far we have stated without proof that "after a burn-in period" and at fixed temperature, states are sampled in proportion to $e^{G(s)/T}$. How do we know that this is true? For particular cases, we can demonstrate the validity of this result via stochastic simulation, and we have done so for several cases, showing results for one specific case in Mirman et al. (in press). The fact that it is true for all cases follows from the fact that the sampling procedure we are using is an instance of a Gibbs sampling procedure, introduced by Geman and Geman (1984). The Gibbs sampler (named after the physicist J. W. Gibbs) is widely used to sample from posterior probability distributions in applications of Bayesian inference.

A Gibbs sampling procedure is a procedure that obtains samples from the joint posterior of a set of random variables by successively updating sampled values of individual probabilistic variables conditional on the values of other variables. Concretely in our case, we are updating the multinomial word variable based on the letter variables and each letter variable based on the word variable and the appropriate position specific feature variables. We can see our procedure as sampling from the conditional distribution of the word variable based on the values of the feature and letter variables on each update at the word level, and as sampling from the conditional distribution of each letter variable, based on the values of the word and feature variables, on each update at the letter level. After burn-in, the overall state after each update is a sample from the joint distribution over all of the variables. The statement that such states are samples from the joint posterior means that the probability of visiting each state (at equilibrium, i.e., after a burn-in period) is equal to the posterior probability of the state.

Two properties of our sampling procedure are necessary to ensure that it accurately samples from the posterior (Hastings, 1970). First, the process must be *ergodic*—it must be possible to get from any state to any other state in a finite number of steps. Taking the feature units' values to be clamped, we are concerned only with states corresponding to a joint specification of a word and four possible letters. The process is ergodic if it is possible to get from any state of the word and letter units to any other state of these units. This property holds in our case, because all of the probabilities encoded in the weights are non-zero, making it possible (a) to visit any possible state of the letter units given an active word unit and a set of active feature values, and (b) to then visit any possible state of the word units given a set of active letter values. In our case, then, it is possible in principle to get from any

[19]Instead of goodness, Hinton and Sejnowski (1986), included a minus sign and called the quantity energy, following (Hopfield, 1982), but the equation is otherwise the same.

state to any other state in one update cycle, consisting of one letter update and one word update. So our model is ergodic[20].

The second critical property is that the updating process exhibits *detailed balance*. A stochastic updating process is said to have detailed balance with respect to a particular probability distribution $\{\pi\} = \{\ldots, \pi_i, \ldots, \pi_j \ldots\}$) over possible states if the probability of being in state $i$ and transitioning to state $j$ is equal to the probability of being in state $j$ and transitioning to state $i$:

$$\pi_i p(i \rightarrow j) = \pi_j p(j \rightarrow i),$$

or equivalently,

$$\frac{p(j \rightarrow i)}{p(i \rightarrow j)} = \frac{\pi_i}{\pi_j},$$

If a stochastic updating process has this property, it will converge to the equilibrium distribution $\{\pi\}$ in which the probabilities of states $i$ and $j$ are $\pi_i$ and $\pi_j$ respectively; ergodicity ensures that we can get to the equilibrium distribution from any starting state[21].

Intuitively, the detailed balance condition can be seen as a way of expressing what it means for a probability distribution to be at equilibrium, or stationary. Referring to the first version of the equation, we can read it as saying that at equilibrium, the probability of being in state $i$ and then transitioning to state $j$ should be equal to the probability of being in state $j$ and transitioning to state $i$. If this is true for all pairs of states, and if we can get from any state to any other state, then the distribution over states will stay constant as we continue to update. It is important to note that it is not the states themselves but the *distribution over states* that is stationary. On each update, the state may change, but the probability distribution over states, conceived of as the proportion of times an infinitely large ensemble of instances of the process is in each possible state, can still remain stationary. This is so because in the ensemble, the detailed balance condition stipulates that the probability of being in state $i$ and transitioning to state $j$ is equal to the probability of being in $j$ and transitioning to $i$.

We have just seen that if we are already at equilibrium (i.e., if the ratios of probabilities of particular states are equal to the ratios of the corresponding transition probabilities) we will stay there. But what if, at a certain time, the distribution of states is not yet at equilibrium? In that case, if the transition probability ratios are equal to the equilibrium probability ratios, the transitions will tend to move the distribution toward the stationary distribution. We will not prove this statement but we will consider an example a bit later showing how movement toward the correct stationary distribution does occur.

To show that our updating procedure will sample from the posterior distribution of the MIA model, we must show that its state transitions are balanced with respect to the posterior probabilities of the paths associated with these states, i.e., that the transition probabilities between states $i$ and $j$ are in balance with the posterior path probabilities. To do so, it is easier to work with the second version of the statement of the detailed balance condition. Working with this version, we would like to show that the ratio of the transition probabilities between any two states is equal to the ratio of the posterior probabilities of the generative paths corresponding to these states. Designating these states and the probabilities of the corresponding paths with the subscripts $i$ and $j$, this corresponds to the expression:

$$\frac{p(S_j \rightarrow S_i)}{p(S_i \rightarrow S_j)} = \frac{\pi_i}{\pi_j}.$$

For concreteness, let's consider a specific case. Suppose that the input features are the correct values of the features of the word TIME, and that the correct word is active at the word level, and the correct letter is active in positions 2, 3, and 4. Let state $S_I$ be the state in which, in addition to the above, the letter I is active in the first position and let state $S_T$ be the state in which, in addition to the above, the letter T is active in the first position, and let $\pi_I$ and $\pi_T$ represent the probabilities of the corresponding paths of the generative model. Using these indices, the above would then correspond to:

$$\frac{p(S_T \rightarrow S_I)}{p(S_I \rightarrow S_T)} = \frac{\pi_I}{\pi_T}.$$

Based on the visual similarity between I and T in the Rumelhart and Siple font, the paths associated with these states should be among the most probable, although state $I$ should be less probable that state $T$. Now, suppose we are in state $I$ and we are about to update the state of the first-position letter variable. We calculate the net input to each letter unit based on the active features and the active letters, and we then select the letter to activate according to the softmax function. The probability of transitioning to state $T$, i.e., of selecting $T$ as the next letter, is $\frac{e^{net_{T_1}}}{\sum_j e^{net_{j_1}}}$, where $net_{T_1}$, the net input to the unit for letter $T$ in position 1, is:

$$\log(p(l_{T_1}|w_{\text{TIME}})) + \sum_d \log(p(f_{v_{d_1}}|l_{T_1}))$$

so that $e^{net_{T_1}}$ is $p(l_{T_1}|w_{\text{TIME}}) \prod_d p(f_{v_{d_1}}|l_{T_1})$. Similarly, suppose we are in state $T$ and we are about to update the state of the first-position letter variable. We proceed as before, and find that the probability of transitioning to state $I$ is $\frac{e^{net_{I_1}}}{\sum_j e^{net_{j_1}}}$, where the net input to the unit for letter $I$ in position 1 is:

$$\log(p(l_{I_1}|w_{\text{TIME}})) + \sum_d \log(p(f_{v_{d_1}}|l_{I_1}))$$

and $e^{net_{I_1}}$ is $p(l_{I_1}|w_{\text{TIME}}) \prod_d p(f_{v_{d_1}}|l_{I_1})$. The ratio of these two transition probabilities, $\frac{p(S_I \rightarrow S_T)}{p(S_T \rightarrow S_I)}$ is then:

$$\frac{p(l_{T_1}|w_{\text{TIME}}) \prod_d p(f_{v_{d_1}}|l_{T_1})}{p(l_{I_1}|w_{\text{TIME}}) \prod_d p(f_{v_{d_1}}|l_{I_1})}$$

[20]It should be noted that some of the transition probabilities can be very small, and thus many of the transitions are highly unlikely. As we shall see below, we will not be relying on moving widely across the state space during processing of a single item.

[21]Being ergodic, as noted in footnote 18, is an in-principle matter, and some of the transition probabilities can be very small, but the starting state we start from—all units off except the clamped feature units—makes for easy access to all of the states that are plausible given the input.

They have the same denominator, which cancels out. This ratio is the same as the ratio of the posterior probabilities of each of the two paths, since the path probabilities share all of the other factors in common as well as the same denominator, and again everything else cancels out.

It would be tedious to repeat the above analysis for all possible pairs of states that might arise in the course of our sampling process. Luckily, there was nothing special about the particular case we just considered. The same argument can be applied for any pair of states differing only by the letter that is active in one of the four letter positions, given any set of clamped features. Furthermore, an analogous argument can be applied for any two states differing only by the word that is active. Since all the transitions are from one letter in a given position to another letter, or from one word to another word, this covers all of the transitions.

This completes the proof that the MIA model exhibits detailed balance, and we previously saw that it was ergodic. It follows, then, that the model samples states with probabilities corresponding to the posterior probabilities of the corresponding paths through the generative model.

In the context of the example we were working with above, we can now observe that the distribution of states tends to move toward the correct equilibrium distribution, at least in a simple specific case. Consider, for concreteness, an ensemble of 1000 separate instances of our network, and let an arbitrary fraction be in state $I$ and the rest be in state $T$ just before we update the multinomial variable for the first letter position in each of these 1000 instances. As one possibility, all of the networks could be in the $I$ state. Now, we note that our update procedure is unaffected by the previous active letter in the first letter position (it depends only on the state of the feature and word units—the other multinomial variables in the system). Relying on the same reasoning we worked through above, it should be clear that the update in each network will put the system in state $T$ with a probability proportional to $\pi_T = p(P_T)$, and in state $I$ with a probability proportional to $\pi_I = p(P_I)$, and thus the ratio of the proportion of networks in states $I$ and $T$ will tend toward $\frac{p(P_T)}{p(P_I)}$ after the update. Thus, in this case, we can move from a distribution far from equilibrium to something much closer to it in just a single update step. We have not considered what would happen in a full range of cases, but perhaps this example helps support the intuition that, in general, the distribution of states will tend to move toward the equilibrium distribution, if the transition probability ratios are in balance with the posterior path probabilities.

***A few practicalities.*** It is important to be aware of two things when using Gibbs sampling and related procedures. First, it takes some time for the settling process to reach the stationary distribution. It is difficult to know how many iterations of settling to allow before taking the state of the network as a valid sample from the stationary distribution, and testing for stationarity is not easy. Second, while in principle it is possible to transition from any state to any other state, in practice adjacent states tend to be correlated, and it may take a long time to make a transition between quite different, but equally good possibilities. For example, for the display in **Figure 8**, the time needed to transition from the interpretation [WORD = FEW; LETTERS = {F,E,W}] to the interpretation [WORD = HEX; LETTERS = {H,E,X}] may be quite long. It may, in fact, be quicker to get a set of decent samples by restarting from a blank starting place several times. This is how we proceeded to sample from the MIA model in Mirman et al. (in press). This is appropriate for our purposes, given that we think of each trial in a perceptual experiment as corresponding to a single case of settling to a perceptual interpretation, corresponding to a single sample from the posterior.

## SUMMARY AND DISCUSSION

The analysis presented above supports the assertion that interactive processing can be consistent with principled Bayesian computation. It is hoped that the analysis will lay to rest the in-principle concern about this matter. It is true that not all versions of interactive models can accurately capture Bayesian computations, but it should now be clear that at least some can. Many questions, or course, remain. In this section I will briefly consider two issues: First, what was wrong with the original IA model? Second, can some of the assumptions made in demonstrating that the MIA model can correctly sample form the posterior of the given generative model be relaxed, and still allow for proper probabilistic computations, or a good approximation to such computations?

### WHAT WAS WRONG WITH THE ORIGINAL IA MODEL?

Complaints about the adequacy of the original IA model (e.g., Massaro, 1989; Norris and McQueen, 2008) have centered on the bi-directional propagation of activation signals, but in fact, the original IA model failed to produce the pattern of logistic additivity one would expect even in the absence of interactive processing: the problem arose even when two sources of bottom-up evidence were combined (McClelland, 1991). This occurred because the particular activation and response selection assumptions used in the original IA model distorted the contributions of two different sources of evidence. Specifically, the original model applied the softmax function, not to the net inputs to units, but to activations of units—activations that had already been subjected to other non-linearities. These non-linearities did not prevent the model (or the TRACE model of speech perception) from capturing qualitatively a wide range of contextual influences on perception, but did contribute to the model's failure to exhibit logistic additivity.

Even if the problem with the original IA model's activation function were corrected, however, there could still be distortions of proper probabilistic computation in a deterministic model like the original IA model, as the IA model's critics claimed. To see this, consider first the following unidirectional model, which would not produce a distortion. In this model, we use the architecture and connection weight values of the MIA model. However, we make two changes: (a) we compute unit activations in the various layers of the model, setting them to continuous values based on the softmax function rather than selecting one to have an activation of 1 and all others to have an activation of 0; and (b) we allow only a unidirectional flow of processing, as in the unidirectional procedure described previously and depicted in **Figure 7A**. The activations so computed will correspond exactly to the probabilistic quantities that we could have computed directly—the net inputs, which are the sums of logs of relevant probabilistic quantities will be turned back into the relevant probabilistic quantities

by the exponentiation operation applied to the net input values in the softmax function. Now consider a version of this model, in which, instead of assumption (b), we allow word level activations to be computed based on letter level activations in all four letter positions, and we then send top-down signals back to the letter level from the word level based on all four letters instead of just three. This will clearly produce a distortion of the resulting activations, unless we take care (as Pearl did in his procedure) to divide back out of the top-down input to each letter position its own contribution to the activation at the word level. Based on these considerations, it appears that the original IA model may have failed to carry out proper Bayesian computations on two counts: it distorted these computations due to its basic activation assumptions and it distorted them due to its failure at lower levels to take back out its own contribution to the signals it received from higher levels.

### RELAXING SOME OF THE ASSUMPTIONS OF THE MIA MODEL

The MIA model makes some assumptions that were helpful in the analysis presented above. Among them are (1) we allowed just one unit to be active at a time in each pool corresponding to a multinomial random variable, (2) unit activation values are restricted to the values 0 and 1, and (3) units are updated according to a strict alternation schedule. None of these features are likely to hold in real neural networks. Would it still be possible to carry out proper probabilistic computations if some or all of these assumptions were relaxed? The exact limits of the conditions under which a (real or artificial) neural network could carry out proper Bayesian computations are not fully known, and further work will be required to further our understanding of this point. For now, I offer the conjecture that perhaps all of these assumptions can be relaxed, based on the following considerations.

First, in McClelland (1991), I presented simulations showing that logistic additivity of factors affecting stimulus and contextual influences on letter identification could be observed in three different variants of the original IA model, whereas the original model violated logistic additivity. Since logistic additivity of stimulus and context effects should be observed under the generative model underlying the MIA model, these findings are consistent with the conjecture above.

One of the three variants I considered in McClelland (1991) was a Boltzmann machine version of the original model. This variant is very similar to the MIA model, with these differences: (a) units within a pool are mutually inhibitory (there are negative connections between them) but they were not strictly mutually exclusive as in the MIA model and (b) unit activations were updated completely at random, as in the standard implementation of a Boltzmann machine. A mathematical analysis presented in McClelland (1991) demonstrated that logistic additivity follows from the assumptions of this model, and in McClelland (1998) I extended this analysis by showing that if the weights and bias terms in this variant of the model are set to the logs of the same probabilistic quantities used in the MIA model, then after settling to equilibrium at a temperature of 1, the relative probabilities of states with exactly one active word unit and exactly one active letter unit in each position would correspond to the relative posterior probabilities of the corresponding paths from the generative model.

The Boltzmann version of the IA model just considered still makes use of binary units. Could samples from the posterior still be obtained in models using continuous activation values for units in the neural network? It seems likely. The other two variants of the original IA model that exhibited logistic additivity in McClelland (1991) did use continuous activation values—in fact, these variants also retained the activation assumptions used in the original IA model. What differentiated these variants from the original model were the assumptions about sources of variability. In the original model, processing was completely deterministic and variability only affected response choices based on activations calculated deterministically, whereas in the two variants considered in McClelland (1991), variability was present either in the external inputs to the model or in the calculation of the net input to each of the units in the network. In both variants, the response choice after a period of settling was determined by selecting the most active unit within a mutually exclusive pool of units (e.g., the units for letters in one of the four letter positions). Yet another variant that used continuous activation values that also exhibits logistic additivity was presented in Movellan and McClelland (2001). These demonstrations of logistic additivity are largely based on simulations; proving that these variants produce logistic additivity is challenging, although some analysis under certain limiting conditions was provided for the third variant in Movellan and McClelland (2001). These findings are consistent with the conjecture that interactive networks that incorporate variability either in their inputs or intrinsic to processing can implement proper probabilistic computations. As previously stated, however, further analysis is required before we can definitively accept or reject this conjecture.

## CONCLUSION AND FUTURE DIRECTIONS

This article has covered a lot of ideas related to Bayesian inference, generative models, and neural networks. The primary goal was to review the ideas necessary to establish the proposition that interactive neural network models and principled probabilistic models of cognition can be compatible with each other. I hope that this review fulfills this goal, and I also hope that it will be of broader use. The probabilistic and neural network concepts considered here are in broad use throughout the psychological, cognitive science, and cognitive neuroscience literatures, and their integration should help advance our understanding of probabilistic computation in perception and its implementation in neural systems.

For the future, there is exciting work to be done. To date the MIA model has been used primarily to establish the basic theoretical point that interactive computations in neural networks are completely consistent with principled Bayesian computations. The ability of the model (or a successor) to capture specific patterns in data, such as those captured by the original IA model of letter perception (McClelland and Rumelhart, 1981) and to capture the many findings in the literature that were not adequately addressed by the original model [e.g., the time-course of stimulus and context effects, as observed in Massaro and Klitzke (1979)] remains to be explored [initial steps in this direction

were described in Khaitan and McClelland (2010)]. For that exploration, it will be necessary to develop, among other things, assumptions about exactly how the visual display conditions used in letter and word perception experiments affect activations of feature units and how this in turn affects the process of settling. Establishing more detailed links with the details of the underlying neurobiology will also be an important direction for the future.

I believe that incorporating learning and distributed representations will also be necessary to fully capture interactive processes in perception as they arise in naturalistic settings. We have seen in this article how an explicit generative model can be embedded in a perceptual system, so that it can sample from the generative model's posterior distribution. For this case, we have had the advantage of working in a domain—the domain of printed words and letters—where the relevant underlying units (the words and letters themselves) and contingent relations between them (letters depend on words, and features on letters)—can be identified, so that an explicit generative model (albeit oversimplified) can be advanced, and instantiated in a neural network. Real scenes that we are called upon to perceive are of course far more complex. There may be several objects in a display at the same time – so rather than a single underlying cause, there can be several. The underlying causes may be partially, but perhaps not completely, independent. The objects may take various poses and scales, be subject to various occluders and misleading lighting effects, etc. The objects themselves might not be fully characterized by mutually exclusive discrete identities, as words and letters are. To handle such cases, one can well imagine that no explicit generative model could ever do full justice to the actual entities or contingent probabilities involved.

A solution to this problem may involve using a network in which the units and connection weights are not pre-assigned, but learned. The network could still be viewed as instantiating a generative model, but without the prior stipulation of the correct set of units or connections. This is the approach taken in the deep belief networks introduced by Hinton and Salakhutdinov (2006). Incorporating these ideas into interactive models addressing the psychological and neural mechanisms of perception provides an exciting future challenge.

## ACKNOWLEDGMENTS

I would like to thank Tom Griffiths, Noah Goodman, Pranav Khaitan, Daniel Mirman, Tim Rogers, and the members of my lab at Stanford for useful comments and questions contributing to the development of this article.

## FUNDING

The author's effort while developing this article was partially supported by Air Force Research Laboratory Grant FA9550-07-1-0537.

## REFERENCES

Aderman, D., and Smith, E. E. (1971). Expectancy as a determinant of functional units in perceptual recognition. *Cogn. Psychol.* 2, 117–129. doi: 10.1016/0010-0285(71)90005-3

Bayes' theorem. (n.d.). In *Wikipedia*. Retrieved May 10, 2013.

Burton, A. M., Bruce, V., and Johnson, R. A. (1990). Understanding face recognition with an interactive activation model. *Br. J. Psychol.* 81, 361–380.

Dean, T. (2005). "A computational model of the cerebral cortex," in *Proceedings of AAAI-05* (Cambridge, MA: MIT Press), 938–943.

Derks, P. L., and Paclisanu, M. (1967). Simple strategies in binary prediction by children and adults. *J. Exp. Psychol.* 73, 278–285. doi: 10.1037/h0024137

Elman, J. L., and McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes. *J. Mem. Lang.* 27, 143–165. doi: 10.1016/0749-596X(88)90071-X

Freeman, J.B., and Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychol. Rev.* 118, 247–279. doi: 10.1037/a0022327

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *J. Exp. Psychol. Hum. Percept. Perform.* 6, 110–125. doi: 10.1037/0096-1523.6.1.110

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 721–741.

Grainger, J., and Jacobs, A. M. (1996). Orthographic processing in visual word recognition: a multiple read-out model. *Psychol. Rev.* 103, 518–565.

Green, C. S., Benson, C., Kersten, D., and Schrater, P. (2010). Alterations in choice behavior by manipulations of world model. *Proc. Natl. Acad. Sci. U.S.A.* 107, 16401–16406. doi: 10.1073/pnas.1001709107

Grossberg, S. A. (1978). "A theory of coding, memory, and development," in *Formal Theories of Visual Perception*, eds E. L. J. Leeuwenberg and H. F. J. M. Buffart (New York, NY: Wiley), 7–26

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109. doi: 10.1093/biomet/57.1.97

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hinton, G. E., and Sejnowski, T. J. (1983). "Optimal perceptual inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC: IEEE Press).

Hinton, G. E., and Sejnowski, T. J. (1986). "Learning and relearning in Boltzmann machines," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* Vol. I, Chap. 7, eds D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Cambridge, MA: MIT Press), 282–317.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554–2558.

Huey, E. B. (1908). *The Psychology and Pedagogy of Reading*. New York, NY: McMillan.

Johnston, J. C., and McClelland, J. L. (1973). Visual factors in word perception. *Percept. Psychophys.* 14, 365–370. doi: 10.3758/BF03212406

Johnston, J. C., and McClelland, J. L. (1974). Perception of letters in words: seek not and ye shall find. *Science* 184, 1192–1194. doi: 10.1126/science.184.4142.1192

Khaitan, P., and McClelland, J. L. (2010). "Matching exact posterior probabilities in the Multinomial Interactive Activation Model," in *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, eds S. Ohlsson and R. Catrambone (Austin, TX: Cognitive Science Society), 623.

Kumaran, D., and McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychol. Rev.* 119, 573–616. doi: 10.1037/a0028681

Lee, T. S., and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* 20, 1434–1448. doi: 10.1364/JOSAA.20.001434

Luce, R. D. (1959). *Individual Choice Behavior*. New York, NY: Wiley.

Massaro, D. W. (1979). Letter information and orthographic context in word perception. *J. Exp. Psychol. Hum. Percept. Perform.* 5, 595–609. doi: 10.1037/0096-1523.5.4.595

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cogn. Psychol.* 21, 398–421. doi: 10.1016/0010-0285(89)90014-5

Massaro, D. W., and Cohen, M. M. (1991). Integration versus interactive activation: the joint influence of stimulus and context in perception. *Cogn. Psychol.* 23, 558–614. doi: 10.1016/0010-0285(91)90006-A

Massaro, D. W., and Klitzke, D. (1979). The role of lateral masking and orthographic structure in letter and word perception. *Acta Psychol.* 43, 413–426. doi: 10.1016/0001-6918(79)90033-7

McClelland, J. L. (1981). "Retrieving general and specific information from stored knowledge of specifics," in *Proceedings of the Third Annual Conference of the Cognitive Science Society*, (Berkeley, CA), 170–172.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cogn. Psychol.* 23, 1–44. doi: 10.1016/0010-0285(91)90002-6

McClelland, J. L. (1998). "Connectionist models and Bayesian inference," in *Rational Models of Cognition*, eds M. Oaksford and N. Chater (Oxford: Oxford University Press), 21–53.

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

McClelland, J. L. Mirman, D., and Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends Cogn. Sci.* 10, 363–369.

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I an account of basic findings. *Psychol. Rev.* 88, 375–407. doi: 10.1037/0033-295X.88.5.375

McCulloch, W., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 7, 115–133. doi: 10.1007/BF02478259

Miller, G. , Heise, G., and Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *J. Exp. Psychol.* 41, 329–335. doi: 10.1037/h0062491

Mirman, D., Bolger, D. J., Khaitan, P., and McClelland, J. L. (in press). Interactive activation and mutual constraint satisfaction. *Cogn. Sci.*

Morton, J. (1969). The interaction of information in word recognition. *Psychol. Rev.* 76, 165–178. doi: 10.1037/h0027366

Movellan, J. R., and McClelland, J. L. (2001). The Morton-Massaro Law of Information Integration: implications for models of perception. *Psychol. Rev.* 108, 113–148. doi: 10.1037/0033-295X.108.1.113

Norris, D., and McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychol. Rev.* 115, 357–395. doi: 10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., and Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behav. Brain Sci.* 23, 299–370. doi: 10.1017/S0140525X00003241

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *J. Exp. Psychol. Learn. Mem. Cogn.* 10, 104–114. doi: 10.1037/0278-7393.10.1.104

Pearl, J. (1982). "Reverend Bayes on inference engines: a distributed hierarchical approach," in *Proceedings of AAAI-82* (Palo Alto, CA: MIT Press), 133–136.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *J. Exp. Psychol.* 81, 274–280. doi: 10.1037/h0027768

Rumelhart, D. E. (1977). "Toward an interactive model of reading," in *Attention and Performance VI,* Chap. 27, ed S. Dornic (Hillsdale, NJ: LEA), 573–603.

Rumelhart, D. E., and McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychol. Rev.* 89, 60–94.

Rumelhart, D. E., and Siple, P. (1974). The process of recognizing tachistoscopically presented words. *Psychol. Rev.* 81, 99–118. doi: 10.1037/h0036117

Salzman, C. D., and Newsome, W. T. (1994). Neural mechanisms for forming a perceptual decision. *Science* 264, 231–237. doi: 10.1126/science.8146653

Thompson, M. C., and Massaro, D. W. (1973). Visual information and redundancy in reading. *J. Exp. Psychol.* 98, 49–54. doi: 10.1037/h0034308

Train, K. (1993). *Qualitative Choice Analysis: Theory, Econometrics, and Application to Automobile Demand*. Cambridge, MA: MIT Press.

Tulving, E., Mandler, G., and Baumal, R. (1964). Interaction of two sources of information in tachistoscopic word recognition. *Can. J. Psychol.* 18, 62–71.

Warren, R. M., and Sherman, G. (1974). Phonemic restorations based on subsequent context. *Percept. Psychophys.* 16, 150–156.

Wheeler, D. (1970). Processes in word recognition. *Cogn. Psychol.* 1, 59–75. doi: 10.1016/0010-0285(70)90005-8

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 12 May 2013; paper pending published: 13 July 2013; accepted: 17 July 2013; published online: 20 August 2013.*
*Citation: McClelland JL (2013) Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review. Front. Psychol. **4**:503. doi: 10.3389/fpsyg.2013.00503*
*This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*

frontiers in PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 20 August 2013
doi: 10.3389/fpsyg.2013.00515

# Modeling language and cognition with deep unsupervised learning: a tutorial overview

***Marco Zorzi*[1,2]*, *Alberto Testolin*[1] *and Ivilin P. Stoianov*[1,3]**

[1] *Computational Cognitive Neuroscience Lab, Department of General Psychology, University of Padova, Padova, Italy*
[2] *IRCCS San Camillo Neurorehabilitation Hospital, Venice-Lido, Italy*
[3] *Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy*

***Edited by:***
*Julien Mayor, University of Geneva, Switzerland*

***Reviewed by:***
*Bradley Love, University College London, UK*
*Angelo Cangelosi, University of Plymouth, UK*

****Correspondence:***
*Marco Zorzi, Computational Cognitive Neuroscience Lab, Department of General Psychology, University of Padova, Via Venezia 12, Padova 35131, Italy*
*e-mail: marco.zorzi@unipd.it*

Deep unsupervised learning in stochastic recurrent neural networks with many layers of hidden units is a recent breakthrough in neural computation research. These networks build a hierarchy of progressively more complex distributed representations of the sensory data by fitting a hierarchical generative model. In this article we discuss the theoretical foundations of this approach and we review key issues related to training, testing and analysis of deep networks for modeling language and cognitive processing. The classic letter and word perception problem of McClelland and Rumelhart (1981) is used as a tutorial example to illustrate how structured and abstract representations may emerge from deep generative learning. We argue that the focus on deep architectures and generative (rather than discriminative) learning represents a crucial step forward for the connectionist modeling enterprise, because it offers a more plausible model of cortical learning as well as a way to bridge the gap between emergentist connectionist models and structured Bayesian models of cognition.

**Keywords: neural networks, connectionist modeling, deep learning, hierarchical generative models, unsupervised learning, visual word recognition**

## INTRODUCTION

A fundamental issue in the study of human cognition is what computations are carried out by the brain to implement cognitive processes. The connectionist framework assumes that cognitive processes are implemented in terms of complex, non-linear interactions among a large number of simple, neuron-like processing units that form a neural network (Rumelhart and McClelland, 1986). This approach has been used in cognitive psychology—often with success—to develop functional models that clearly represent a great advance over previous verbal-diagrammatic models because they can produce simulations of learning, skilled performance, and breakdowns of processing after brain damage. One paradigmatic example is the connectionist modeling of visual word recognition and reading aloud, which has often provided key theoretical and methodological advances with broad influences well-beyond the language domain (e.g., McClelland and Rumelhart, 1981; Seidenberg and McClelland, 1989; Plaut and Shallice, 1993; Plaut et al., 1996). Connectionist models of the reading processes can produce highly detailed simulations of human performance, accounting for a wide range of empirical data that include reaction times and accuracy of skilled readers at the level of individual words, the development of reading skills in children, and the impaired performance of dyslexic individuals (Plaut et al., 1996; Zorzi et al., 1998; Harm and Seidenberg, 1999, 2004; Perry et al., 2007, 2010, 2013). Despite significant progress in the attempt to improve the architectural and learning principles incorporated in neural network models (see O'Reilly, 1998; O'Reilly and Munakata, 2000), much modeling work in psychology is still based on the classic neural network with one layer of hidden units (i.e., a "shallow" architecture) and error backpropagation (Rumelhart et al., 1986) as learning algorithm—a choice that is typically seen as a compromise to achieve efficient learning of complex cognitive tasks. We argue below that a key step forward for connectionist modeling is the use of networks with a "deep" architecture (Hinton, 2007, 2013) and where most of the learning is generative rather than discriminative (**Box 1**).

The shallow architecture of the prototypical multi-layer neural network (Rumelhart et al., 1986) does not capture the hierarchical organization of the cerebral cortex. Hierarchical processing is thought to be a fundamental characteristic of cortical computation (Hinton, 2007; Clark, 2013) and it is a key feature of biologically inspired computational models of vision (Riesenhuber and Poggio, 1999). The idea of a deep network with a hierarchy of increasingly complex feature detectors can be traced back to the Interactive Activation Model (IAM) of letter and word perception (McClelland and Rumelhart, 1981), but this seminal proposal did not transfer to connectionist learning models because the error backpropagation algorithm had little success in training networks with many hidden layers (Hinton, 2007, 2013). Another key assumption of the IAM that did not readily transfer to connectionist learning models is the mixing of bottom–up and top–down processing through recurrent feedback. Finally, the widespread use of the error backpropagation algorithm in connectionist modeling, leaving aside its lack of biological plausibility (O'Reilly, 1998), implies subscription to the dubious assumption that learning is largely discriminative (e.g., classification or function learning) and that an external teaching signal is available at each learning event (that is, all training data is labeled). This

**Box 1 | Glossary.**

**BOLTZMANN MACHINE**
Stochastic neural network of symmetrically connected, neuron-like units whose dynamics is governed by an energy function. The input to the network is given through a layer of visible units, while another layer of hidden units is used to model the latent causes of the data. A variant known as Restricted Boltzmann Machine (RBM) is obtained by removing within-layer lateral connections to form a bipartite graph, allowing to perform efficient inference and learning.

**CONTRASTIVE DIVERGENCE**
Objective function that allows to efficiently train RBMs by approximating the log-likelihood gradient, without requiring to run a Markov chain to convergence.

**DEEP BELIEF NETWORK**
Hierarchical generative model composed of a stack of RBMs, which can be greedily trained layer-wise in an unsupervised fashion. The whole network can be eventually fine-tuned with supervised learning to perform discriminative tasks.

**DEEP LEARNING**
Machine learning framework that exploits multiple layers of hidden units to build hierarchical internal representations of the input data.

**DISCRIMINATIVE LEARNING**
Learning approach whose objective is to map the observed variables $X$ into corresponding output variables $Y$, usually by modeling the conditional distribution $P(Y|X)$, optimizing classification boundaries, or by approximating a function $Y = f(X)$. This approach requires labeled examples (i.e, a teaching signal for supervised learning).

**GENERATIVE LEARNING**
Learning approach whose objective is to model the joint distribution $P(X, Y)$ of observed and latent variables, typically using a likelihood-based criterion. This approach does not require labeled data (i.e., learning is unsupervised).

**GRAPHICAL MODELS**
Probabilistic models in which the topology of a graph defines conditional independecies between random variables, allowing to efficiently represent complex joint distributions through factorization.

learning regimen is exceptional in the real world. Reinforcement learning (Sutton and Barto, 1998) is a plausible alternative, but there is a broad range of situations where learning is fully unsupervised and its only objective is that of building rich internal representations of the sensory world (Hinton and Sejnowski, 1999). Notably, the learned internal model can then be used to infer causes and make predictions (Dayan et al., 1995; Hinton and Ghahramani, 1997; Friston, 2005; Hinton, 2010b; Huang and Rao, 2011; Clark, 2013).

Unsupervised learning has a long history, but the classic learning algorithms have important limitations. Some develop a representation that is distributed but also linear (Oja, 1982), which implies that higher-order information remains invisible. Others develop a representation that is non-linear but also localist, that is one in which each observation is associated to a single hidden unit (Rumelhart and Zipser, 1985; Kohonen, 1990). For these reasons, their application to modeling complex cognitive functions has been limited. An important breakthrough in unsupervised learning is the use of statistical principles such as maximum likelihood and Bayesian estimation to develop generative models that discover representations that are both distributed and non-linearly related to the input data (Hinton and Ghahramani, 1997). A generative model is a probabilistic model that captures the hidden (latent) causes of the data, thereby providing a sensible objective function for unsupervised learning. In other words, the "learner" estimates a model, without any supervision or reward, that represents the probability distribution of the data. Generative models are appealing because they make strong suggestions about the role of feedback connections in the cortex and are consistent with neurobiological theories that emphasize the mixing of bottom–up and top–down interactions in the brain: bottom–up inputs convey sensory information, whereas internal representations form a generative model that predicts the sensory input via top–down activation (Hinton and Ghahramani, 1997). Learning can be viewed as maximizing the likelihood of the observed data under the generative model, which is equivalent to discovering efficient ways of coding the sensory data (Ghahramani et al., 1999). Notably, the application of these algorithms to natural images has been shown to generate receptive field properties similar to those observed in the visual cortex (Rao and Ballard, 1999).

Generative learning can be implemented in the framework of recurrent stochastic neural networks with hidden units (Hinton, 2002). However, one hidden layer can be insufficient for modeling structured and high-dimensional sensory data. In contrast, a network with many hidden layers, that is a deep network, can learn a more powerful *hierarchical generative model* (Hinton and Salakhutdinov, 2006; Hinton et al., 2006). Note that a good generative model of the data can be a very useful starting point for later discriminative learning (Hinton, 2007; Stoianov and Zorzi, 2012). The internal representations obtained from generative learning can be the input to a variety of classification or function learning tasks, thereby exploiting re-use of learned features (Bengio et al., 2012). Moreover, the internal model might be refined through supervised learning to strengthen the features that are most informative for solving a specific classification task (Hinton and Salakhutdinov, 2006; also see Love et al., 2004, for a related modeling approach to category learning). Indeed, it has been shown that human category learning implies flexibility in the use and creation of perceptual features (Schyns et al., 1998) and that different types of features might be extracted according to the nature of the learning task (e.g., unsupervised vs. supervised; Love, 2002).

The goal of the present article is to provide a tutorial overview of generative learning in deep neural networks to highlight its appeal for modeling language and cognition. We start with a brief review of the theoretical foundations of generative

learning and deep networks. We then discuss various practical aspects related to training, testing and analyzing deep networks, using the classic letter and word perception problem of McClelland and Rumelhart (1981) as a tutorial example. The emergence of a hierarchy of orthographic representations through deep unsupervised learning is particularly interesting (also see Di Bono and Zorzi, under review) because it can revisit the hard-wired architecture of the IAM. The idea that perception of written words involves the sensitivity to increasingly larger orthographic units is also supported by recent neuroimaging findings (Dehaene et al., 2005; Vinckier et al., 2007).

## LEARNING A GENERATIVE MODEL: RESTRICTED BOLTZMANN MACHINES

Here we consider a class of neural networks known as Boltzmann Machines (hereafter BM; Ackley et al., 1985). These are stochastic associative networks that observe and model data by using local signals only. BMs can be interpreted as undirected graphical models (Jordan and Sejnowski, 2001; see **Box 2**) where learning corresponds to fitting a generative model to the data. Despite the appeal of BMs as plausible models of cortical learning, their use was strongly discouraged by the very high computational demand of the original learning algorithm, until the recent development of *contrastive divergence (CD) learning* (Hinton, 2002). CD makes learning of BMs practical, even for large networks (see below).

BMs consist of a set of stochastic units, fully connected with symmetric weights and without self-connections, where each unit fires with a probability depending on the weighted sum of its inputs. Data patterns are represented by the activation of "visible" units. An additional layer of "hidden" units captures high-order statistics and represent the latent causes of the data. Inspired by statistical mechanics, the model behavior is driven by an energy function $E$ that describes which configurations of the units are more likely to occur by assigning them a certain probability value:

$$p(v,\ h) = \frac{e^{-E(v,\ h)}}{Z}$$

where $v$ and $h$ are, respectively, the visible and hidden units and $Z$ is a normalizing factor known as *partition function*, which ensures that the values of $p$ constitute a legal probability distribution (i.e., summing up to 1). The network state changes in a way that allows the gradual decrease of the associated energy, modulated by a "temperature" parameter $T$ so that at higher temperatures an occasional increase of energy is also permitted to avoid local minima. To achieve local energy minimum (equilibrium), $T$ is

**Box 2 | Probabilistic Graphical Models.**

The framework of probabilistic graphical models (Koller and Friedman, 2009) provides a general approach to model arbitrarily complex statistical distributions, which can involve a large number of stochastic variables that interact together. Graphical models allow us to describe complex relations between variables by exploiting the *structure* of their joint distribution, since in general their interactions are not globally defined but instead each variable is only influenced by a limited subset of "neighbors." The topology of a graphical model explicitly defines the scope of interaction of each variable (represented by a node in the graph) by highlighting the set of independecies that hold in the distribution. This allows to *factorize* a joint probability distribution using local conditional probabilities.

Graphical models can have *directed* connections between variables, such as in Bayesian networks (**Figure 1A**), or *undirected* connections, such as in Markov networks (**Figure 1B**). Both types of connections might be present in the same graph, thus forming a *hybrid* model. Although they share the same underlying theoretical framework, Bayesian and Markov networks have rather different representational and computational characteristics. In directed models, the semantic of connections defines a "parent of" relationships between linked variables, while in undirected models the connections are symmetric and therefore only encode a sort of "degree of affinity" between linked variables. This leads to a different representation of independencies between nodes of the graph: in Bayesian networks, each node is conditionally independent from all the others given its parents, its children and the parents of its children, while in undirected models each node is conditionally independent from all the others given the nodes directly connected to it [i.e., its "Markov blanket" (Pearl, 1988), highlighted in **Figure 1**]. In both cases, these conditional independencies can be exploited to derive efficient inference and learning procedures even in the presence of a large number of variables, because only the Markov blanket of a certain node is required in order to sample from its conditional distribution.

In the case of undirected graphical models, each edge is associated with a certain function, known as *factor*, which takes as input the values of the nodes connected by the edge and gives as output a scalar value that represents the affinity between them: a high value indicates that the two variables are likely to be strongly related, while a low value indicates a weak relation. The joint distribution of all the variables in the graph can be efficiently defined as a product of such local factors:

$$P(X_1, X_2, \ldots, X_n) = \frac{1}{Z}\prod_i \phi_i(D_i)$$

where $D_i$ represents the scope of each factor $\phi_i$ (i.e., which variables it involves) and $Z$ is a global normalization constant called *partition function*, which ensures dealing with legal probabilities summing up to 1.

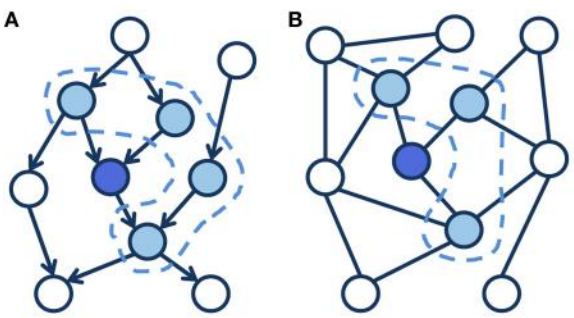



**FIGURE 1 | (A)** A directed graphical model, also known as Bayesian network. **(B)** An undirected graphical model, also known as Markov network. In both graphs, the dashed line highlights the Markov blanket of the blue node.

gradually decreased (*simulated annealing*). The learning procedure minimizes the Kullback-Liebler divergence between the data distribution and the model distribution. Accordingly, for each pattern the network performs a data-driven, *positive phase* (+) and a model-driven, *negative phase* (–). In the positive phase the visible units are clamped to the current pattern and the hidden layer settles to a stable activation state. In the negative phase all units are unclamped and the network is run [using a Markov Chain Monte Carlo (MCMC) algorithm; see **Box 3**] until it settles on a stable activation state over visible and hidden units, which reflects the model beliefs. After each phase, correlations between the activations of each pair of connected units are collected and used to update the network weights. Note that learning is unsupervised (i.e., the network does not learn an input–output mapping like typical multilayer networks trained with error backpropagation) and it uses only local signals and Hebbian rules. A similar form of contrastive Hebbian learning is also used in the generalized recirculation algorithm and in Leabra (O'Reilly, 1998, 2001). Learning the connection weights in the original BM is based on a maximum likelihood learning rule that is very simple and locally optimal, but unfortunately the learning algorithm is also very slow because it implies running a Markov chain until convergence (which may require an exponential time).

The breakthrough that led to CD learning (Hinton, 2002; also see Welling and Hinton, 2002; for a mean field version) is the finding that the negative phase does not need to be run until equilibrium (i.e., full convergence). If sampling starts from the hidden unit state computed in the positive phase (i.e., driven by the data), correlations computed after a fixed number of steps in the Markov chain are sufficient to drive the weights toward a state in which the input data will be accurately reconstructed. Hence, CD learning approximates the gradient of the log-likelihood of the learning data by performing only few iterations, which in practice gives good results even with a single step (*CD-1*). After computing the model's reconstruction, weights are updated by contrasting visible-hidden correlations computed on the data vector ($v^+h^+$) with visible-hidden correlations computed on the reconstruction ($v^-h^-$):

$$\Delta W = \eta(v^+h^+ - v^-h^-)$$

where $\eta$ is the learning rate. Importantly, a restriction to the architecture of the BM by not allowing intra-layer connections (RBM; Hinton, 2002) makes learning extremely fast. The energy function for RBMs is defined as:

$$E(v, h) = -b^T v - c^T h - h^T W v$$

where $W$ is the matrix of connections weights and $b$ and $c$ are the biases of visible and hidden units, respectively. In RBMs, the update of units in one layer no longer requires any iterative settling because they are conditionally independent given the state of the other layer. That is, the sampling process is speeded up by performing block Gibbs sampling (see **Box 3**) over visible and hidden units (i.e., all units in a layer are sampled in a single step).

Examples of application of CD learning in connectionist modeling studies include numerical cognition (Stoianov et al., 2002, 2004; Zorzi et al., 2005) and space coding for sensorimotor transformations (De Filippo De Grazia et al., 2012).

**Box 3 | Block Gibbs sampling in RBMs.**

In a probabilistic graphical model, we are often interested in generating samples from the model distribution. A general-purpose, powerful method is the Gibbs sampling algorithm, which generates a sequence of observations that progressively approximate a specified multivariate probability distribution (Geman and Geman, 1984). Gibbs sampling belongs to the family of MCMC methods, which draw samples from a probability distribution by constructing a Markov chain that has the desired distribution as its equilibrium distribution (Andrieu et al., 2003). Under certain conditions, after an initial *burn-in* phase the Markov chain will converge to the stable distribution. The basic idea of Gibbs sampling is to construct the Markov chain so that one particular variable is sampled at each step *given* the current values of *all the other variables*. After repeating this process iteratively for enough time, the chain will generate samples from the target joint distribution. Notably, Gibbs sampling can exploit the structure of the graph (i.e., the conditional independecies between variables) to speed up this process: since the value of each node is only influenced by its Markov blanket (see **Box 2**), if two variables are conditionally independent given the current evidence (i.e., their Markov blanket is observed) they can be sampled at the same time. This variant of the algorithm is known as *block Gibbs sampling*.

In the case of Boltzmann Machines, learning requires sampling from the joint distribution of visible and hidden variables in order to compute visible-hidden correlations on the model expectations. If the connectivity of the network is restricted, as in the RBM, the sampling process can be significantly speeded up by using block Gibbs sampling. Indeed, the units of the same layer become conditionally independent if there are no intra-layer connections; that is, in RBMs the Markov blanket of a hidden unit corresponds to the visible layer, and vice versa (**Figure 2**). This allows to sample all units of the same layer in parallel.

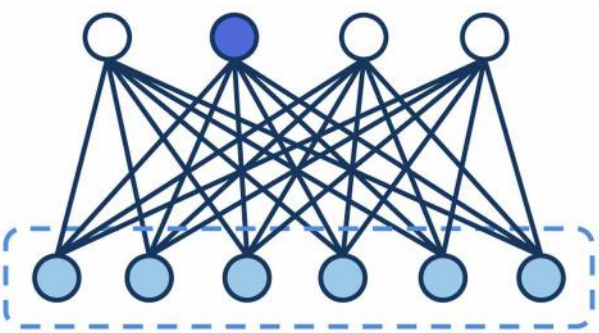


**FIGURE 2 | Graphical representation of a Restricted Boltzmann Machine.** The dashed line highlights the Markov blanket of the blue hidden unit, which corresponds to the whole layer of visible units.

## LEARNING A HIERARCHICAL GENERATIVE MODEL: DEEP BELIEF NETWORKS

RBMs can be used as building blocks of more complex architectures, where the hidden variables of the generative model can be organized into layers of a hierarchy (**Figure 3A**). The resulting architecture is referred to as a "deep network." In particular, the Deep Belief Network (DBN; Hinton and Salakhutdinov, 2006; Hinton et al., 2006) is a stack of RBMs that can be trained layer by layer in a greedy, unsupervised way. The main intuition behind deep learning is that, by training a generative model at level *l* using as input the hidden causes discovered at level *l–1*, the network will progressively build more structured and abstract representations of the input data. Importantly, architectures with multiple processing levels permit an efficient encoding of information by exploiting re-use of features among different layers: simple features extracted at lower levels can be successively combined to create more complex features, which will eventually unravel the main causal factors underlying the data distribution. Indeed, it has been shown that functions that can be compactly represented by a depth *k* architecture might require an exponential number of computational elements to be represented by a depth *k–1* architecture (Bengio, 2009). Moreover, adding a new layer to the architecture increases a lower bound on the log-likelihood of the generative model (Hinton et al., 2006), thus improving the overall capacity of the network. After learning of all layers, the deep architecture can be used as a generative model by reproducing the data when sampling from the model, that is by feeding the activations of the deepest layer all the way back to the input layer. Note that the hierarchical structure of the internal representations is an emergent property of the learning algorithm. In contrast, hierarchy in classic connectionist models is typically built in by stipulating the representations to be used at more than one layer (e.g., Rumelhart and Todd, 1993; Perry et al., 2013); indeed, training of deep multi-layer perceptrons using error backpropagation is very difficult because the error gradient tends to vanish when propagated backwards through more than one hidden layer (see Hinton, 2013, for further discussion).

An important advantage of deep unsupervised learning is that the internal representations discovered by the network are not tied to a particular discriminative task, because the objective of learning is only to model the hidden causes of the data. However, once the system has developed expressive abstract representations, possible supervised tasks can be carried out by introducing additional modules, which directly operate on such high-level representations of the data and can therefore yield excellent performance in classification or function learning (**Figure 3B**). For example, on a popular handwritten digit recognition problem (MNIST dataset; LeCun et al., 1998), high discriminative accuracy can be obtained even by a linear classifier applied on the top-level internal representations of a DBN that was only trained to reconstruct the digit images (Testolin et al., 2013; examples of digits reconstructed by the network are reported in **Figure 3C**). Within this perspective, the use of an additional fine-tuning phase of the whole deep network using error backpropagation (as done in Hinton and Salakhutdinov, 2006) might be unwarranted, not only because of the biological implausibility of the learning algorithm, but also because the network would become specifically tuned to a particular task. Indeed, the idea that high-level representations obtained from (unsupervised) model learning should be usable across several tasks (**Figure 3B**) is referred to as "transfer learning" and it is a hot topic for the machine learning community (Bengio, 2009; Bengio et al., 2012). It is worth mentioning that machine learning researchers have recently investigated deep networks built through greedy layer-wise training of stacked autoencoders, where each autoencoder is a multilayer perceptron trained to auto-associate the input (Bengio and Lamblin, 2007; Baldi, 2012). This approach has been successful in terms of machine learning benchmarks, but it is less appealing than DBNs for cognitive modeling purposes because learning is based on error backpropagation and it is not grounded in a

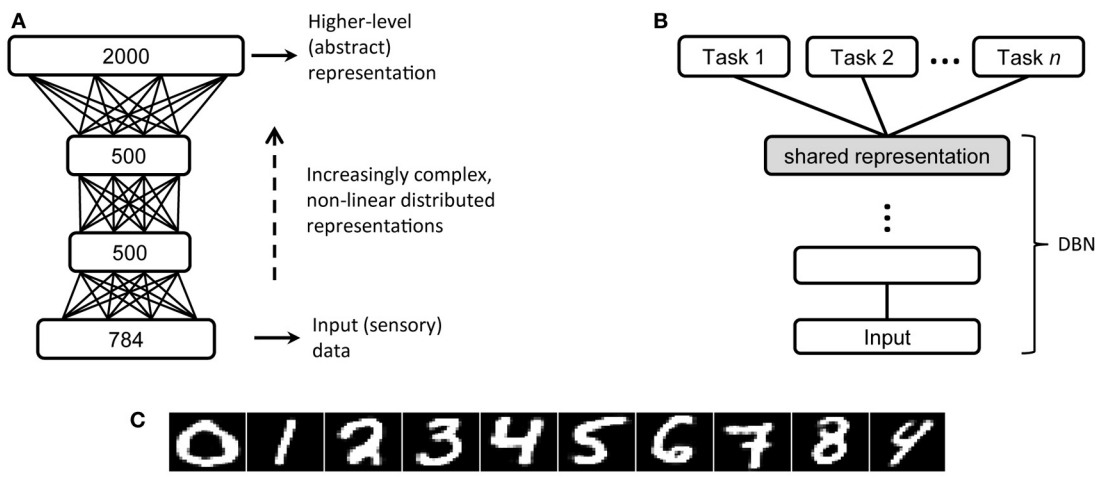



**FIGURE 3 | (A)** Architecture of the DBN with three hidden layers used in the MNIST handwritten digit recognition problem (Hinton and Salakhutdinov, 2006). **(B)** A typical transfer learning scenario, on which high-level, abstract representations are first extracted by deep unsupervised learning and then used to perform a variety of supervised tasks [adapted from Bengio et al. (2012)]. **(C)** Reconstructions of MNIST digit images made by the deep network.

sound probabilistic framework. Moreover, deep autoencoders are not used as generative models to produce predictions based on top–down signals.

A final consideration concerns the computational complexity of deep learning: thanks to its efficiency, the algorithm proposed by Hinton et al. (2006) solves the problem of learning in densely connected networks that have many hidden layers. If implemented on multicore hardware, deep learning is practical even with *billions* of connections, thereby allowing the development of very-large-scale simulations (Raina et al., 2009; Dean et al., 2012; Le et al., 2012). Medium-to-large-scale simulations can even be performed on a desktop PC equipped with a low-cost graphic card (Testolin et al., 2013; see below).

## CONNECTIONIST MODELING WITH DEEP NETWORKS: A TUTORIAL

In this section we provide a practical overview on how to construct a complete DBN simulation. We illustrate how to train, test and analyze a deep network model using the classic letter and word perception problem of McClelland and Rumelhart (1981). Written word perception is particularly representative because it can be linked to one of the most influential models of language processing, McClelland and Rumehart's IAM, and more specifically to its two key assumptions: (1) a hierarchical organization of the network, with increasingly more complex levels of representation, and (2) the mixing of bottom–up and top–down processing (i.e., interactivity) to resolve ambiguity of the sensory input. Interestingly, a recent re-formulation of the IAM as a probabilistic generative model (Khaitan and McClelland, 2010) was shown to perform optimal Bayesian inference, thereby supporting the appeal of the hierarchical interactive architecture (Mirman et al., in press). A deep learning model would therefore represent an important step forward, because the hard-wired architecture of the IAM might be replaced by the hierarchical generative model learned in a DBN. In this regard, learning word perception can be seen as a stochastic inference problem where the goal is to estimate the posterior distribution over latent variables given the image of a word as input.

Though written word perception is an excellent candidate for deep learning, the complexity of the problem makes realistic simulations difficult to handle. For example, high-resolution images of whole words would require a very large network, with tens of thousands of visible units (e.g., 20,000 units for a 400 by 50 pixels image), many hidden layers and billions of connections (see Krizhevsky et al., 2012, for deep learning on a realistic object recognition problem). One possible simplification would be to split words into letter constituents and first model the perception of single letters. This might lead to sensible internal letter representations that are invariant to position, size, rotation, and noise (i.e., abstract letter identities; McClelland and Rumelhart, 1981). Alternatively, written words can be represented using small resolution images, with letters encoded as combinations of simple geometric features (the "Siple" font; McClelland and Rumelhart, 1981). We employed the latter solution for the simulations presented here.

In this tutorial we also consider deep learning of handwritten digits (MNIST database; LeCun et al., 1998) and visual numerosity estimation (Stoianov and Zorzi, 2012) in relation to the analysis of DBNs, because they represent more realistic perception problems that involve training on thousands of images. Training on a large dataset can be important for the emergence of a richer hierarchical structure of features.

### TRAINING A DBN

As in other connectionist models, input to the network is provided as pattern of activations over visible units. Note that 2D images are vectorized; this implies that the spatial structure remains only implicit in the co-activation of neighboring visible units, but it can emerge during learning in the form of statistical regularities (see examples below). Learning a generative model does not require labeled data, that is, unlike supervised learning, each pattern does not need to possess a class label or any other form of associated target state. Nevertheless, this kind of information might still be useful for testing and analyzing the network. Note that realistic, large-scale simulations often imply abundance of unlabeled data and only a limited sample of pre-classified learning examples (see Le et al., 2012, for deep learning on millions of images randomly extracted from videos on the Internet).

A ready-to-use parallel implementation of deep unsupervised learning on graphic cards is described in Testolin et al. (2013), and it is publicly available for download[1].

#### *Network architecture*

The learning algorithm tunes the parameters (i.e., weights) of a DBN with a given structure that should be specified after establishing the input domain. Here we only consider network architectures with fully connected pairs of layers (**Figure 3A**), but alternatives based on weights sharing like convolutional networks (LeCun et al., 1998) can simplify the learning problem by assuming identical processing applied to different portions of the image, thereby reducing the number of parameters of the model. In general, the size of a given hidden layer might be proportional to the expected number of features describing the data at a certain processing level. Intuitively, many hidden units will allow for the encoding of more specific characteristics of the data, whereas fewer units imply a greater compression of the representation and hence increase the generality of the features. A more neutral strategy with regard to the architectural choices is to keep the size of few consecutive layers constant. Finally, a large top hidden layer can be useful to unfold categories and classes, thereby facilitating linear associations to categories or other processing domains (as we will discuss in the following sections). At any rate, we advise to try several architectures, gradually increasing the number of layers and units per layer, until satisfactory results are obtained.

#### *Learning tasks*

We illustrate the tutorial with examples of increasing complexity. The first toy example is the visual perception of single letters with input consisting of black and white (b/w) images of size

[1]A variety of multicore implementations (MATLAB and Python on graphic cards; Octave/MPI on a multi-core cluster) is described in Testolin et al. (2013) and the source codes can be found at: http://ccnl.psy.unipd.it/research/deeplearning

7 × 7 pixels (i.e., patterns over 49 visible units). The dataset contains the images of 26 capital letters created with the schematic "Siple" font, composed of 14 basic visual features (Rumelhart and Siple, 1974). We found that a small two-layer DBN network with as few as 10 units in the first layer and 30 units in the second layer was sufficient to discover the underlying visual features. The second example extends the problem above to the visual perception of four-letter words, using the classic dataset of 1180 words employed by McClelland and Rumelhart (1981) in the IAM. Input are b/w images of size 28 × 7 pixels (i.e., patterns over 196 visible units) of words printed with the Siple font. This problem required a DBN with more hidden units: 120 in the first hidden layer and 200 in the second one (see **Figure 4**).

Two additional examples approach realistic problems: the perception of handwritten digits and visual numerosity perception. The training datasets for these problems contain thousands of samples per category (i.e., digits or numerosity levels) and provide a rich variety of different instances. In the handwritten digit recognition problem, input data consists of 50,000 vectorized gray-level images of size 28 × 28 pixels (i.e., patterns over 784 visible units) that contain handwritten digits from zero to nine (MNIST dataset; LeCun et al., 1998). A robust model of this data would benefit from a hierarchical process that extracts increasingly more complex features (e.g., Gabor filters at the first level, edge detectors in the following layers, etc.). We used the DBN architecture proposed by Hinton and Salakhutdinov (2006) for this task, with three hidden layers of size 500, 500, and 2000 units, respectively. The data of the numerosity perception problem consists of 51,200 vectorized b/w images of size 30 × 30 pixels (i.e., patterns over 900 visible units) that contain up to 32 rectangular objects of variable size. We used the DBN architecture proposed by Stoianov and Zorzi (2012), consisting of two hidden layers of size 80 and 400 units, which was shown to extract abstract numerosity information.

#### *Learning parameters*

The DBN learning algorithm is governed by few meta parameters. First, the learning rate should be small, typically in the range 0.01–0.1. Second, the use of a momentum coefficient (i.e., a fraction of the previous weight update) is also critical to avoid local minima, and it is usually set to 0.5 at the beginning of training and then increased up to 0.9. Third, network weights should be regularized, that is kept relatively small, by applying a constant weight decrease in the form of a small weight-decay factor of about 0.0001. Finally, weights should be initialized with small random values drawn from a zero-mean Gaussian distribution with standard deviation of 0.01. The initial values of the bias can be set to zero. These and other issues related to training RBMs are discussed in a comprehensive practical guide by Hinton (2010a).

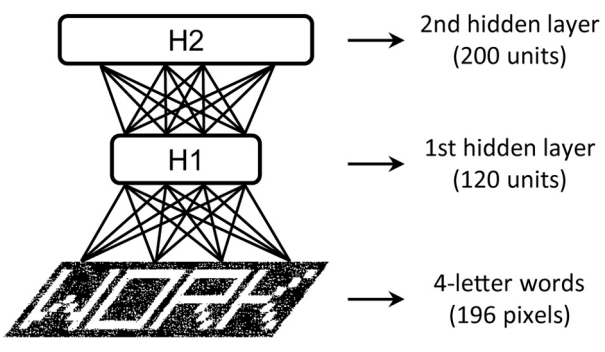



**FIGURE 4 | Architecture of the DBN with two hidden layers used in the written word perception problem.**

DBNs are trained with the CD learning algorithm, one RBM layer at a time, using as input either the sensory data (first RBM) or the activations of the previous hidden layer (deeper RBMs). This greedy, layer-wise learning procedure can be performed in a completely iterative way, by updating the network weights after each pattern (*on-line* learning). A complete sweep over all training patterns constitutes a learning epoch. In batch (*off-line*) learning, instead, weights updates are computed over the whole training set. A good compromise between these two approaches is to use a *mini-batch* learning scheme, in which the dataset is partitioned into small subsets (i.e., mini-batches) and the weights are updated with the average gradient computed on each subset (Neal and Hinton, 1998). This latter strategy is highly recommended, because it improves the quality of learning by avoiding local minima and it also allows to significantly speed-up the learning phase on multicore parallel implementations (see Testolin et al., 2013, for a mini-batch GPU implementation of deep networks). The mini-batch size should be set between 10 and few hundred patterns.

#### *Monitoring learning*

The learning progress can be monitored by analyzing the reconstruction error on the training patterns. The mean reconstruction error on the entire training set should fall rapidly at the beginning of learning and then gradually stabilize. However, this measure can be misleading because it is not the objective function optimized by the CD-$n$ algorithm, especially for large $n$ (Hinton, 2010a). A more precise measure of the performance of the network is to compare the free energy of the training data with that of a sample of held-out patterns (Hinton, 2010a). A final approach to monitor the quality of learning is to regularly perform an additional discriminative task over the learned internal representations, as we will discuss at length below.

#### *Sparsity constraints on internal representations*

An interesting variant of standard RBMs (and, consequently, DBNs) consists in forcing the network's internal representations to rely on a limited number of active hidden units. In this case the network develops sparse distributed representations, which have many useful properties and appear to be a coding strategy adopted by the brain (Olshausen and Field, 1996; see Olshausen and Field, 2004, for review). Forcing sparseness within a network's hidden layer can be interpreted in terms of inhibitory competition between units (O'Reilly, 2001). A sparse-coding version of the RBM encourages the development of more orthogonal features, which can allow a better pattern discriminability and a more intuitive interpretation of what each unit is representing. In RBMs, sparsity can be obtained by driving the probability $q$ of a unit to be active to a certain desired (low) probability $p$ (Lee et al., 2008;

Nair and Hinton, 2009). For logistic units, this can be practically implemented by first calculating the quantity *q-p*, which is then multiplied by a scaling factor and added to the biases (and, possibly, to each incoming weight) of the hidden units at every weight update. Depending on the number of hidden units, the desired sparsity level ($p$) can be set in the range of 0.01–0.1. Monitoring the distribution of the hidden units activity can be useful to verify that the desired sparsity level is obtained and that the scaling factor is correctly set so that the probability that a unit is active is close to $p$ while learning is not hindered (Hinton, 2010a).

#### TESTING A DBN: READ-OUT OF INTERNAL REPRESENTATIONS

When performing a discriminative task, one of the simplest methods is to exploit a linear classifier (e.g., Rosenblatt, 1958), to assign a certain class to each input pattern. The classifier makes a decision by using a linear combination of the input features and this represents its main limitation (Minsky and Papert, 1969). In the case of real sensory signals, this shortcoming is exacerbated by the fact that the feature vectors are high-dimensional and usually lie on highly curved and tangled manifolds (DiCarlo et al., 2012). However, deep belief networks perform a non-linear projection of the feature vector at each hidden layer, gradually building increasingly more complex and abstract representations of the data that eventually make explicit the latent causes of the sensory signal. This hierarchical organization suggests that a linear "read-out" of hidden unit representations should become increasingly more accurate as a function of layer depth. In this perspective, accuracy of linear read-out can be considered as a coarse measure of how well the relevant features are explicitly encoded at a given depth of the hierarchical generative model (see, e.g., Stoianov and Zorzi, 2012; Di Bono and Zorzi, under review). As noted above, linear read-out can also be used to monitor the quality of the representations developed by the deep network during unsupervised generative learning.

The linear read-out on internal representations can be easily implemented using another connectionist module, such as a linear network trained with the delta rule, thereby preserving the biological plausibility of the model. The linear network can also be considered as a response module that supports a particular behavioral task, so that its responses can be assessed against the human data (e.g., numerosity perception in Stoianov and Zorzi, 2012, or location-invariant visual word recognition in Di Bono and Zorzi, under review). For example, Stoianov and Zorzi applied this approach to simulate human behavior in a numerosity comparison task after training a DBN on thousands of images of sets of objects. The internal representations at the deepest layer provided the input to a linear network trained to decide whether the numerosity of the input image was larger or smaller than a reference number. Notably, the responses of this decision module were described by a psychometric function that was virtually identical to that of human adults, with the classic modulation by numerical ratio that is the signature of Weber's law for numbers.

From a practical point of view, delta rule learning can be conveniently replaced by an equivalent method that is computationally more efficient, which relies on the calculation of a pseudo-inverse matrix (Hertz et al., 1991). Formally, data patterns $P = \{P_1, P_2, \ldots, P_n\}$ can be associated with desired categories $L = \{L_1, L_2, \ldots, L_n\}$ by means of the following linear association:

$$L = WP$$

where $P$ and $L$ are matrices containing $n$ column vectors that correspondingly code patterns $P_i$ (sensory data or internal representations) and binary class labels $L_i$, and $W$ is the weight matrix of the linear classifier. If an exact solution to this linear system does not exist, a least-mean-square approximation can be found by computing the weight matrix as:

$$W = LP^{+}$$

where $P^{+}$ is the Moore-Penrose pseudo-inverse (Albert, 1972)[2]

As an example, we applied the read-out DBN testing method on the internal representations learned for the images of the four-letter words used in McClelland and Rumelhart (1981). We tested two different discriminative problems. The first required the identification of each of the four letters composing a word, using as label a binary vector with one-hot (i.e., localistic) coding of the target letter. The second problem consisted in the identification of the word itself, using as label a binary vector with one-hot coding of the target word. To investigate the quality of the features extracted by deep learning, we compared the classification accuracy on the representations learned at each of the levels of a two-layer DBN ($H1 = 120$ units, $H2 = 200$ units) with that of the representations learned by a single RBM with as many hidden units as the top layer of the DBN ($H = 200$ units). As a baseline, we also measured the classification accuracy obtained by trying to directly categorize the raw input vectors. Note that the read-out of the original data is trivial, due to lack of variability (and noise) in the coding of letters and words (i.e., there is a unique pattern for each letter and word). Indeed, the raw data vectors are linearly separable as shown by the perfect accuracy of the read-out. However, if the input patterns are degraded by adding a certain amount of noise, one should expect a progressive decrease of the classification accuracy when the input representation does not include high-level, invariant features. Indeed, **Figure 5** shows that when each word image was corrupted by randomly setting to zero a certain percentage of its pixels, read-out accuracy on the raw pixel data dropped even with a small amount of noise and it approached zero in the word recognition task. As expected, the DBN extracted robust internal representations that were less sensitive to noise. Indeed, both hidden layers supported good discrimination accuracy for letters, whereas only the deepest hidden layer adequately supported word discrimination. Notably, the shallow generative model (RBM) with as many hidden units as the top DBN layer did not unfold word-level information, thereby failing to support robust word recognition (especially for larger noise levels). These results are consistent with the seminal proposal of hierarchical feature processing to

[2]In some high-level programming languages, this operation is readily available. For example, in MATLAB/Octave we can use the *pinv()* function, $W = L^{*}pinv(P)$, or the left matrix divide (a.k.a. "backslash") operator, $W = (P'\backslash L')'$.

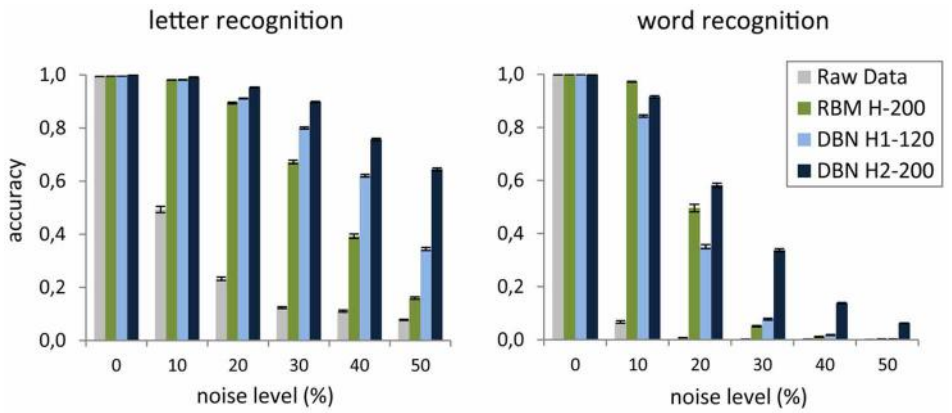



**FIGURE 5 | Mean accuracy of the linear classifier on the task of recognizing each letter of a word (left) and the whole word (right) as a function of noise level applied to the raw images.** Accuracy is averaged over 20 random noise injections and it is computed over the entire dataset of words. Error bars represent SEM. The results are shown for read-out from the two hidden layers of a deep network (DBN), a shallow network (RBM), and raw images.

yield abstract representations of written words (McClelland and Rumelhart, 1981).

## ANALYZING A DBN

### *Discovering learned representations*

In the previous section we illustrated how it is possible to assess the quality of the internal representations learned at each layer of the hierarchy of a deep belief network by performing a discriminative task. However, this information is tied to a given classification task and is therefore limited in scope. Moreover, the supervised classifier operates on the pattern of activity over an entire hidden layer, that is a distributed representation encoding a variety of micro-features (Hinton et al., 1986) representing task-independent statistical regularities of the data. A very simple but informative approach to investigate the role of a particular unit in the network consists of visualizing its connection weights using the original structure of the data (e.g., the 2D image in our visual perception examples). This is particularly intuitive for the first hidden layer, where the weight matrix defines how the visible units contribute to the activation of each hidden unit. We can therefore visualize the "receptive field" of each hidden unit by plotting the strength of its visible-to-hidden connections. The same principle can be applied to the deeper layers of the DBN, by combining their weight matrix with those of the lower layers. A straightforward way is to use a linear combination of the weight matrices, possibly imposing a threshold on the absolute values of the weights in order to select only strong connections. This allows to visualize the receptive field learned at a layer $k$ as a weighted linear combination of the receptive fields learned at level $k$-$1$ (Lee et al., 2008, 2009). The main drawbacks of this technique are that one has to manually choose threshold values and that non-linearities between layers are not considered, with the risk of losing relevant information. Nevertheless, this method can provide good visualization of the learned features even without imposing a threshold on the weights (see **Figure 6**).

Using the above method, we analyzed the receptive fields of the hidden units of DBNs trained on images of letters as well on the handwritten digits of the MNIST dataset. In the letter perception task, we found that most of the units of the first hidden layer were tuned to basic geometric features, whereas most of the units of the second hidden layer were tuned to a composition of these features (see examples in **Figure 6A**). The greater image resolution and variability of the handwritten digits pose a much more complex visual problem, which induced the emergence of a more structured hierarchy of features in the DBN. As shown in **Figure 6B**, the first hidden layer learned simple and localized visual features (mostly Gaussian and Gabor filters), resembling those found in the primary visual cortex. The second hidden layer combined these features into edges, lines, and strokes detectors. Finally, the third hidden layer extracted even more complex visual features that resemble parts of digits. Note that the finding of low-level visual features (basis functions) in the first hidden layer is common to many problems that involve a large variability in the training images (see, e.g., Lee et al., 2008; Stoianov and Zorzi, 2012).

Applying sparsity constraints on the internal representations further improves the quality of the emerging features. For example, a sparse DBN trained on patches of natural images developed complex receptive fields (e.g., T-junctions) in the second hidden layer that were very similar to those found in area V2 of the visual cortex (Lee et al., 2008). Our sparse DBN simulations also resulted in an increase of the complexity of the emergent features. For example, the letter perception network encoded more letter-like features in the second hidden layer (**Figure 6C**) and the handwritten digit perception network learned shape-specific detectors in the third hidden layer (**Figure 6D**).

A more sophisticated approach to investigate the features encoded by a hidden unit is to find its preferred input stimuli, as done by neurophysiologists in single-cell recording studies. The basic idea is to probe the network on a variety of input patterns, each time recording the neural response and then looking for possible regularities. This approach can be very effective if we have an idea about which type of patterns are more likely to elicit specific responses (for example, responses to bigrams after training on words; Di Bono and Zorzi, under review). However, if we cannot make assumptions about the nature of the preferred stimuli, this

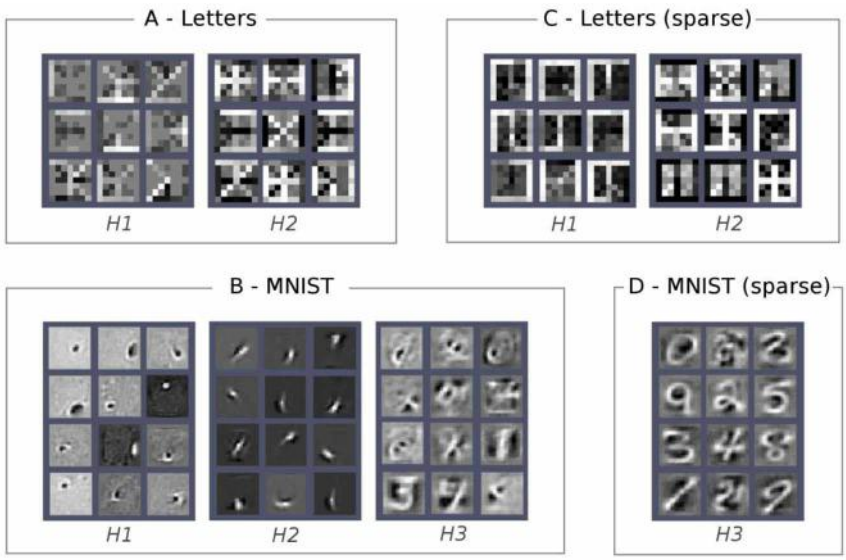



**FIGURE 6 | Visualization of features learned at different hidden layers ($H_i$).** Each square within a layer represents the receptive field of one hidden unit. Excitatory connections are shown in white, whereas inhibitory connections are in black. **(A)** H1 and H2 on single letters (pixelated "Siple font"). **(B)** H1, H2 and H3 on MNIST. **(C)** Sparse H1 and H2 on single letters. **(D)** Sparse H3 on MNIST. From left to right: H1 on single letters (pixelated "Siple font"); H2 on single letters; H1 on MNIST; H2 on MNIST; H3 on MNIST; sparse H1 on single letters; sparse H2 on single letters; sparse H3 on MNIST.

method becomes computationally intractable because it would require testing the network on an exponential number of possible input patterns. Nevertheless, this problem can be solved by formulating it as an *optimization problem*, where the goal is to find the input pattern that maximizes the activation of a certain hidden unit given the processing constraints imposed by the network (Erhan et al., 2009). Formally, if θ denotes the deep network parameters (weights and biases) and $h_{ij}(\theta, x)$ is the activation of a given unit $i$ from a given layer $j$ in the network, then $h_{ij}$ is a function of both θ and the input sample $x$. Assuming that the vector $x$ has a bounded norm and after learning the parameters are fixed, then the problem of maximizing the unit activation is:

$$x^* = \arg\max_x h_{ij}(\theta,\ x)$$

Although this is a non-convex optimization problem, it has been empirically shown that good local minima can be found (Erhan et al., 2009). This method has been recently used to investigate whether high-level, class-specific feature detectors can emerge in very-large-scale deep unsupervised learning (i.e., using millions of images for training; Le et al., 2012). The impressive result was that it is indeed possible to learn highly complex and abstract features at the deepest layers, such as prototypical faces (Le et al., 2012).

A different approach can be used if we expect monotonic response of some hidden units to a given property of the data. The individuation of these detectors is based on regressing the property of interest (or even multiple properties) onto the response of each hidden unit. A high absolute value of the normalized regression coefficient indicates sensitivity of the hidden unit to the property of interest; this might also indicate selectivity when combined with small (near-zero) regression coefficients for other properties. Using this method, Stoianov and Zorzi (2012) discovered detectors in the second hidden layer of their DBN tuned to visual numerosity but insensitive to other important visual properties like cumulative area. Di Bono and Zorzi (under review) also used this method to investigate word selectivity in their DBN model of visual word recognition. After finding the preferred word for a given hidden unit, its word selectivity was assessed by recording the response to all other training words and performing a regression analysis using the orthographic (i.e., Levenshtein) distance from the preferred word as predictor.

#### *Sampling from the generative model*

Up to this point, we only discussed methods that investigate the *bottom–up* processing of sensory data. However, a deep belief network is a *generative* model, and it can be very useful to assess the *top–down* generation of sensory data, as well as the mixing of bottom–up and top–down signals during inference in a noisy situation. In one scenario, we can provide to the model a noisy input pattern (e.g., randomly corrupted or partially occluded) and let the network find the most likely interpretation of the data under the generative model. This process requires the iteratively sampling of the states of the network until an equilibrium activation state is reached, which in DBNs can be efficiently done using block Gibbs sampling (see **Box 3**). As an example, in **Figure 7** we show the result of inference in the word perception DBN when four different noisy versions of the same image are given as input to the model. Note that the visible units settle onto an activation state corresponding to the correct word image.

We can also study the generative capability of a DBN when the visible units are not clamped to an initial state, and the network is therefore let free to autonomously produce a sensory pattern through a completely top–down process. This generative process can be constrained to produce "class prototypes" by

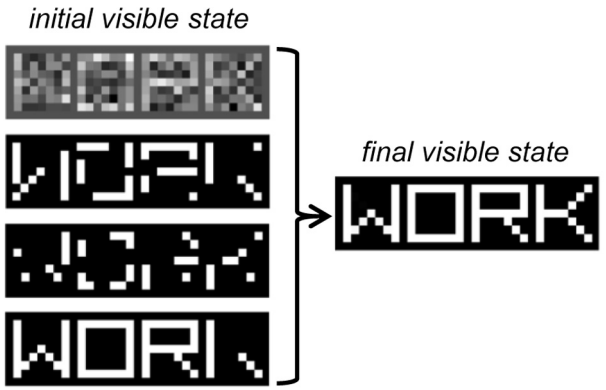



**FIGURE 7 | Inference in the word perception DBN when the word image "WORK" is presented as input under different types of noise.** From top to bottom: Gaussian noise, binary noise (30%), binary noise, (50%), occlusion noise. The final state of the visible units, identical across the four noise conditions, is shown on the right.

adding a multimodal RBM on the top of the network hierarchy (Hinton et al., 2006), which is jointly trained using two input sources, one containing the internal representation learned by the DBN and the other encoding the corresponding label. For example, in the handwritten digit recognition model, input to the multimodal RBM is provided by the second hidden layer (500 units) and by 10 units representing the image label (one unit for each possible digit class) (see **Figure 8A**). After learning, the label units can be clamped to a certain state (e.g., with only the unit corresponding to the class "7" active) and the top RBM settles to equilibrium, thereby recovering the internal representation of the given digit class. The generative connections of the DBN can then be used to obtain an image on the visible layer in a single top–down pass. The image generated can be thought of as the model's prototype for the corresponding abstract representation.

Here we propose an interesting, more simple variant of the top–down generation of the learned prototypes. Instead of jointly training the top-level RBM using the internal representation of images and the corresponding class label, and then performing Gibbs sampling until equilibrium with the label units clamped to a certain class, we can try to directly map the class label and the internal representation through a linear projection (see **Figure 8B**). This mapping is analogous to the read-out module previously discussed but it works in the opposite direction. Prototype generation can thus be performed by associating the class vectors $L$ with the internal representations $P$ learned by the DBN through a weight matrix $W_2$:

$$P = W_2 L$$

$$W_2 = PL^+$$

As in Hinton et al. (2006), after computing the internal state $P$ at the deepest layer, a single top–down pass through the generative connections of the DBN produces the prototype for the specific class. **Figure 8C** shows the prototypes generated for each digit class of the MNIST dataset using this linear projection method. Note that this method can be readily extended to more complex scenarios that involve a mapping between internal representations learned by different networks (which may reflect knowledge about different domains or sensory modalities).

Finally, it is worth noting that the quality of inference when sampling from the generative model can be improved if the single top–down pass is replaced by an interactive process, as proposed in a recent variant of the DBN known as Deep Boltzmann Machine (Salakhutdinov and Hinton, 2009).

## DISCUSSION

Understanding how cognition and language might emerge from neural computation is certainly one of the most exciting frontiers in cognitive neuroscience. In this tutorial overview we discussed a recent step forward in connectionist modeling, which allows the emergence of hierarchical representations in a deep neural network learning a generative model of the sensory data. We started by reviewing the theoretical foundations of deep learning, which rely on the framework of probabilistic graphical models to derive efficient inference and learning algorithms over hierarchically organized energy-based models. We then provided a step-by-step tutorial on how to practically perform a complete deep learning simulation, covering the main aspects related to the training, testing and analysis of deep belief networks. In our presentation we focused on examples that require the progressive extraction of abstract representations from sensory data and that are therefore representative of a wide range of cognitive processes. In particular, we showed how deep learning can be applied to the classic letter and word perception problem of McClelland and Rumelhart (1981). In addition to providing a useful toy example of modeling based on deep learning, the emergent properties of the model revisit key aspects of the seminal IAM and suggest a very promising research direction for developing a full-blown deep learning model of visual word recognition. Indeed, up-scaling the present toy model is likely to be successful because deep learning is particularly suited to capture features hierarchies over large training datasets with great pattern variability. This aspect was present in two additional problems that complemented our tutorial with more realistic simulations, that is, handwritten digit recognition (LeCun et al., 1998) and visual numerosity perception (Stoianov and Zorzi, 2012). Together, the various simulations illustrate the strength of the deep learning approach to cognitive modeling.

Deep unsupervised learning extracts increasingly more abstract representations of the world, with the important consequence that explanatory factors behind the sensory data can be shared across tasks. The hierarchical architecture captures higher order structure of input data that might be invisible at the lower levels and it efficiently exploits features re-use. The idea that learned internal representations at the deepest layers can be easily "read-out" is consistent with the notion of "explicitness of information" articulated by Kirsh (1990), who argued that explicitness is tightly related to the processing system which uses it. Within this perspective, the degree of explicitness is better linked to the

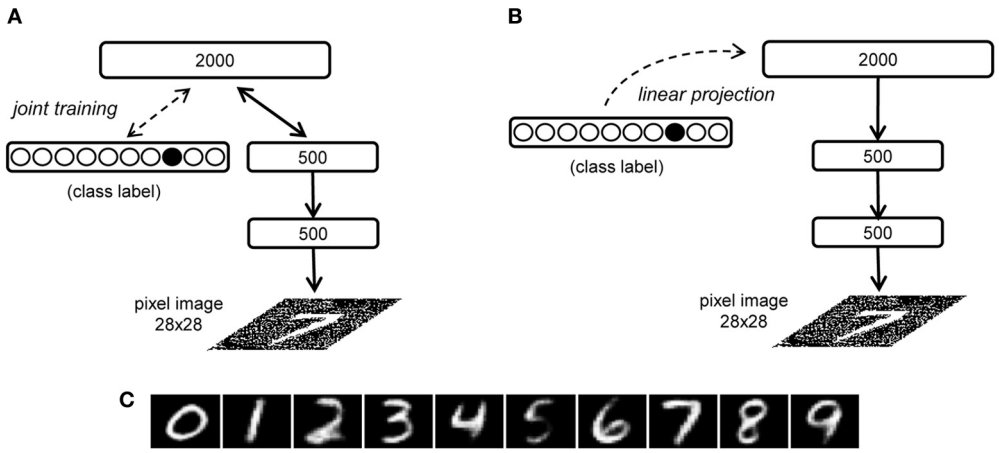



**FIGURE 8 | Illustration of the prototype generation methods in the handwritten digit recognition model. (A)** The RBM involving the third hidden layer is jointly trained on the internal representation of the second hidden layer and an additional set of units representing the digit classes (Hinton et al., 2006). **(B)** Our linear projection method: class label units are only added after the complete DBN training and are associated to the third hidden layer representations by means of a linear mapping. **(C)** Digit prototypes generated using the linear projection method.

usability of information rather than to its form (i.e., how quickly it can be accessed, retrieved or in some other manner put to use). This idea has been further extended by Clark (1992), who proposed to take into account also the multi-track usability of stored information: "Truly explicit items of information should be usable in a wide variety of ways, that is, not restricted to use in a single task" (p. 198). Note that this conception of abstract representations that can be shared across tasks or even across domains is particularly useful in the context of modeling language processing.

Efficient generative learning in neural networks is a recent breakthrough in machine learning and its potential has yet to be fully unfolded. In particular, the extension of RBMs to the temporal domain (Sutskever et al., 2008; Taylor and Hinton, 2007) is a very promising avenue for research. Indeed, generative networks that learn the temporal dynamics of the data could anticipate relevant events in the environment, using the history of the system as context to make accurate predictions about the incoming information, as proposed by the predictive coding framework (Huang and Rao, 2011; Clark, 2013). Learning and processing of sequential information is also a key aspect of cognition and it is particularly ubiquitous in language processing (Elman, 1990). An initial exploration of this direction is the use of the Recurrent Temporal RBM (Sutskever et al., 2008) for learning orthographic structure from letter sequences (Testolin et al., 2012, submitted).

It is worth noting that deep generative network models of cognition can offer a unified theoretical framework that encompasses classic connectionism and the structured Bayesian approach to cognition. Structured Bayesian models of cognition (for reviews see Chater et al., 2006; Griffiths et al., 2010) assume that human learning and inference approximately follow the principles of Bayesian probabilistic inference and they have been used in the last few years to address a number of issues in cognitive science, including language processing (Chater and Manning, 2006, for review). However, Bayesian models are typically formulated at the level of "computational theory" (Marr, 1982) rather than at the process level that characterizes other cognitive modeling paradigms like connectionism (for further discussion see McClelland et al., 2010; Jones and Love, 2011). This implies limits on the phenomena that can be studied with the Bayesian approach, because only problems of inductive inference or that contain an inductive component are naturally expressed in Bayesian terms (Griffiths et al., 2008). In contrast, computational models of cognition based on deep neural networks and generative learning implement the probabilistic approach in a neural-like architecture and can provide an emergentist explanation of structured representations that is in line with the connectionist tradition (McClelland et al., 2010). Their probabilistic formulation not only allows to deal with ambiguity of sensory input and with the intrinsic uncertainty of environmental dynamics, but it also provides a coherent theory about how learning can integrate new evidence to refine beliefs of the model. Importantly, there is no need to have an external signal that guides learning, because the aim is to reproduce incoming information as accurately as possible by discovering its hidden causes (that is, learning can be seen as a stochastic inference problem).

In conclusion, we believe that the focus on deep architectures and generative learning represents a crucial step forward for the connectionist modeling enterprise, because it offers a more plausible model of cortical learning as well as way to bridge the gap between emergentist connectionist models and structured Bayesian models of cognition.

## ACKNOWLEDGMENTS

This study was supported by the European Research Council (grant no. 210922 to Marco Zorzi).

## REFERENCES


Ackley, D., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cogn. Sci.* 9, 147–169. doi: 10.1207/s15516709cog0901_7

Albert, A. (1972). *Regression and the Moore-Penrose pseudoinverse*. New York, NY: Academic Press.

Andrieu, C., De Freitas, N., Doucet, A., Jordan, M. I., and Freitas, N. De, (2003). An introduction to MCMC for machine learning. *Mach. Learn.* 50, 5–43. doi: 10.1023/A:1020281327116

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. *J. Mach. Learn. Res.* 27, 37–50.

Bengio, Y. (2009). Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2, 1–127. doi: 10.1561/2200000006

Bengio, Y., Courville, A., and Vincent, P. (2012). Representation learning: a review and new perspectives. *arXiv* 1206.5538, 1–34.

Bengio, Y., and Lamblin, P. (2007). Greedy layer-wise training of deep networks. *Adv. Neural Inform. Process. Syst.* 19, 153–170.

Chater, N., and Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends Cogn. Sci.* 10, 335–344. doi: 10.1016/j.tics.2006.05.006

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: conceptual foundations. *Trends Cogn. Sci.* 10, 287–291. doi: 10.1016/j.tics.2006.05.007

Clark, A. (1992). The presence of a symbol. *Connect. Sci.* 4, 193–205.

Clark, A. (2013). Whatever next. Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz machine. *Neural Comput.* 7, 889–904. doi: 10.1162/neco.1995.7.5.889

Dean, J., Corrado, G. S., Monga, R., Chen, K., Devin, M., Le, Q. V., et al. (2012). Large scale distributed deep networks. *Adv. Neural Inform. Process. Syst.* 24, 1–9.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition. *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

De Filippo De Grazia, M., Cutini, S., Lisi, M., and Zorzi, M. (2012). Space coding for sensorimotor transformations can emerge through unsupervised learning. *Cogn. Process.* 13(Suppl. 1), 141–146. doi: 10.1007/s10339-012-0478-4

Dehaene, S., Cohen, L., Sigman, M., and Vinckier, F. (2005). The neural code for written words: a proposal. *Trends Cogn. Sci.* 9, 335–341. doi: 10.1016/j.tics.2005.05.004

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211.

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). "Visualizing higher-layer features of a deep network," in *Technical Report, Univeristé de Montréal*, 1–13.

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 721–741. doi: 10.1109/TPAMI.1984.4767596

Ghahramani, Z., Korenberg, A., and Hinton, G. E. (1999). "Scaling in a hierarchical unsupervised network," in *International Conference on Artificial Neural Networks*, (Edinburgh), 13–18.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci.* 14, 357–364. doi: 10.1016/j.tics.2010.05.004

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). "Bayesian models of cognition," in *Cambridge Handbook of Computational Cognitive Modeling*, ed R. Sun (Cambridge, MA: Cambridge University Press), 59–100.

Harm, M. W., and Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychol. Rev.* 106, 491–528. doi: 10.1037/0033-295X.106.3.491

Harm, M. W., and Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychol. Rev.* 111, 662–720. doi: 10.1037/0033-295X.111.3.662

Hertz, J. A., Krogh, A. S., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Weasley.

Hinton, G. (2013). Where do features come from? *Cogn. Sci.* doi: 10.1111/cogs.12049. [Epub ahead of print].

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* 14, 1771–1800. doi: 10.1162/089976602760128018

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends Cogn. Sci.* 11, 428–434. doi: 10.1016/j.tics.2007.09.004

Hinton, G. E. (2010a). "A practical guide to training restricted boltzmann machines," in *Technical Report UTML TR 2010-003, University of Toronto*. 9, 1.

Hinton, G. E. (2010b). Learning to represent visual input. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 177–184. doi: 10.1098/rstb.2009.0200

Hinton, G. E., and Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 352, 1177–1190. doi: 10.1098/rstb.1997.0101

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. (1986). "Distributed representations," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Cambridge, MA: MIT Press), 77–109.

Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi: 10.1162/neco.2006.18.7.1527

Hinton, G. E., and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hinton, G. E., and Sejnowski, T. J. (1999). *Unsupervised Learning: Foundations of Neural Computation*. Cambridge, MA: MIT Press.

Huang, Y., and Rao, R. P. N. (2011). Predictive coding. Wiley interdisciplinary reviews. *Cogn. Sci.* 2, 580–593. doi: 10.1002/wcs.142

Jones, M., and Love, B. C. (2011). Bayesian fundamentalism or enlightenment. on the explanatory status and theoretical contributions of bayesian models of cognition. *Behav. Brain Sci.* 34, 169–88; disuccsion 188–231.

Jordan, M. I., and Sejnowski, T. J. (2001). *Graphical Models: Foundations of Neural Computation*. Cambridge, MA: MIT Press.

Khaitan, P., and McClelland, J. L. (2010). "Matching exact posterior probabilities in the multinomial interactive activation model," in *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, eds S. Ohlsson and R. Catrambone (Austin, TX: Cognitive Science Society), 623.

Kirsh, D. (1990). "When is information explicitly represented?" in *The Vancouver Studies in Cognitive Science,* ed P. Hanson (Vancouver, BC: UBC Press), 340–365.

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE* 78, 1464–1480. doi: 10.1109/5.58325

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: The MIT Press.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 24, 1–9.

Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., et al. (2012). "Building high-level features using large scale unsupervised learning," in *International Conference on Machine Learning*, (Edinburgh).

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Lee, H., Ekanadham, C., and Ng, A. Y. (2008). Sparse deep belief net models for visual area V2. *Adv. Neural Inform. Process. Syst.* 20, 873–880.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *International Conference on Machine Learning* (New York, NY: ACM Press), 609–616.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychon. Bull. Rev.* 9, 829–835. doi: 10.3758/BF03196342

Love, B. C., Medin, D. L., and Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychol. Rev.* 111, 309–332. doi: 10.1037/0033-295X.111.2.309

Marr, D. (1982). *Vision: a Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W. H. Freeman and Company.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., et al. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci.* 14, 348–356. doi: 10.1016/j.tics.2010.06.002

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychol. Rev.* 88, 375. doi: 10.1037/0033-295X.88.5.375

Minsky, M., and Papert, S. (1969). *Perceptrons: an Introduction to Computational Geometry.* Cambridge, MA: MIT Press.

Mirman, D., Bolger, D. J., Khaitan, P., and McClelland, J. L. (in press). Interactive activation and mutual constraint satisfaction. *Cogn. Sci.*

Nair, V., and Hinton, G. E. (2009). 3D Object Recognition with Deep Belief Nets. *Adv. Neural Inf. Process. Syst.* 21, 1339–1347.

Neal, R. M., and Hinton, G. E. (1998). "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*, ed M. I. Jordan (Dordecht, The Netherlands: Kluwer Academic Publishers), 355–368.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *J. Math. Biol.* 1, 267–273. doi: 10.1007/BF00275687

Olshausen, B. A, and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Olshausen, B. A, and Field, D. J. (2004). Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487. doi: 10.1016/j.conb.2004.07.007

O'Reilly, R. (1998). Six principles for biologically based computational models of cortical cognition. *Trends Cogn. Sci.* 2, 455–462.

O'Reilly, R. C. (2001). Generalization in interactive networks: the benefits of inhibitory competition and Hebbian learning. *Neural Comput.* 13, 1199–1241.

O'Reilly, R., and Munakata, Y. (2000). *Computational Exploration in Cognitive Neuroscience*. Cambridge, MA: MIT Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann.

Perry, C., Ziegler, J. C., and Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychol. Rev.* 114, 273–315. doi: 10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J. C., and Zorzi, M. (2010). Beyond single syllables: large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cogn. Psychol.* 61, 106–151. doi: 10.1016/j.cogpsych.2010.04.001

Perry, C., Ziegler, J. C., and Zorzi, M. (2013). A computational and empirical investigation of graphemes in reading. *Cogn. Sci.* 37, 800–828. doi: 10.1111/cogs.12030

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115. doi: 10.1037/0033-295X.103.1.56

Plaut, D., and Shallice, T. (1993). Deep dyslexia: a case study of connectionist neuropsychology. *Cogn. Neuropsychol.* 10, 377–500. doi: 10.1080/02643299308253469

Raina, R., Madhavan, A., and Ng, A. Y. (2009). "Large-scale deep unsupervised learning using graphics processors," in *International Conference on Machine Learning* (New York, NY: ACM Press), 1–8.

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386. doi: 10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., and Williams, R. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Rumelhart, D., and McClelland, J. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1: Foundations. Cambridge, MA: MIT Press.

Rumelhart, D. E., and Siple, P. (1974). Process of recognizing tachistoscopically presented words. *Psychol. Rev.* 81, 99–118. doi: 10.1037/h0036117

Rumelhart, D. E., and Todd, P. M. (1993). "Learning and connectionist representations," in *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, eds D. Meyer and S. Kornblum (Cambridge, MA: MIT Press), 3–30.

Rumelhart, D. E., and Zipser, D. (1985). Feature discovery by competitive learning. *Cogn. Sci.* 9, 75–112. doi: 10.1207/s15516709cog0901_5

Salakhutdinov, R., and Hinton, G. E. (2009). "Deep boltzmann machines," in *International Conference on Artificial Intelligence and Statistics*, (Clearwater Beach, FL), 448–455.

Schyns, P., Goldstone, R., and Thibaut, J. (1998). The development of features in object concepts. *Behav. Brain Sci.* 21, 1–17. doi: 10.1017/S0140525X98000107

Seidenberg, M. S., and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychol. Rev.* 96, 523–568. doi: 10.1037/0033-295X.96.4.523

Stoianov, I., and Zorzi, M. (2012). Emergence of a "visual number sense" in hierarchical generative models. *Nat. Neurosci.* 15, 194–196. doi: 10.1038/nn.2996

Stoianov, I., Zorzi, M., Becker, S., and Umiltá, C. (2002). "Associative arithmetic with Boltzmann Machines: The role of number representations," in *Lecture Notes in Computer Science: ICANN 2002*, ed J. Dorronsoro (Berlin: Springer), 277–283.

Stoianov, I., Zorzi, M., and Umiltà, C. (2004). The role of semantic and symbolic representations in arithmetic processing: insights from simulated dyscalculia in a connectionist model. *Cortex* 40, 194–196. doi: 10.1016/S0010-9452(08)70948-1

Sutskever, I., Hinton, G. E., and Taylor, G. (2008). The recurrent temporal restricted Boltzmann machine. *Adv. Neural Inform. Process. Syst.* 20, 1601–1608.

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning*. Cambridge, MA: MIT Press.

Taylor, G., and Hinton, G. E. (2007). Modeling human motion using binary latent variables. *Adv. Neural Inform. Process. Syst.* 19, 1345–1352.

Testolin, A., Sperduti, A., Stoianov, I., and Zorzi, M. (2012). "Assessment of sequential boltzmann machines on a lexical processing task," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, (Bruges), 275–280.

Testolin, A., Stoianov, I., De Filippo De Grazia, M., and Zorzi, M. (2013). Deep unsupervised learning on a desktop PC: a primer for cognitive scientists. *Front. Psychol.* 4:251. doi: 10.3389/fpsyg.2013.00251

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., and Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: dissecting the inner organization of the visual word-form system. *Neuron* 55, 143–156. doi: 10.1016/j.neuron.2007.05.031

Welling, M., and Hinton, G. E. (2002). "A new learning algorithm for mean field boltzmann machines," in *Lecture Notes in Computer Science: ICANN 2002*, ed J. Dorronsoro (Berlin: Springer), 351–357.

Zorzi, M., Houghton, G., and Butterworth, B. (1998). Two routes or one in reading aloud. a connectionist dual-process model. *J. Exp. Psychol. Hum. Percept. Perfor* 24, 1131–1161. doi: 10.1037/0096-1523.24.4.1131

Zorzi, M., Stoianov, I., and Umiltà, C. (2005). "Computational modeling of numerical cognition," in *Handbook of Mathematical Cognition*, ed J. Campbell (New York, NY: Psychology Press), 67–84.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 31 May 2013; paper pending published: 08 July 2013; accepted: 20 July 2013; published online: 20 August 2013.*
*Citation: Zorzi M, Testolin A and Stoianov IP (2013) Modeling language and cognition with deep unsupervised learning: a tutorial overview. Front. Psychol.* ***4****:515. doi: 10.3389/fpsyg.2013.00515*
*This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 02 September 2013
doi: 10.3389/fpsyg.2013.00563

# Spoken word recognition without a TRACE

***Thomas Hannagan[1]*, James S. Magnuson[2,3] and Jonathan Grainger[1]***

[1] *Laboratoire de Psychologie Cognitive, CNRS/Aix-Marseille University, Marseille, France*
[2] *Department of Psychology, University of Connecticut, Storrs, CT, USA*
[3] *Haskins Laboratories, New Haven, CT, USA*

***Edited by:***
*Julien Mayor, University of Geneva, Switzerland*

***Reviewed by:***
*Matthew H. Davis, MRC Cognition and Brain Sciences Unit, UK*
*Ulrich H. Frauenfelder, University of Geneva, Switzerland*

***Correspondence:**
*Thomas Hannagan, Laboratoire de Psychologie Cognitive, CNRS/Université Aix-Marseille, 5 place Victor Hugo, 13331 Marseille, France*
*e-mail: thom.hannagan@gmail.com*

How do we map the rapid input of spoken language onto phonological and lexical representations over time? Attempts at psychologically-tractable computational models of spoken word recognition tend either to ignore time or to transform the temporal input into a spatial representation. TRACE, a connectionist model with broad and deep coverage of speech perception and spoken word recognition phenomena, takes the latter approach, using exclusively time-specific units at every level of representation. TRACE reduplicates featural, phonemic, and lexical inputs at every time step in a large memory trace, with rich interconnections (excitatory forward and backward connections between levels and inhibitory links within levels). As the length of the memory trace is increased, or as the phoneme and lexical inventory of the model is increased to a realistic size, this reduplication of time- (temporal position) specific units leads to a dramatic proliferation of units and connections, begging the question of whether a more efficient approach is possible. Our starting point is the observation that models of visual object recognition—including visual word recognition—have grappled with the problem of spatial invariance, and arrived at solutions other than a fully-reduplicative strategy like that of TRACE. This inspires a new model of spoken word recognition that combines time-specific phoneme representations similar to those in TRACE with higher-level representations based on string kernels: temporally independent (time invariant) diphone and lexical units. This reduces the number of necessary units and connections by several orders of magnitude relative to TRACE. Critically, we compare the new model to TRACE on a set of key phenomena, demonstrating that the new model inherits much of the behavior of TRACE and that the drastic computational savings do not come at the cost of explanatory power.

**Keywords: spoken word recognition, time-invariance, TRACE model, symmetry networks, string kernels**

## 1. INTRODUCTION

There is a computational model of spoken word recognition whose explanatory power goes far beyond that of all known alternatives, accounting for a wide variety of data from long-used button-press tasks like lexical decision (McClelland and Elman, 1986) as well as fine-grained timecourse data from the visual world paradigm (Allopenna et al., 1998; Dahan et al., 2001a,b; see Strauss et al., 2007, for a review). This is particularly surprising given that we are not talking about a recent model. Indeed, the model we are talking about—the TRACE model (McClelland and Elman, 1986)—was developed nearly 30 years ago, but successfully simulates a broad range of fine-grained phenomena observed using experimental techniques that only began to be used to study spoken word recognition more than a decade after the model was introduced.

TRACE is an interactive activation (IA) connectionist model. The essence of IA is to construe word recognition as a hierarchical competition process taking place over time, where excitatory connections between levels and inhibitory connections within levels result in a self-organizing resonance process where the system fluxes between dominance by one unit or another (as a function of bottom–up and top–down support) over time at each level. The levels in TRACE begin with a pseudo-spectral representation of acoustic-phonetic features. These feed forward to a phoneme level, which in turn feeds forward to a word level. The model is interactive in that higher levels send feedback to lower levels (though in standard parameter settings, only feedback from words to phonemes is non-zero). **Figure 1** provides a conceptual schematic of these basic layers and connectivities, although the implementational details are much more complex.

The details are more complex because of the way the model tackles the extremely difficult problem of recognizing series of phonemes or words that unfold over time, at a sub-phonemic grain. The solution implemented in TRACE is to take the conceptual network of **Figure 1** and reduplicate every feature, phoneme, and word at successive timesteps. Time steps are meant to approximate 10 ms, and feature units are duplicated at every slice, while phonemes and words are duplicated every third slice. Thus, the phoneme layer can be visualized as a matrix with one row per phoneme and one column per time slice (i.e., a phonemes × slices matrix). However, units also have temporal extent—features for a given phoneme input extend over 11 time slices, ramping on and off in intensity. The same scheme is used at the lexical level, which can be visualized as a words × time slices matrix. Word

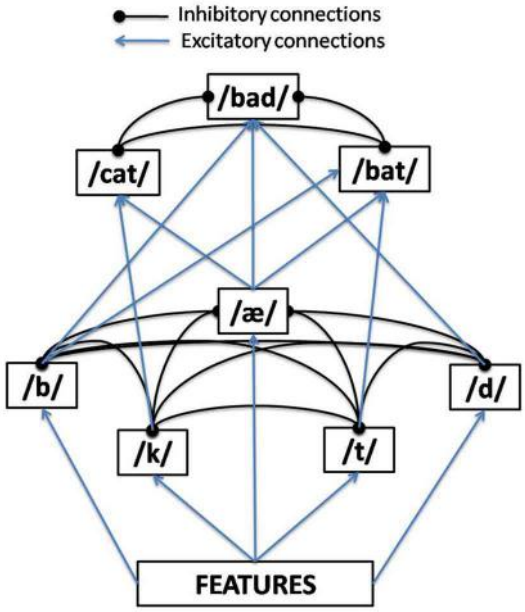



**FIGURE 1 | One time-slice of the TRACE model of spoken word recognition.**

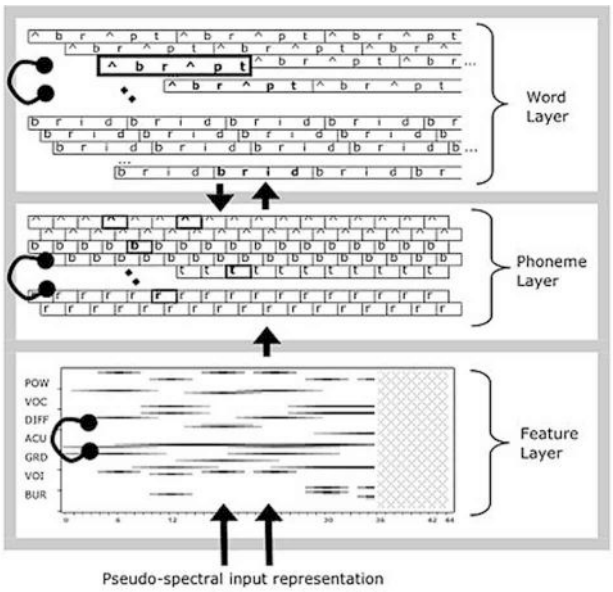



**FIGURE 2 | The detailed structure of the TRACE model of spoken word recognition (adapted from McClelland and Elman, 1986).**

lengths are not the simple product of constituent phoneme durations because phoneme centers are spaced six slices apart. This also gives TRACE a coarse analog to coarticulation; the features for successive phonemes overlap in time (but this is a weak analog, since feature patterns simply overlap and sometimes sum; but real coarticulation actually changes the realization of nearby and sometimes distant articulatory gestures). Each feature unit has forward connections to all phoneme units containing that feature that are aligned with it in time. Each phoneme unit has a forward connection to and a feedback connection from each word unit that "expects" that phoneme at that temporal location (so a /d/ unit at slice *s* has connections to /d/-initial words aligned near [at or just before or after] slice *s*, /d/-final words whose offsets are aligned at or adjacent to *s*, etc.). This more complex structure is shown in **Figure 2**.

The input to the model is transient; activation is applied to feature units "left-to-right" in time, as an analog of real speech input. Features that are activated then send activation forward. In IA networks, activation persists even after the removal of bottom–up input, as activation decays gradually rather than instantaneously. So as time progresses beyond the moment aligned with slice *s*, units aligned at slice *s* can continue to be active. A unit's activation at a time step, *t*, is a weighted sum of its bottom–up input, its top–down input, and its own activation at time *t-1*, minus a decay constant. The crucial point in understanding TRACE is that time is represented in two different ways. First, stimulus time unfolds step-by-step, with bottom–up inputs for that step applied only in that step. Second, time-specific units at each level are aligned with a specific time step, *t*, but their activation can continue to wax and wane after the bottom–up stimulus has been applied at time *t*. This is because the model will only receive external input at time *t*, but activation will continue to flow among units aligned with time *t* as a function of bottom–up, top–down, and lateral connections within the model. This is what inspires the name "TRACE": activation of a unit at time *t* is a constantly updating memory of what happened at time *t* modulated by lateral and top–down input.

In the original TRACE paper, McClelland and Elman presented results demonstrating how TRACE accounts for about 15 (depending on how one counts) crucial phenomena in human speech perception and spoken word recognition (see also Strauss et al., 2007 for a review). McClelland (1991) demonstrated how the addition of stochastic noise allowed TRACE to account properly for joint effects of context and stimulus (in response to a critique by Massaro, 1989). More recently, TRACE has been successfully applied to the fine-grained time-course of effects of phonological competition (Allopenna et al., 1998), word frequency (Dahan et al., 2001a), and subcateogorical (subphonemic) mismatches (Dahan et al., 2001b), using the visual world paradigm (Tanenhaus et al., 1995). In this paradigm, eye movements are tracked as participants follow spoken instructions to interact with real or computer-displayed arrays of objects (see Cooper, 1974, for an earlier, passive-task variant of the paradigm, the potential of which was not recognized at the time). While participants make only a few saccades per trial, by averaging over many trials, one can estimate the fine-grained time course of lexical activation and competition over time.

While some models have simulated aspects of visual world results (e.g., ShortlistB, Norris and McQueen, 2008), none has simulated the full set TRACE simulates, nor with comparable precision (although this assertion is based largely on absence of evidence—most models have not been applied to the full range of phenomena TRACE has; see Magnuson et al., 2012, for a review).

While TRACE is not a learning model, its ability to account for such a variety of findings in a framework that allows one to test highly specific hypotheses about the general organization of spoken word recognition (for instance TRACE's assumption of localist and separated levels of representations makes it easier to consider the impact of perturbing specific levels of organization, i.e., sublexical or lexical). However, while TRACE does an excellent job at fitting many phenomena, its translation of time to space via its time-specific reduplications of featural, phonemic and lexical units is notably inefficient (indeed, McClelland and Elman, 1986 noted it themselves; p. 77). In fact, as we shall describe in detail below, extending TRACE to a realistic phoneme inventory (40 instead of 14) and a realistic lexicon size (20,000 instead of 212 words) would require approximately 4 million units and 80 billion connections. To us, this begs a simple question: is it possible to create a model that preserves the many useful aspects of TRACE's behavior and simplicity while avoiding the apparent inefficiency of reduplication of time-specific units at every level of the model? As we explain next, we take our inspiration from solutions proposed for achieving spatial invariance in visual word recognition in order to tackle the problem of temporal invariance in spoken word recognition.

### 1.1. TIME AND TRACE: MAN BITES GOD

Visual words have several advantages over spoken words as objects of perception. All their elements appear simultaneously, and they (normally) persist in time, allowing the perceiver to take as much time as she needs, even reinspecting a word when needed. In a series of words, spaces indicate word boundaries, making the idea of one-at-a-time word processing (rather than letter-by-letter sequential processing) possible. In speech, the components of words cannot occur simultaneously (with the exception of single-vowel words like "a"). Instead, the phonological forms of words must be recovered from the acoustic outcomes of a series of rapidly performed and overlapping (coarticulated) gymnastic feats of vocal articulators. A spoken word's parts are transient, and cannot be reinspected except if they are held in quickly decaying echoic memory. In a series of words, articulation and the signal are continuous; there are no robust cues to word boundaries, meaning the perceiver must somehow simultaneously segment and recognize spoken words on the fly. Any processing model of spoken word recognition will need some way to code the temporal order of phonemes and words in the speech stream. There are four fundamental problems the model will have to grapple with.

First, there is the "temporal order problem," which we might call the "*dog or god*" problem. If, for example, a model simply sent activation to word representations whenever any of their constituent phonemes were encountered without any concern for order, the sequences /dag/, /gad/, /agd/ (etc.) would equally and simultaneously activate representations of both *dog* and *god*. TRACE solves this by having temporal order built into lexical level units: a unit for *dog* is a template detector for the ordered pattern /d/-/a/-/g/, whereas a *god* unit is a template detector for /g/-/a/-/d/.

Second, there is the "multi-token independence problem," or what we might call the "*do/dude*" or "*dog eats dog*" problem: the need to encode multiple instances of the same phoneme (as in words like *dude*, *dad*, *bib*, *gig*, *dread*, or *Mississippi*) or word (as in *dog eats dog*). That is, a model must be able to treat the two instances of /d/ in *dude* and the two instances of *dog* in *dog eats dog* as independent events. For example, if we tried having a simple model with just one unit representing /d/, the second /d/ in *dude* would just give us more evidence for /d/ (that is, more evidence for *do*), not evidence of a new event. The same would be true for *dog eats dog*; a single *dog* unit would just get more activated by the second instance without some way of treating the two tokens as independent events. TRACE achieves multi-token independence by brute force: it has literally independent detectors aligned at different time slices. If the first /d/ is centered at slice 6, the /a/ (both /a/ and /ae/ are represented by /a/ in TRACE) will be centered at slice 12 and the final /d/ will be centered at slice 18. The two /d/ events will activate completely different /d/ phoneme units. Thus, TRACE achieves multi-token independence (the ability to "recognize" two temporally distant tokens of the same type as independent) by having time-specific detectors.

Third is the "*man bites dog*" problem, which is the temporal order problem extended to multi-word sequences. The model must have some way to code the ordering of words; knowing that the words *dog*, *man*, and *bites* have occurred is insufficient; the model must be able to tell *man bites dog* from *dog bites man*. Putting these first three problems together, we might call them the "*man bites god*" problem—without order, lexical ambiguities will lead to later phrasal ambiguities. TRACE's reduplicated units allow it to handle all three.

Finally, there is the "segmentation problem." Even if we ignore the primary segmentation problem in real speech (the fact that phonemes overlap due to coarticulation) and make the common simplifying assumption that the input to spoken word recognition is a series of already-recognized phonemes, we need a way to segment words. It may seem that this problem should be logically prior to the "*man bites dog*" problem, but many theories and models of spoken word recognition propose that segmentation emerges from mechanisms that map phonemes to words. For example, in the Cohort model (Marslen-Wilson and Tyler, 1980), speech input in the form of phoneme sequences is mapped onto lexical representations (ordered phonological forms) phoneme-by-phoneme. When a sequence cannot continue to be mapped onto a single word, a word boundary is postulated (e.g., given *the dog*, a boundary would be postulated at /d/ because it could not be appended to the previous sequence and still form a word). TRACE was inspired largely by the Cohort model, but rather than explicitly seeking and representing word boundaries, segmentation is emergent: lateral inhibition among temporally-overlapping word units forces the model to settle on a series of transient, temporary "winners"—word units that dominate at different time slices in the "trace."

Solving several problems at once is compelling, but the computational cost is high. Specifically, because TRACE relies on reduplication at every time slice of features, phonemes, and words, the number of units in the model will grow linearly as a function of the number of time slices, features, phonemes, and words. But because units in TRACE have inhibitory links to all overlapping units at the same level, the number of connections grows quadratically as units at any level increase. Scaling up the

14 phonemes in the original TRACE model to the approximately 40 phonemes in the English inventory would not in itself lead to an explosive increase in units or connections (see Appendix A). Moving from the original TRACE lexicon of just 212 words to a realistically-sized lexicon of 20,000 words, however, would. In fact, the original TRACE model, with 14 phonemes and 212 words would require 15,000 units and 45 million connections. Increasing the phoneme inventory would change the number of units to approximately 17,000 and the number of connections to 45.4 million. Increasing the lexicon to 20,000 words would result in 1.3 million units and 400 billion connections. How might we construct a more efficient model?

### 1.2. VISUAL AND SPOKEN WORD RECOGNITION

There are several reasons to believe that visual and spoken word recognition could share more mechanisms than is usually appreciated. To be sure, very salient differences exist between the visual and auditory modalities. One signal has a temporal dimension, the other is spatially extended. The former travels sequentially (over time) through the cochlear nerve, the latter in parallel through the optic nerve. In addition, just as in spoken word recognition, researchers in the field of visual word recognition have to ponder an invariance problem. Although a unique fixation near the center of a word is usually enough for an adult to recognize it (Starr and Rayner, 2001), ultimately this fixation has only stochastic precision and will rarely bring the same stimulus twice at exactly the same place on the retina, resulting in dissimilar retinal patterns. A credible model of the visual word recognition system should find a way to overcome this disparity in a word's many location exemplars, and to summon a unique lexical meaning and a unique phonology independently of wherever the visual stimulus actually fell on the retina.

### 1.3. STRING KERNELS

In the machine learning literature, one computational technique that has been very successful at comparing sequences of symbols independently of their position goes under the name of string kernels (Hofmann et al., 2008). Symbols could be amino-acids, nucleotides, or letters in a webpage: in every case the gist of string kernels is to represent strings (such as "TIME") as points in a high-dimensional space of symbol combinations (for instance as a vector where each component stands for a combination of two symbols, and only the components for "TI," "TM," "TE," "IM," "IE," "ME" would be non-zero). It is known that this space is propitious to linear pattern separations and yet can capture the (domain-dependent) similarities between them. String kernels have also been very successful due to their computability: it is not always necessary to explicitly represent the structures in the space of symbol combinations in order to compute their similarity (the so-called "kernel trick," which we will not use here).

It has been argued that string kernels provide a very good fit to several robust masked priming effects in visual word recognition, such as for instance letter transposition effects (the phenomenon that a letter transposition like *trasnpose* better primes the original word than a stimulus with letter replacements, such as *tracmpose*), and are thus likely involved at least in the early stages of visual word encoding (Hannagan and Grainger, 2012). To our knowledge, however, there have been no published investigations of string kernels in the domain of spoken word recognition. While the notion of an open biphone may at first blush sound implausible, keep in mind that the open bigram string kernel approach affords spatial invariance for visual word recognition. Might it also provide a basis for temporal invariance for spoken words?

## 2. TISK, THE TIME INVARIANT STRING KERNEL MODEL OF SPOKEN WORD RECOGNITION: MATERIALS AND METHODS

### 2.1. GENERAL ARCHITECTURE AND DYNAMICS

Our extension of the string kernel approach to spoken words is illustrated in **Figure 3**. It uses the same lexicon and basic activation dynamics as the TRACE model, but avoids a massive reduplication of units, as it replaces most time-specific units from TRACE with time-invariant units. It is comprised of four levels: inputs, phonemes, nphones (single phones and diphones) and words. Inputs consist of a bank of time-specific input units as in TRACE, through which a wave of transient activation travels. However, this input layer is deliberately very simplified compared to its TRACE analog. The input is like the Dandurand et al. (2010) input layer, though in our case, it is a time slice × phoneme matrix rather than a spatial slot × letter matrix. Thus, for this initial assay with the model, we are deferring an implementation like TRACE's pseudo-spectral featural level and the details it affords (such as TRACE's rough analog to coarticulation, where feature patterns are extended over time and overlap). With our localist phoneme inputs, at any time there is always at most one input unit active—inputs do not overlap in time, and do not code for phonetic similarity (that is, the inputs are orthogonal localist nodes). Note that the use of time-specific nodes at this level is a matter of computational convenience without theoretical commitment or consequence; these nodes provide a computationally

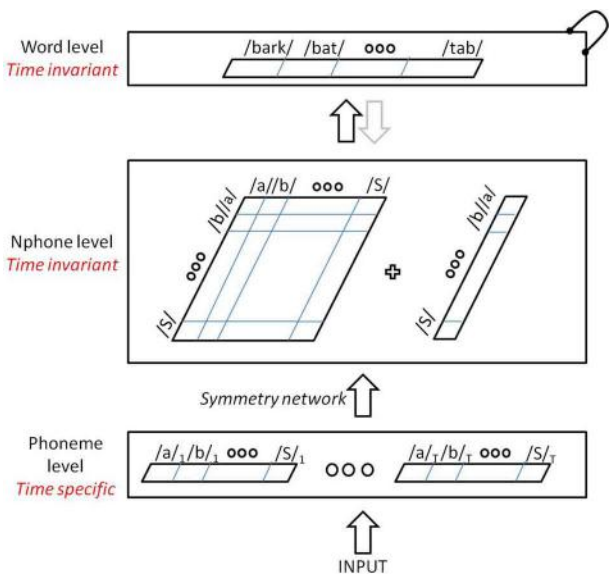



**FIGURE 3 | The TISK model—a time-invariant architecture for spoken word recognition.**

expedient way to pass sequences of phonemic inputs to the model, and could conceivably be replaced by a single bank of input nodes (but this would require other additions to the model to allow inputs to be "scheduled" over time). As in the TRACE model, one can construe these input nodes as roughly analogous to echoic memory or a phonological buffer. As we shall see, these simplifications do not prevent the model from behaving remarkably similarly to TRACE.

For our initial simulations, the model is restricted to ten slices (the minimum number needed for single-word recognition given the original TRACE lexicon), each with 14 time-specific phoneme units (one for each of the 14 TRACE phonemes). The input phoneme units feed up to an nphone level with one unit for every phoneme and for every ordered pairing of phonemes. The nphone units are time-invariant; there is only one /d/ unit at that level and only one /da/ diphone unit. Finally, nphone units feed forward to time-invariant (one-per-word) lexical units.

A critical step in the model is the transition between the time-specific phoneme input level and the time-invariant nphone level. This is achieved via entirely feedforward connections, the weights of which are set following certain symmetries that we will describe shortly. The nphone level implements a string kernel and consists of 196 + 14 units, one for each possible diphone and phoneme given the TRACE inventory of 14 phonemes. Units at this level can compete with one another via lateral inhibition, and send activation forward to the time invariant word level through excitatory connections, whose weights were normalized by the number of nphones of the destination word. The word level consists of 212 units (the original TRACE lexicon), with lateral inhibitory connections only between those words that share at least one phoneme at the level below. For this preliminary investigation, feedback connections from words to nphones were not included.

Units in the model are leaky integrators: at each cycle $t$, the activation $A_i$ of unit $i$ will depend on the net input it receives and on its previous activation, scaled down by a decay term, as described in Equation (1):

$$A_i(t) = \begin{cases} A_i(t-1) * (1 - \text{Decay}) \\ \quad + \text{Net}_i(t) * (1 - A_i(t-1)), & \text{if Net}_i > 0 \\ A_i(t-1) * (1 - \text{Decay}) \\ \quad + \text{Net}_i(t) * A_i(t-1), & \text{if Net}_i \leq 0 \end{cases} \quad (1)$$

where the net input of unit $i$ at time $t$ is given by:

$$\text{Net}_i = \sum_{j=1}^{k} w_{ij} A_j(t) \quad (2)$$

Python code for the model is available upon request to the first author, and a list of parameters is provided below as supplemental data. In the next section, we describe in detail the connections between time-specific phonemes and time-invariant nphones.

### 2.2. FROM TIME-SPECIFIC TO TIME-INVARIANT UNITS: A SYMMETRY NETWORK FOR PHONOLOGICAL STRING KERNELS

We now describe the transition phase between time-specific phonemes and time-invariant nphones in the TISK model. It is clear that unconstrained (that is, unordered) "open diphone" connectivity would be problematic for spoken words; for example, if *dog* and *god* activated exactly the same diphones (/da/, /dg/, /ag/, /ga/, /gd/, /ad/), the system would be unable to tell the two words apart. The challenge is to activate the correct diphone /da/, but not /ad/, upon presentation of a sequence of phonemes like $[/d/_t, /a/_{t+1}]$, that is, phoneme /d/ at time $t$ and phoneme /a/ subsequently. Thus, the goal is to preserve activation of non-adjacent phonemes as in an open diphone scheme (for reasons explained below) with the constraint that only observed diphone sequences are activated—that is, *dog* should still activate a /dg/ diphone (as well as /da/ and /ag/) because those phonemes have been encountered in that sequence, but not /gd/, while *god* should activate /gd/ but not /dg/. This would provide a basis for differentiating words based on sequential ordering without using time-specific units "all the way up" through the hierarchy of the model.

The issue of selectivity (here, between "anadromes": diphones with the same phonemes in different order) vs. invariance (here, to position-in-time) has long been identified in the fields of visual recognition and computer vision, and has recently received attention in a series of articles investigating invariant visual word recognition (Dandurand et al., 2010, 2013; Hannagan et al., 2011).

Directly relevant to this article, Dandurand et al. (2013) trained a simple perceptron network (that is, an input layer directly connected to an output layer, with weights trained using the delta rule) to map location-specific strings of letters to location-invariant words. To their surprise, not only did this simplistic setup succeed in recognizing more than 5000 words, a fair fraction of which were anagrams, it also produced strong transposition effects. By introducing spatial variability—the "i" in *science* could occur in many different absolute positions rather than just one—tolerance for slight misordering in relative position emerged. When Dandurand et al. (2013) investigated how the network could possibly succeed on this task in the absence of hidden unit representations, they observed that during the course of learning, the "Delta learning rule" had found an elegant and effective way to keep track of letter order by correlating connection strengths with the location of the unit. More precisely, the connections coming from all "e" units and arriving at word *live* had their weights increasing with the position, whereas the connections from the same units to the word *evil* had their weights decreasing with position. In this way, connection weights became a proxy for the likelihood of a word given all letters at all positions. This simple scheme enabled the network to distinguish between anagrams like *evil* and *live*. We describe next how this solution found by the delta rule can be adapted to map time-specific phonemes to time-invariant diphones or single phonemes.

The network in **Figure 4** has two symmetries: firstly, weights are invariant to changes in input phoneme *identity* at *any given time*. This is manifest in **Figure 4** by the symmetry along the medial vertical axis: for any $t$, $a_t$ and $b_t$ can exchange their weights. Secondly, weights are invariant to changes in input phonemes *identity* across *opposite times* (in **Figure 4**), a central symmetry with center midway through the banks of input phonemes: for any $t \leq T$, $a_t$ and $b_{T-t}$ are identical, and so are

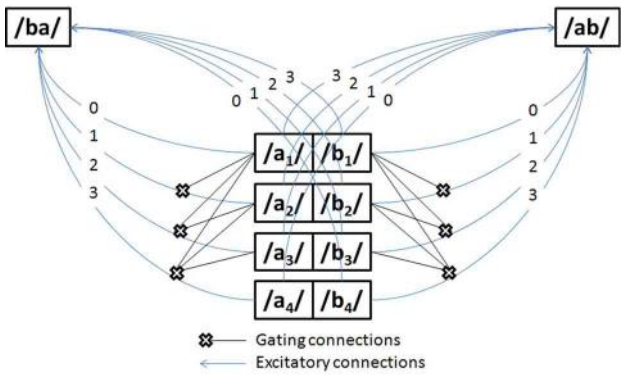



**FIGURE 4 | A symmetry network for time-invariant nphone recognition that can distinguish anadromes.** The units in the center of the diagram (e.g., $/a/_1$) represent time-specific input nodes for phonemes */a/* and */b/* at time steps 1–4. The */ba/* and */ab/* nodes represent time-invariant diphone units.

$b_t$ and $a_{T-t}$. Although the first symmetry concerns both excitatory (arrows) and gating connections (crosses, which will be shortly explained), the second symmetry concerns only excitatory connections.

What is the point of these symmetries? Consider a network where the weights have been set up as in **Figure 4**. Then at all possible times *t*, presenting the input sequence $[/a/_t, /b/_{t+1}]$ by clamping the appropriate units to 1 will always result in a constant net input for */ab/*, here a net input of 4, and it will always result in a smaller constant net input to */ba/*, here a net input of 2. A common activation threshold for every diphone unit can then be set anywhere between these two net inputs (for instance, a threshold of 3), that will ensure that upon perceiving the sequence $[/a/_t, /b/_{t+1}]$ the network will always recognize */ab/* and not */ba/*. The same trick applies for the complementary input sequence $[/b/_t, /a/_{t+1}]$, by setting the weights from these phoneme units to the transposed diphone */ba/* in exactly the opposite pattern. A subtlety, however, is that in order to prevent sequences with repeated phonemes like $[/b/_1, /a/_2, /b/_3]$ from activating large sets of irrelevant nphones like */br/* or */bi/*, it is necessary to introduce gating connections (cross-ended connections in **Figure 4**), whereby upon being activated, unit $/b/_1$ will disable the connection between all future $/b/_t > 1$ and all diphones */*b/* (where "∗" stands for any phoneme but *b*).

The use of gating connections is costly, as the number of connections needed is proportional to the square of the number of time slices, but less naïve gating mechanisms exist with explicit gating units that would be functionally equivalent at a much smaller cost (linear with increasing numbers of time slices). More generally, other mappings between time-specific phonemes and time-invariant n-phones are possible. However, our approach is cast within the theory of symmetry networks (Shawe-Taylor, 1993), which ensures that several mathematical tools are available to carry out further analysis. The particular symmetry network introduced here arguably also has a head-start in learnability, given that it builds on a solution found by the delta rule. Specifically, in a perceptron trained to recognize visual words (Dandurand et al., 2013), the Delta rule found the "central symmetry through time" visible in **Figure 4**. We do not know if pressure to represent temporal sequences would allow the model to discover the "axial" symmetry and necessity for gating connections, but this is a question we reserve for future research. We note that some studies have reported the emergence of symmetry networks in more general settings than the delta rule and word recognition, that is, under unsupervised learning algorithms and generic visual inputs (Webber, 2000). Perhaps the best argument for this architecture is that it is reliable, and allows for the activation of the kind of "string kernels" recently described by Hannagan and Grainger (2012), at a computational cost that can be regarded as an upper-bound and yet is not prohibitive.

## 3. RESULTS

### 3.1. PERFORMANCE ON SINGLE WORD RECOGNITION

We begin with a comparison of TISK and TRACE in terms of the recognition time of each word in the original 212-word TRACE lexicon. If TISK performs like TRACE, there should be a robust correlation between the recognition time for any particular word in the two models. We operationalized spoken word recognition in three different ways: an absolute activation threshold ($R_{\text{abs}}$), a relative activation threshold ($R_{\text{rel}}$) and a time-dependent criterion ($R_{\text{tim}}$). The first criterion states that a word is recognized if its activation reaches an absolute threshold, common to all words. In the second criterion, recognition is granted whenever a word's activation exceeds that of all other words by a threshold (0.05 in the simulations). Finally the time-dependent criterion defines recognition as a word's activation exceeding that of all other words for a certain number of cycles (10 cycles in the simulations).

Spoken word recognition accuracy for TRACE is consistently greater than that for TISK in these simulations, although both models obtain high performance under all criteria. TRACE exhibits close to perfect recognition with the three criteria ($T_{\text{abs}} =$ 97%, $T_{\text{rel}} = 99\%$, $T_{\text{tim}} = 99\%$). TISK on the other hand operates less well under an absolute criterion, but recognition is improved using a relative threshold, and it rises to TRACE-like level with a time-dependent threshold ($T_{\text{abs}} = 88\%$, $T_{\text{rel}} =$ 95%, $T_{\text{tim}} = 98\%$). Also, mean recognition cycles are similar for TRACE ($T_{\text{abs}} = 38$ cycles, $T_{\text{rel}} = 32$ cycles, $T_{\text{tim}} = 40$ cycles) and for TISK ($T_{\text{abs}} = 45$ cycles, $T_{\text{rel}} = 38$ cycles, $T_{\text{tim}} = 40$ cycles). At the level of individual items, performance is very similar for the two models, as revealed by high correlations between recognition times (for correctly recognized items) under all three recognition definitions (r for each definition: $T_{\text{abs}} = 0.68$, $T_{\text{rel}} = 0.83$, $T_{\text{tim}} =$ 0.88). **Figure 5** illustrates the correlation between response times in the case of $T_{\text{tim}}$. In the rest of this article we will use the time-dependent criterion $T_{\text{tim}}$, as the one with which models achieved both the best performance and the most similar performance.

It is also instructive to consider the two words on which TISK failed, /triti/ (*treaty*) and /stˆdid/ (*studied*). Indeed the model confused these words with their respective embedded cohort competitors /trit/ (*treat*) and /stˆdi/ (*study*). For the model these are the most confusable pairs of words in the lexicon, because in each case almost exactly the same set of nphones is activated for the target and the cohort competitor, except for one or two n-phones (the only additional diphone for *treaty* compared to

*treat* is /*ii*/; *studied* activates two additional diphones compared to *study*: /dd/ and /id/). It is certainly possible to fine-tune TISK so as to overcome this issue. Note also that TISK recognizes correctly the vast majority of words containing embeddings, including word-onset embeddings.

But these particular failures are perhaps more valuable in that they point to the type of learning algorithm that could be used in the future, in TISK as in TRACE, to find the connection weights in a more principled manner. Namely, they strongly suggest that a learning algorithm should attribute more weight to these connections that are the most diagnostic given the lexicon (e.g., connection /ii/ to /triti/).

### 3.2. TIME COURSE OF LEXICAL COMPETITORS

As previously observed, what is impressive about the TRACE model is less its ability to recognize 212 English words than the way it does so, which captures and explains very detailed aspects of lexical competition in human spoken word recognition. Consider the so-called "Visual World Paradigm" (Tanenhaus et al., 1995), in which subjects' eye movements are tracked as they follow verbal instructions to manipulate items in a visual display. When the items include objects with similar sounding names (e.g., so-called "cohort" items with the same word onset, such as *beaker* and *beetle*, or rhymes, such as *beaker* and *speaker*) as well as unrelated items to provide a baseline, eye movements provide an estimate of activation of concepts in memory over time. That is, the proportion of fixations to each item over time maps directly onto phonetic similarity, with early rises in fixation proportions to targets and cohorts and later, lower fixation proportions to rhymes (that are still fixated robustly more than unrelated items; Allopenna et al., 1998). Allopenna et al. also conducted TRACE simulations with items analogous to those they used with human subjects, and found that TRACE accounted for more than 80% of the variance in the over-time fixation proportions.

In order to assess how TISK compares to TRACE in this respect, we subjected the model to simulations analogous to those used by Allopenna et al. (1998). However, rather than limiting the simulations to the small subset of the TRACE lexicon used by Allopenna et al., we actually conducted one simulation for every (correctly recognized) word in the TRACE lexicon with both TRACE and TISK. We then calculated average target activations over time, as well as the over-time average activation of all cohorts of any particular word (words overlapping in the first two phonemes), any rhymes, and words that embed in the target (e.g., for *beaker*, these would include *bee* and *beak*, whereas for *speaker*, these would be *speak*, *pea*, *peek*). Rather than selecting a single word to pair with each word as its unrelated baseline, we simply took the mean of all words (including the target and other competitors); because most words are not activated by any given input, this hovers near resting activation levels (−0.2 for TRACE, 0 for TISK). The results are shown in **Figure 6**.

Readers familiar with the Allopenna et al. article will notice some differences in our TRACE simulation results compared to theirs. First, we have activations below zero, while they did not. This is because Allopenna et al. followed the standard practice of treating negative activations as zero. Second, our rhyme activations remain below zero, even though they are robustly higher than those of the mean activation baseline. Having robustly positive rhyme activations in TRACE requires the use of a carrier phrase like the one used by Allopenna et al. (or a transformation to make all activations above resting level positive); without this, because there is strong bottom–up priority in TRACE, cohorts will be so strongly activated that rhyme activation will be difficult to detect. However, what really matters for our purposes is the relative activations of each competitor type, which are clearly consistent between the two models.

### 3.3. LEXICAL FACTORS INFLUENCING RECOGNITION

Let's return to item level recognition times. We can probe the models more deeply by investigating how recognition times vary

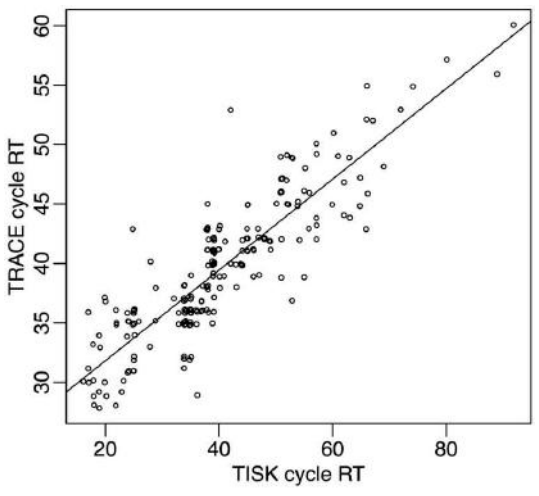



**FIGURE 5 | Response times in TISK (*x*-axis) and TRACE (*y*-axis) for all 212 words in the lexicon, when a time threshold is used for recognition.**

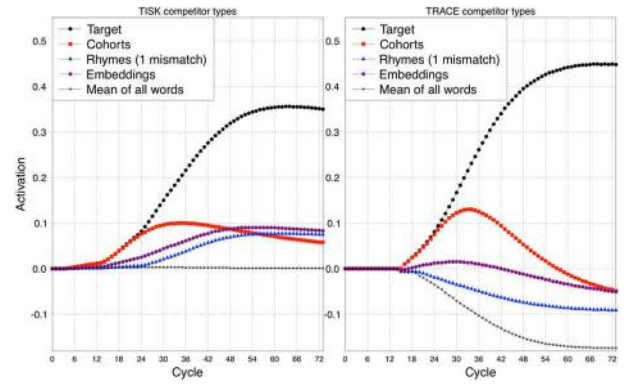



**FIGURE 6 | Comparison between TISK (left panel) and TRACE (right panel) on the average time-course of activation for different competitors of a target word.** Cohort: initial phonemes shared with the target. Rhymes (1 mismatch): all phonemes except the first shared with the target. Embeddings: words that embed in the target. The average time course for all words (Mean of all words) is presented as a baseline.

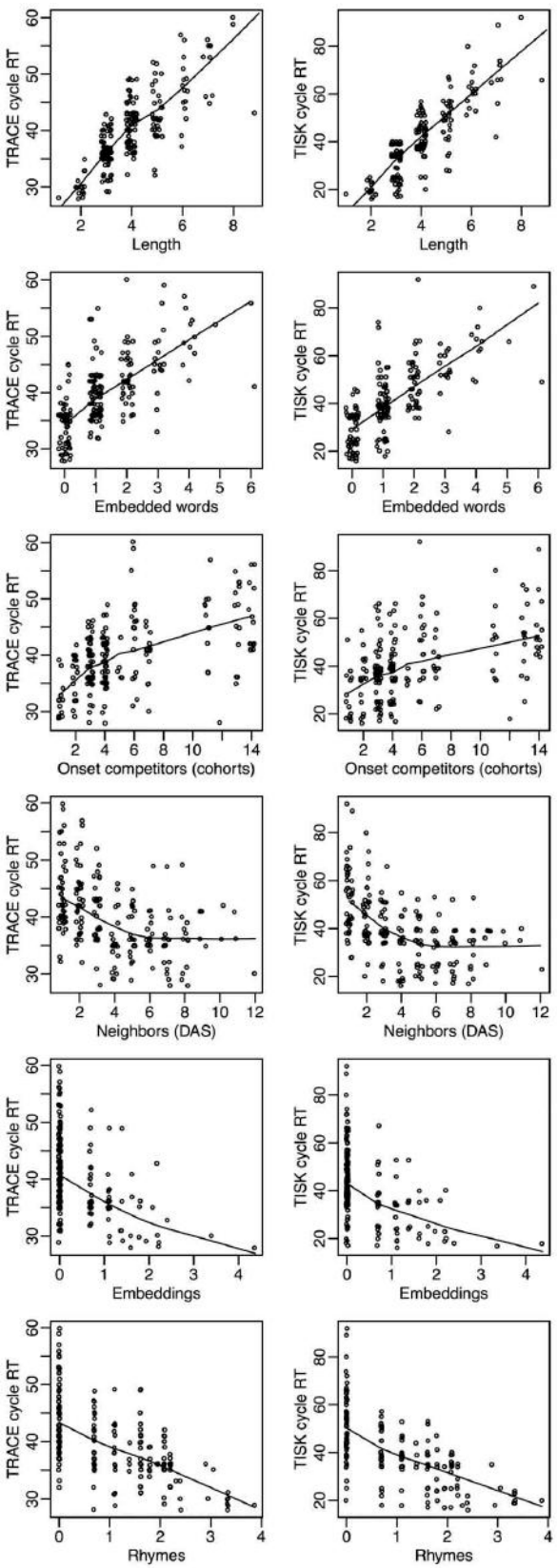



**FIGURE 7 | Continued**

**FIGURE 7 | An overview of how recognition cycles correlate with other lexical variables in TRACE (left column) and in TISK (right column).** Length: target length. Embedded words: number of words that embed in the target. Onset competitors (Cohorts): number of words that share two initial phonemes with the target. Neighbors (DAS): count of deletion/addition/subsitution neighbors of the target. Embeddings: logarithm of the number of words the target embeds in. Rhymes: logarithm of the number of words that overlap with the target with first phoneme removed.

in each model with respect to the lexical dimensions that have attracted the most attention in the spoken word recognition literature. **Figure 7** presents the correlation between recognition cycles and six standard lexical variables: the length of the target (Length), how many words it embeds in (Embeddings), how many words embed in it (Embedded), how many deletion/addition/subsitution neighbors it has (Neighbors), the number of words with which it shares 2 initial phonemes (Cohorts), and the number of words that overlap with it when its first phoneme is removed (Rhymes).

**Figure 7** shows that among the six lexical dimensions considered, three are inhibitory dimensions (Length, Embedded words and Cohorts) and three are clearly facilitatory dimensions (Neighbors, Embeddings, and Rhymes). Crucially, precisely the same relationships are seen for both models, with an agreement that is not only qualitative but also quantitative.

Facilitatory variables are perhaps the most surprising, as neighborhood has long been construed as an inhibitory variable for spoken word recognition. Although the precise details are not relevant for this initial presentation of TISK, further inspection of these neighborhood effects reveals that there is an interaction of neighborhood with word length; for longer words, neighbors begin to have a facilitative effect. The crucial point is that one can see that TRACE and TISK behave in remarkably similar ways—and both make intriguing, even counter-intuitive, but testable predictions.

### 3.4. COMPUTATIONAL RESOURCES

We will end this comparison with an assessment of the resources needed in both models. **Table 1** shows the number of connections and units in TRACE and TISK, as calculated in Appendix C. The figures for TRACE are obtained by considering the number of units required per slice in the model (starting from the phoneme level, for a fair comparison with TISK which doesn't use a featural level): 14 phonemes, and, in the basic TRACE lexicon, 212 words, for 226 units. Now assuming an average of 3 phonemes per word, and allowing for connections between units at adjacent time slices, TRACE needs approximately 225,000 connections per time slice. If we make the trace 200 time slices long (which assuming 10 ms per slice would amount to 2 s, the duration of echoic memory), we need approximately 15,000 units and 45 million connections. Increasing the lexicon to a more realistic size of 20,000 words and the phoneme inventory to 40, these figures reach approximately 1.3 million units and 400 billion connections.

Next let us consider the situation in TISK. With a 2 s layer of time-specific input units (again, corresponding to the approximate limit of echoic memory), 14 phonemes and 212 words as in

**Table 1 | Estimates of the number of units and connections required in TRACE and TISK for 212 or 20,000 words, 14 or 40 phonemes, an average of four phonemes per word, and assuming 2 s of input stream.**

| | 212 words 14 phonemes | | 212 words 40 phonemes | | 20,000 words 40 phonemes | |
|---|---|---|---|---|---|---|
| | TRACE | TISK | TRACE | TISK | TRACE | TISK |
| Units | 15, 067 | 3222 | 16, 800 | 9852 | 1, 336, 000 | 29, 640 |
| Connections | 45, 049, 733 | 3, 737, 313 | 45, 401, 600 | 31,718,357 | >4E + 11 | 348, 783, 175 |

TRACE, TISK requires 3.2 thousand units and 3.7 million connections. This represents a 5-fold improvement over TRACE for units, and a 15-fold improvement for connections. With 20,000 words and 40 phonemes, TISK would require approximately 29,000 units (TRACE requires 45 times more) and 350 million connections (TRACE requires 1.1 thousand times more).

**Figure 8** presents an overview of the number of connections as a function of trace duration (number of time slices) and lexicon size in TISK and in TRACE. The most striking feature already apparent in **Table 1** is that TRACE shows an increase in connections which dwarfs the increase in TISK. But **Figure 8** also shows that in TRACE this increase is quadratic in lexicon size and steeply linear in time slices, while connection cost in TISK looks linear in both variables with very small slopes. While Appendix B demonstrates that both functions are actually quadratic in the number of words (due to lateral inhibition at the lexical level in both models), there is still a qualitative difference in that the quadratic explosion due to the word level is not multiplied by the number of time slices in TISK, like it is in TRACE—decoupling trace duration and lexicon size was, after all, the whole point of this modeling exercise.

What is the significance of this computational economy for spoken word recognition? We would argue that it makes it easier to examine the behavior of the model at large scales. The 400 billion connections required in TRACE currently discourage any direct implementation with a realistic lexicon. However, word recognition behavior in IA models like TRACE and TISK is exquisitely sensitive to the nature of lexical competition. One should therefore not be content with effects obtained using an artificial sample of 200 words but should aim at running the model on the most realistic lexicon possible.

Depending on the precise linking assumptions one is willing to make between units and connections on the one hand, and actual neurons and synapses on the other hand (see, for instance, de Kamps and van der Velde, 2001 for a well-motivated attempt), one may or may not find that for some large but still reasonable lexicon size the connection cost in TRACE becomes larger than the sum total of all available synapses in the brain, whereas **Figure 8** and Appendix B suggest that the cost in TISK would be orders of magnitude smaller and may barely make a dent in the synaptic budget.

But even leaving aside this possibility, the notion that wiring cost should come into consideration when modeling cognitive systems appears to be rather safe. Firing neurons and maintaining operational synapses has a high metabolic cost, and the pressure to perform such a ubiquitous task as spoken word recognition would seem to demand an implementation that balances cost and efficiency in the best possible way. Although the connections in TRACE or TISK are meant to be functional rather than biological, metabolic costs at the biological level constrain connectivity at the functional level: numerous functional networks as derived from human brain imaging achieve economical trade-offs between wiring cost and topological (connectivity) efficiency (Bullmore and Sporns, 2012). Recent investigations with artificial neural networks have also shown that minimizing the number of connections can improve performance by favoring the emergence of separate levels of representations (Clune et al., 2006).

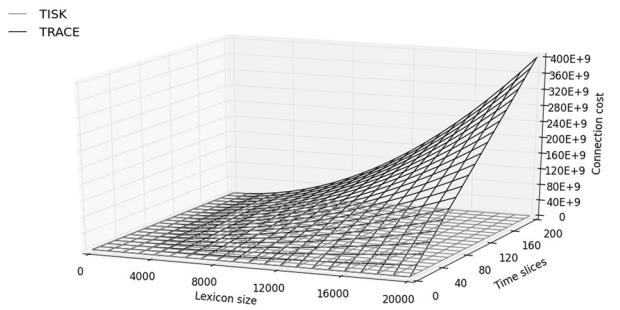



**FIGURE 8 | Number of connections (*y*-axis, "connection cost") as a function of time slices and lexical size in TISK (gray surface) and TRACE (black surface).**

## 4. DISCUSSION

### 4.1. SPOKEN AND VISUAL WORD RECOGNITION: A BRIDGE BETWEEN ORTHOGRAPHY AND PHONOLOGY

In 1981, McClelland and Rumelhart presented an interactive-activation model of visual word recognition that was to be a major inspiration for the TRACE model of spoken word recognition (McClelland and Elman, 1986) and an inspiration for future generations of reading researchers. Most important is that in **Figure 1** of their article, McClelland and Rumelhart sketched an overall architecture for visual and auditory word perception, describing interconnections between the two in the form of reciprocal letter-phoneme connections. This architecture clearly predicts that visual word recognition should be influenced on-line by phonological knowledge and spoken word recognition should be influenced by orthographic knowledge. Support for these predictions has since been provided by a host of empirical investigations (see Grainger and Ziegler, 2008 for a review). Strangely enough, however, attempts to implement such a bi-modal architecture have been few and far between. Research on visual word recognition has come the closest to achieving this, with the development

of computational models that include phonological representations (Seidenberg and McClelland, 1989; Plaut et al., 1996; Coltheart et al., 2001; Perry et al., 2007).

With respect to spoken word recognition, however, to our knowledge no computational model includes orthographic representations, and although our TISK model of spoken word recognition is not an improvement in this respect, it was nevertheless designed with the constraint of eventually including such representations in mind. As such, TISK not only provides an answer to McClelland and Elman's question of how to avoid duplication in TRACE, but also picks up on McClelland and Rumelhart's challenge to develop a truly bimodal model of word recognition. One model has been developed along the lines initially suggested by McClelland and Elman (1986)—this is the bimodal interactive-activation model (Grainger et al., 2003; Grainger and Holcomb, 2009), recently implemented by Diependaele et al. (2010). Future extensions of this work require compatibility in the way sublexical form information is represented for print and for speech. The present work applying string kernels to spoken word recognition, along with our prior work applying string kernels to visual word recognition (Hannagan and Grainger, 2012), suggest that this particular method of representing word-centered positional information provides a promising avenue to follow. Indeed, string kernels provide a means to represent order information independently of whether the underlying dimension is spatial or temporal, hence achieving spatial invariance for visual words and temporal invariance for spoken words.

### 4.2. TESTING FOR TEMPORAL INVARIANCE IN SPOKEN WORD RECOGNITION

Researchers interested in the neural representations for visual words are blessed with the Visual Word Form Area, a well-defined region in the brain that sits at the top of the ventral visual stream, and is demonstratively the locus of our ability to encode letter order in words or in legal non-words (Cohen et al., 2000; Gaillard et al., 2006) but is not selectively activated for spoken words. Until recently, the common view was that by the mere virtue of its situation in the brain, if not by its purported hierarchical architecture with increasingly large receptive fields, the VWFA was bound to achieve complete location invariance for word stimuli. However, recent fMRI studies show that, and computational modeling explains why, a significant degree of sensitivity to location is present in the VWFA (Rauschecker et al., 2012). A trained, functional model of location invariance for visual words explains why this can be so (Hannagan and Grainger, in press). In this model the conflicting requirements for location invariant and selectivity conspire with limited resources, and force the model to develop in a symmetry network with broken location symmetry on its weights (Hannagan et al., 2011). This in turn produces "semi-location invariant" distributed activity patterns, which are more sensitive to location for more confusable words (Hannagan and Grainger, in press). Thus brain studies have already been highly informative and have helped constrain our thinking on location invariance in visual words.

But attempts to proceed in the same way for the auditory modality quickly run into at least two brick walls. The first is that a clear homologue of the VWFA for spoken words has remained elusive. This might be because the speech signal varies in more dimensions than the visual signal corresponding to a visual object; a VWFA homologue for speech might need to provide invariance not just in temporal alignment, but also across variation in rate, speaker characteristics, etc. However, one study points to the left superior temporal sulcus as a good candidate for an Auditory Word Form Area (AWFA) on the grounds that this region only responded for auditory words and showed repetition suppression when the same word was spoken twice (Cohen et al., 2004), and there have been reports of invariance for temporal alignment or speaker characteristics and/or multi-dimensional sensitivity in the superior (Salvata et al., 2012) and medial (Chandrasekaran et al., 2011) temporal gyri. The second issue is that paradigms for testing temporal invariance are less easily designed than those which test location invariance in the visual case. Speculating from Rauschecker et al. (2012), however, we can propose a task that tests for the presence of time-specific word representations, in which subjects would be presented with a sequence of meaningless sounds where one spoken word would be embedded. By manipulating the position of this word in the sequence, one could then test whether a "blind" classifier could be trained to discriminate by their positions-in-time the different fMRI activation patterns evoked in the superior temporal sulcus. Because this decoding procedure can be applied to signals recorded from several disconnected regions of interest, this procedure would be agnostic to the existence of a well-circumscribed AWFA. TRACE and TISK both predict that the classifier should succeed with fMRI patterns evoked early on in the processing stream, i.e., at the time-specific phoneme level, but only TISK predicts that time-invariant representations should be found downstream, for lexical representations. Although the necessity for testing the existence of time-specific units is obvious in the light of the TISK model, we would argue that this has long been an urgent experimental question to ask. TRACE has been the most successful model of spoken word recognition for almost three decades now, and therefore it might be worth taking seriously the most striking hypothesis it makes of the existence of time-specific units, an hypothesis which even TISK does not succeed in completely avoiding at the phoneme level.

### 4.3. PREVIOUS MODELS AND ALTERNATIVE APPROACHES TO TEMPORAL ORDER

We claimed previously that TRACE has the greatest breadth and depth of any extant model of spoken word recognition. Of course, there are models whose proponents argue that they have solved key problems in spoken word recognition without using TRACE's inefficient time-specific reduplication strategy. We will review a few key examples, and consider how they compare with TRACE and TISK.

Norris (1994), Norris et al. (2000), and Norris and McQueen (2008) introduced Shortlist, Merge, and Shortlist B, the first two being IA network models and the latter a Bayesian model of spoken word recognition. All three models share basic assumptions, and we refer to them collectively as "the Shortlist models." Contrary to TRACE, the Shortlist models are entirely feedforward. They also make a critical distinction between words and tokens, the latter being time-specific entities that instantiate the

former, time-invariant lexical templates. The reduplication of the lexical level that afflicts TRACE is avoided in these models by assuming that only a "short list" of tokens is created and wired on-the-fly into a "lattice" of lexical hypotheses. These models have a sizable lexicon (even a realistic 20,000 word lexicon in the case of Shortlist B), and although they have not been applied to the full range of phenomena that TRACE has, they have successfully simulated phenomena such as frequency and neighborhood effects. Unfortunately, because no computational mechanism is described that would explain how the on-the-fly generation and wiring of tokens could be achieved, the account of spoken word recognition provided by Shortlist is still essentially promissory.

Other approaches to temporal order use fundamentally different solutions than TRACE's reduplication of time-specific units. Elman's (1990) simple recurrent network (SRN) may be foremost among these in the reader's mind. The SRN adds a simple innovation to a standard feedforward, backpropagation-trained two-layer network: a set of context units that provide an exact copy of the hidden units at time step *t*-1 as part of the input at time *t*, with fully connected, trainable weights from context to hidden units. This feedback mechanism allows the network to learn to retain (partial) information about its own state at preceding time steps, and provides a powerful means for sequence learning. However, while SRNs have been applied to speech perception and spoken word recognition (most notably in the Distributed Cohort Model: Gaskell and Marslen-Wilson, 1997, but for other examples see Norris, 1990, and Magnuson et al., 2000, 2003), so far as we are aware, no one has investigated whether SRNs can account for the depth and breadth of phenomena that TRACE does (though SRNs provide a possible developmental mechanism since they are learning models, and the Distributed Cohort Model has been applied to semantic phenomena beyond the scope of TRACE).

Another approach is the cARTWORD model of Grossberg and Kazerounian (2011), where activity gradients specific to particular sequences can differentiate orderings of the same elements (e.g., ABC vs. ACB, BAC, etc.). However, this mechanism cannot represent sequences with repeated elements (for example, it cannot distinguish ABCB from ABC, as the second B would simply provide further support for B rather than a second B event), which makes it incapable of representing nearly one third of English lemmas. Furthermore, it is premature to compare this approach to models like TRACE, since it has been applied to a single phenomenon (phoneme restoration) with just a few abstract input nodes and just a few lexical items; thus, we simply do not know whether it would scale to handle realistic acoustic-phonetic representations and large lexicons, let alone the broad set of phenomena TRACE accounts for (see Magnuson, submitted, for detailed arguments and simulations showing that the supposed failures of TRACE to account for phoneme restoration phenomena reported by Grossberg and Kazerounian, 2011, were the result of flawed simulations, not a problem with TRACE). Note that a similar activity gradient approach in visual word recognition (Davis, 2010) has also been attempted, with similar limitations.

### 4.4. THE UTILITY OF INTERACTIVE ACTIVATION MODELS

Because spoken word recognition is a slowly acquired skill in humans, any model of it should eventually strive to incorporate some kind of learning algorithm that explains how the representations necessary to solve the task have matured. Unlike SRNs though, models such as TRACE and TISK do not comply to this requirement. On the other hand and until proven the contrary TRACE vastly outperforms SRNs in explanatory power while having the advantage of being more transparent. We would argue that IA models and learning models like SRNs should be construed as complementary approaches to spoken word recognition. Imagine SRNs were demonstrated to account for similar depth and breadth as TRACE. We would still be left with the puzzle of how they do so. Unpacking the complex composites of cooperative and competitive wiring patterns that would develop would be no mean feat. This is where we find interactive activation models like TRACE and TISK especially useful. The IA framework allows one to construct models with levels of organization (the representational levels) with inter- and intralevel interaction governed by discrete parameters. This allows one to generate hypotheses about which aspects of the model are crucial for understanding some phenomenon (e.g., by investigating which model parameters most strongly generate a key behavior), or about which level of organization may be perturbed in a particular language disorder (Perry et al., 2010; Magnuson et al., 2011). One modeling approach that is likely to be productive is to use simpler frameworks like IA models to generate hypotheses about key model components in some behavior or disorder, and then to seek ways that such behaviors or disruptions might emerge in a more complex model, such as an SRN or another type of attractor network (cf. Magnuson et al., 2012). Similarly, TISK provides a testbed for investigating whether a string kernel scheme is a robust basis for spoken word recognition. For example, the success of string kernel representations in TISK might suggest that we should investigate whether the complex wiring SRNs learn approximates string kernels.

### 4.5. RELATIONSHIP BETWEEN TRACE AND TISK

One might be surprised that TISK and TRACE display such similar behavior despite the lack of feedback in the former and its presence in the latter. Feedback in models of spoken word recognition is a controversial topic (McClelland et al., 2006; McQueen et al., 2006; Mirman et al., 2006a), which we do not address here; our aim is to see whether a model with a radically simpler computational architecture compared to TRACE can (begin to) account for a similar range of phenomena in spoken word recognition. However, this resemblance despite feedback is less surprising than it may seem. Indeed, it has been known for several years that the feedback contribution to word recognition in TRACE is limited given noise-free input (Frauenfelder and Peeters, 1998): simulations show that feedback makes the model more efficient and robust against noise (Magnuson et al., 2005). It also provides an implicit sensitivity to phonotactics—the more often a phoneme or n-phone occurs in lexical items, the more feedback it potentially receives—and it is the mechanism by which top–down lexical effects on phoneme decisions are explained in TRACE. None of these effects were considered in this article, which focused on core word recognition abilities and lexical competition effects. We acknowledge that without feedback, TISK will not be able to simulate many top–down phenomena readily

simulated in TRACE. Future research with TISK will explore the impact of feedback connections.

### 4.6. LIMITATIONS AND NEXT STEPS

The aim of this study was to improve on one particularly expensive aspect of the TRACE model without drastically affecting its lexical dynamics, or diminishing its explanatory power. We have demonstrated that a radically different approach to sequence representation, based on string kernels, provides a plausible basis for modeling spoken word recognition. However, our current model has several obvious limitations.

First, to apply TISK to the full range of phenomena to which TRACE has been applied will require changes, for example, in the input representations for TISK. As we mentioned above, we used single-point inputs for TISK rather than the on- and off-ramping, over-time inputs in TRACE that also give the model a coarse analog to coarticulation. An input at least this grain will be required to apply TISK to, for example, subcategorical mismatch experiments that TRACE accounts for (Dahan et al., 2001b).

Second, TISK's levels and representations are stipulated rather than emergent. Our next step will be to examine whether codes resembling string kernels emerge when intra-level weights are learned rather than stipulated. What learning algorithm could find the set of weight values under which TISK and TRACE have been shown to achieve close to perfect recognition? Is there more than one such set, and do they make different predictions from the existing fine-tuned solutions? There are a few results that suggest the way forward. For instance, there are demonstrations that Hebbian learning applied at the lexical level in TRACE can help explain short term phenomena in spoken word recognition (Mirman et al., 2006b). If Hebbian learning is indeed active on short scales, there are no reasons to doubt that it will be involved on longer time-scales, slowly shaping the landscape of inhibition between words, which forms the basis for much of the behaviors explored in this article.

Third, a problem shared by all models of word recognition is that it is not clear how to scale from a model of word recognition to higher levels, e.g., to a model of sentence comprehension. Because TISK's word level is time-invariant, there is no obvious way to generate ngrams at the word level. However, TISK and TRACE, like other models capable of activating a series of words over time given unparsed input (i.e., word sequences without word boundary markers) should be linkable to parsing approaches like "supertagging" (Bangalore and Joshi, 1999; Kim et al., 2002) or the self-organizing parser (SOPARSE) approach of Tabor et al. (e.g., Tabor and Hutchins, 2004). Note that a common intuition is that SRNs provide a natural way of handling sequential inputs from acoustics to phonemes to words. However, it is not clear that this translates into a comprehensive model of the entire speech chain. It is not apparent that you could have a single recurrent network that takes in acoustics and somehow achieves syntactic parsing (let alone message understanding) while producing human-like behavior at phonetic, phonological, lexical levels. These are non-trivial and unsolved problems, and despite the intuitive appeal of recurrent networks, remain unanswered by any extant model.

Finally, it is notable that we have not implemented feedback yet in TISK. This renders TISK incapable of accounting for top–down lexical effects on phoneme decisions. However, as Frauenfelder and Peeters (1998) and Magnuson et al. (2005) have demonstrated, feedback plays little role in recognition given clear inputs. When noise is added to a model like TRACE, feedback preserves speed and accuracy dramatically compared to a model without feedback. While feedback also provides a mechanistic basis for understanding top–down effects, it is also remarkable that at least one effect attributed to feedback in TRACE (rhyme effects; Allopenna et al., 1998) emerges in TISK without feedback. This suggests that in fact examining which, if any (other), putatively top–down effects emerge without feedback in TISK will be a useful enterprize. Given, however, the remarkable fidelity to TRACE that TISK demonstrates over a broad swath of phenomena, it is clear that feedback need not be included in this first assay with TISK.

## 5. CONCLUSION

Twenty-seven years after Elman and McClelland introduced the TRACE model, we have endeavored to answer the question of how to dispense with time-duplication, and have presented an alternative that preserves TRACE-like performance on spoken word recognition while using orders of magnitude less computational resources. Perhaps more importantly, the particular structures and mechanisms that achieve time-invariance in TISK construct new and intriguing bridges between visual and spoken word recognition.

## FUNDING

Thomas Hannagan and Jonathan Grainger were supported by ERC research grant 230313.

## ACKNOWLEDGMENTS

We thank Emily Myers, Lori Holt, and David Gow Jr., for stimulating discussions.

## REFERENCES

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition: evidence for continuous mapping models. *J. Mem. Lang.* 38, 419–439. doi: 10.1006/jmla.1997.2558

Bangalore, S., and Joshi, A. (1999). Supertagging: an approach to almost parsing. *Comput. Linguisti.* 25, 238–265.

Bowers, J. S., Damian, M. F. E., and Davis, C. J. (2006). A fundamental limitation of the conjunctive codes learned in PDP models of cognition: comments on Botvinick and Plaut. *Psychol. Rev.* 116, 986–997. doi: 10.1037/a0017097

Bullmore, E., and Sporns, O. (2012). The economy of brain network organization *Nat. Rev. Neurosci.* 13, 336–349.

Chandrasekaran, B., Chan, A. H. D., and Wong, P. C. M. (2011). Neural processing of what and who information during spoken language processing. *J. Cogn. Neurosci.* 23, 2690–2700. doi: 10.1162/jocn.2011.21631

Clune, J., Mouret, J. B., Lipson, H. (2013). The evolutionary origins of modularity. *Proc. R. Soc. B* 280:20122863. doi: 10.1098/rspb.2012.2863

Cohen, L., Dehaene, S., Naccache, L., Lehericy, S., Dehaene-Lambertz, G., Henaff, M., et al. (2000). The visual word-form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain* 123, 291–307. doi: 10.1093/brain/123.2.291

Cohen, L., Jobert, A., Le Bihan, D., and Dehaene, S. (2004). Distinct unimodal and multimodal

regions for word processing in the left temporal cortex. *Neuroimage* 23, 1256–1270. doi: 10.1016/j.neuroimage.2004.07.052

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204–256. doi: 10.1037/0033-295X.108.1.204

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cogn. Psychol.* 6, 84–107. doi: 10.1016/0010-0285(74)90005-X

Dahan, D., Magnuson, J. S., and Tanenhaus, M. K. (2001a). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cogn. Psychol.* 42, 317–367. doi: 10.1006/cogp.2001.0750

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., and Hogan, E. M. (2001b). Tracking the time course of subcategorical mismatches: evidence for lexical competition. *Lang. Cogn. Process.* 16, 507–534. doi: 10.1080/01690960143000074

Dandurand, F., Grainger, J., and Dufau, S. (2010). Learning location invariant orthographic representations for printed words. *Connect. Sci.* 22, 25–42. doi: 10.1080/09540090903085768

Dandurand, F., Hannagan, T., and Grainger, J. (2013). Computational models of location-invariant orthographic processing. *Connect. Sci.* 25, 1–26. doi: 10.1080/09540091.2013.801934

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychol. Rev.* 117, 713–758. doi: 10.1037/a0019738

de Kamps, M., and van der Velde, F. (2001). From artificial neural networks to spiking neuron populations and back again. *Neural Netw.* 14, 941–953. doi: 10.1016/S0893-6080(01)00068-5

Diependaele, K., Ziegler, J., and Grainger, J. (2010). Fast phonology and the bi-modal interactive activation model. *Eur. J. Cogn. Psychol.* 22, 764–778. doi: 10.1080/09541440902834782

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Frauenfelder, U. H., and Peeters, G. (1998). "Simulating the time course of spoken word recognition: an analysis of lexical competition in TRACE," in *Localist Connectionist Approaches to Human Cognition*, eds J. Grainger and A. M. Jacobs (Mahwah, NJ: Erlbaum), 101–146.

Gaillard, R., Naccache, L., Pinel, P., Clémenceau, S., Volle, E., Hasboun, D., et al. (2006). Direct intracranial, FMRI, and lesion evidence for the causal role of left inferotemporal cortex in reading. *Neuron* 50, 191–204. doi: 10.1016/j.neuron.2006.03.031

Gaskell, M. G., and Marslen-Wilson, W. D. (1997). Integrating form and meaning: a distributed model of speech perception. *Lang. Cogn. Process.* 12, 613–656. doi: 10.1080/016909697386646

Grainger, J., Diependaele, K., Spinelli, E., Ferrand, L., and Farioli, F. (2003). Masked repetition and phonological priming within and across modalities *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 1256–1269. doi: 10.1037/0278-7393.29.6.1256

Grainger, J., and Holcomb, P. J. (2009). Watching the word go by: on the time-course of component processes in visual word recognition. *Lang. Linguist. Compass* 3, 128–156. doi: 10.1111/j.1749-818X.2008.00121.x

Grainger, J., and Jacobs, A. M. (1996) Orthographic processing in visual word recognition: a multiple readout model. *Psychol. Rev.* 103, 518–565. doi: 10.1037/0033-295X.103.3.518

Grainger, J., and Ziegler, J. (2008). "Cross-code consistency effects in visual word recognition," in *Single-Word Reading: Biological and Behavioral Perspectives*, eds E. L. Grigorenko and A. Naples (Mahwah, NJ: Lawrence Erlbaum Associates), 129–157.

Grossberg, S., and Kazerounian, S. (2011). Laminar cortical dynamics of conscious speech perception: a neural model of phonemic restoration using subsequent context. *J. Acoust. Soc. Am.* 130, 440. doi: 10.1121/1.3589258

Grossberg, S., and Myers, C. W. (2000). The resonant dynamics of speech perception: interword integration and duration-dependent backward effects. *Psychol. Rev.* 107, 735–767. doi: 10.1037/0033-295X.107.4.735

Hannagan, T., Dandurand, F., and Grainger, J. (2011). Broken symmetries in a location invariant word recognition network. *Neural Comput.* 23, 251–283. doi: 10.1162/NECO_a_00064

Hannagan, T., and Grainger, J. (2012). Protein analysis meets visual word recognition: a case for String kernels in the brain. *Cogn. Sci.* 36, 575–606. doi: 10.1111/j.1551-6709.2012.01236.x

Hannagan, T., and Grainger, J. (in press). The lazy Visual Word Form Area: computational insights into location-sensitivity. *PLoS Comput. Biol.*

Hofmann, T., Schökopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *Ann. Stat.* 36, 1171–1220. doi: 10.1214/009053607000000677

Jones, M. N., and Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychol. Rev.* 114, 1–37. doi: 10.1037/0033-295X.114.1.1

Kim, A., Srinivas, B., and Trueswell, J. C. (2002). "The convergence of lexicalist perspectives in psycholinguistics and computational linguistics," in *Sentence Processing and the Lexicon: Formal, Computational and Experimental Perspectives*, eds P. Merlo and S. Stevenson (Philadelphia, PA: John Benjamins Publishing), 109–135.

Magnuson, J. S., Kukona, A., Braze, B., Johns. C. L., Van Dyke, J., Tabor, W., et al. (2011). "Phonological instability in young adult poor readers: time course measures and computational modeling," in *Dyslexia Across Languages: Orthography and the Brain-Gene-Behavior Link*, eds P. McCardle, B. Miller, J. R. Lee, and O. Tseng (Baltimore: Paul Brookes Publishing), 184–201.

Magnuson, J. S., Mirman, D., and Harris, H. D. (2012). "Computational models of spoken word recognition," in *The Cambridge Handbook of Psycholinguistics*, eds M. Spivey, K. McRae, and M. Joanisse (Cambridge: Cambridge University Press), 76–103. doi: 10.1017/CBO9781139029377.008

Magnuson, J. S., Strauss, T. J., and Harris, H. D. (2005). "Interaction in spoken word recognition models: feedback helps," in *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, eds B. G. Bara, L. W. Barsalou, and M. Bucciarelli, (Stresa), 1379–1394.

Magnuson, J. S., Tanenhaus, M. K., and Aslin, R. N. (2000). Simple recurrent networks and competition effects in spoken word recognition. *Univ. Rochester Work. Pap. Lang. Sci.* 1, 56–71.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., and Dahan, D. (2003). The time course of spoken word recognition and learning: studies with artificial lexicons. *J. Exp. Psychol. Gen.* 132, 202–227. doi: 10.1037/0096-3445.132.2.202

Marslen-Wilson, W. D., and Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition* 8, 1–71.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cogn. Psychol.* 21, 398–421. doi: 10.1016/0010-0285(89)90014-5

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cogn. Psychol.* 23, 1–44. doi: 10.1016/0010-0285(91)90002-6

McClelland, J. L., and Elman, J. L. (1986). The trace model of speech perception. *Cogn. Psychol.* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

McClelland, J. L., Mirman, D., and Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends Cogn. Sci.* 10, 363–369. doi: 10.1016/j.tics.2006.06.007

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: part 1. an account of basic findings. *Psychol. Rev.* 88, 375–407. doi: 10.1037/0033-295X.88.5.375

McQueen, J., Norris, D., and Cutler, A. (2006). Are there really interactive processes in speech perception? *Trends Cogn. Sci.* 10, 533. doi: 10.1016/j.tics.2006.10.004

Mirman, D., McClelland, J. L., and Holt, L. L. (2005). Computational and behavioral investigations of lexically induced delays in phoneme recognition. *J. Mem. Lang.* 52, 424–443. doi: 10.1016/j.jml.2005.01.006

Mirman, D., McClelland, J. L., and Holt, L. L. (2006a). Theoretical and empirical arguments support interactive processing. *Trends Cogn. Sci.* 10, 534. doi: 10.1016/j.tics.2006.10.003

Mirman, D., McClelland, J. L., and Holt, L. L. (2006b). Interactive activation and Hebbian learning produce lexically guided tuning of speech perception. *Psychon. Bull. Rev.* 13, 958–965. doi: 10.3758/BF03213909

Norris, D. (1990). "A dynamic-net model of human speech recognition," in *Cognitive Models of Speech Processing: Psycholinguistic and Computational Persepectives*, ed G. T. M. Altmann (Cambridge: MIT press), 87–104.

Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition* 52, 189–234. doi: 10.1016/0010-0277(94)90043-4

Norris, D., and McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychol. Rev.* 115, 357–395. doi: 10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., and Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behav. Brain Sci.* 23, 299–325. doi: 10.1017/S0140525X00003241

Perry, C., Ziegler, J. C., and Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychol. Rev.* 114, 273–315. doi: 10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J. C., and Zorzi, M. (2010). Beyond single syllables: large-scale modelling of reading aloud with the connectionist dual process (CDP++) model. *Cogn. Psychol.* 61, 106–151. doi: 10.1016/j.cogpsych.2010.04.001

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115. doi: 10.1037/0033-295X.103.1.56

Rauschecker, A. M., Bowen, R. F., Parvizi, J., and Wandell, B. A. (2012). Position sensitivity in the visual word form area. *Proc. Natl. Acad. Sci. U.S.A.* 109, 9244–9245. doi: 10.1073/pnas.1121304109

Rey, A., Dufau, S., Massol, S., and Grainger, J. (2009). Testing computational models of letter perception with item-level ERPs. *Cogn. Neurospsychol.* 26, 7–22. doi: 10.1080/09541440802176300

Salvata, C., Blumstein, S. E., and Myers, E. B. (2012). Speaker invariance for phonetic information: an fMRI investigation. *Lang. Cogn. Process.* 27, 210–230. doi: 10.1080/01690965.2011.594372

Seidenberg, M. S., and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychol. Rev.* 96, 523–568. doi: 10.1037/0033-295X.96.4.523

Shawe-Taylor, J. (1993). Symmetries and discriminability in feedforward network architectures. *IEEE Trans. Neural Netw.* 4, 816–826. doi: 10.1109/72.248459

Starr, M. S., and Rayner, K. (2001). Eye movements during reading: some current controversies. *Trends Cogn. Sci.* 5, 156–163. doi: 10.1016/S1364-6613(00)01619-3

Strauss, T. J., Harris, H. D., and Magnuson, J. S. (2007). jTRACE: a reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behav. Res. Methods* 39, 19–30. doi: 10.3758/BF03192840

Tabor, W., and Hutchins, S. (2004). Evidence for self-organized sentence processing: digging in effects. *J. Exp. Psychol. Learn. Mem. Cogn.* 30, 431–450. doi: 10.1037/0278-7393.30.2.431

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

Webber, C. J. S. (2000). Self-organization of symmetry networks: transformation invariance from the spontaneous symmetry-breaking mechanism. *Neural Comput.* 12, 565–596. doi: 10.1162/089976600300015718

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 30 April 2013; accepted: 08 August 2013; published online: 02 September 2013.*

*Citation: Hannagan T, Magnuson JS and Grainger J (2013) Spoken word recognition without a TRACE. Front. Psychol.* ***4****:563. doi: 10.3389/fpsyg.2013.00563*

*This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*

# APPENDIX

## A. PARAMETERS OF THE MODEL

| Name | Value | Description |
|---|---|---|
| Times | 10 | Number of time-specific slots (for input and time specific phonemes) |
| Istep | 10 | Pace of input stream (a new input is introduced every "istep" cycles) |
| Deadline | 100 | Deadline |
| DecayP | 0.01 | Decay rate for time-specific phonemes |
| DecayNP | 0.01 | Decay rate for time-invariant nphones |
| DecayW | 0.05 | Decay rate for time-invariant words |
| Gap | max | Authorized gap between phonemes in time-invariant nphones (e.g., if gap = 1, "/bark/" = "/ba/," "/ar/," "/rk/"; if gap = 2, "/bark/"= '/ba/," "/br/," "/ar/," "/ar/," "/ak/," "/rk/"). |
| PtoNPexc | 0.1 | Time-specific phoneme to time-invariant nphone excitation |
| PtoNPthr | 6 | Time-invariant nphone activation threshold |
| NPtoNPinh | 0 | Lateral inhibition between nphones |
| NPtoWexc | 0.05 | Excitation from time-invariant nphone ("/ba/") to words ("/bark/") |
| NPtoWscale | Wordlength | Scaling factor for NPtoW connections (here, set to word length) |
| WtoNPexc | 0 | Excitation from words ("/bark/") to time-invariant nphone ("/ba/") |
| 1PtoWexc | 0.01 | Excitation from 1-phone ("/a/") to words ("/bark/") |
| Wto1PExc | 0 | Excitation from words ("/bark/") to 1-phone ("/a/") |
| WtoWinh | −0.005 | Lateral inhibition between words |

## B. SIZING TRACE

Recall that TRACE duplicates each feature, phoneme, and word unit at multiple time slices. Features repeat every slice, while phonemes and words repeat every three slices. **Figure 2** illustrates reduplication and temporal extent of each unit type. For completeness we will include the feature level in our sizing of TRACE, although it will not be taken into account in our comparison with TISK. In the following, $S$, $F$, $P$, and $W$ will, respectively stand for the number of time slices, features, phonemes and words in the model.

### *B.1 Counting units*

Because there is a bank of $F$ features aligned with every slice, there are $SF$ feature units. For phonemes, given that we have $P$ time-specific units every three slices, for a total of $P(S/3)$. For words, we have $W$ time-specific units every three slices, for a total of $W(S/3)$.

The total number of units as a function of $S$, $F$, $P$, and $W$ can therefore be written: $SF + P(S/3) + W(S/3) = S(F + P/3 + W/3)$ We see that the cost in units is linear in all of these variables, and that for 201 time slices, 212 words, 14 phonemes, and 64 feature units the TRACE model requires 12,633 + 938 + 14,204 = 27,805 units.

### *B.2 Counting connections*

We start by counting the feature-phoneme connections. There are seven features per phoneme on average (vowels, fricatives and liquids don't use the burst parameter, but some phones take two values within a feature level). Let us count how many phoneme units overlap with each slice. From **Figure 2**, we can see that two copies of each phoneme overlap with each time slice. Therefore, there are seven (features) *2 (copies) $*P$ feature-phoneme connections per slice, which results in 14 PS feature-phoneme connections in the model.

Let us proceed to phoneme-word and word-phoneme connections. Words are four phonemes long on average, and there are $W(S/3)$ word units. But each of those units receives input not just from the four phonemes that are maximally aligned with it, but also the phonemes to the left and right of the maximally aligned phonemes. Thus, the total number of phoneme-word connections will be $3(S/3)Wp = SWp$, where $p$ is the number of phonemes per word. There will be an equal number of feedback connections from words to phonemes, for a total count of $4SW$ Phoneme-phoneme connections.

Next we consider the phoneme–phoneme connections. Each phoneme unit has an inhibitory link to each phoneme unit with which it overlaps. We can see from **Figure 2** that three copies of each phoneme overlap any given slice. So for each phoneme unit aligned at a given slice, it will have $3P - 1$ outgoing inhibitory links (we subtract 1 for the unit itself). We do not need to count incoming connections; these are included when we multiply by the number of phoneme units. This results in a total count of $PS(P - 1/3)$ word–word connections.

Just like phonemes, each word unit has an inhibitory link to each word unit with which it overlaps. The number of copies of any word that will overlap with a given slice will vary with word length, as can be seen in **Figure 2**. We can also see from **Figure 2** that words span six slices per phoneme. Recall that words are duplicated every three slices. For the 2- and 3-phoneme long examples in **Figure 2**, we can determine that the number of copies of each word of length $p$ that overlap with a given third slice (that is, an alignment slice, or a slice where one copy of the word is actually aligned) would be $1 + 2(2p - 1)$ (the first 1 is for the unit aligned at the slice), i.e., $4p - 1$. So a word unit at an alignment slice will have $(4p - 1)W - 1$ outgoing inhibitory connections. Therefore we arrive at a count of $W(S/3)((4p - 1)W - 1)$ word–word connections, which for an average word length of four phonemes amounts to $SW(5W - 1/3)$. All in all, we arrive at the following formula for the total connection count in TRACE: Total $= 14PS + 4SW + PS(P - 1/3) + SW(5W - 1/3) = S(14P + W + P(P - 1/3) + W(5W - 1/3)) = S(P(P + 41/3) + W(5W + 2/3)) = S[(P^2 + 41/3P) + (5W^2 + 2/3W)]$.

$$\begin{aligned} c_{\text{TRACE}} &= 14PS + 4SW + PS(P - 1/3) + SW(5W - 1/3) \\ &= S(14P + W + P(P - 1/3) + W(5W - 1/3)) \end{aligned}$$

$$= S(P(P + 41/3) + W(5W + 2/3))$$
$$= S[(P^2 + 41/3P) + (5W^2 + 2/3W)] \quad (3)$$

According to our calculations, the cost in connections is therefore a quadratic function of $P$ and $W$ (due to lateral inhibition at the phoneme and word levels), and a linear function of $S$ (due to limited overlap of units over time slices). In particular, with the standard parameters of 212 words, 14 phonemes, a mean word length of 4 phonemes, and 67 alignment units the TRACE model requires 45,573,266 connections.

### C. SIZING TISK

TISK has three levels: a time specific phoneme level, a time-invariant string kernel level (tisk, after which the model is named), and a time-invariant word level. TISK doesn't have a feature level, and instead the output of such a level is emulated by a wave of net inputs that arrives to the time-specific phoneme level at a regular pace. A feedforward symmetry network operates the transition between the time-specific phoneme level and the nphone level. There are positive feedforward and feedback connections between the nphone and word levels, and lateral inhibitory connections within them, although in practice only the word level has non-zero inhibitory connections and they are restricted to neighbors. The cuts in computational resources are mostly due to the symmetry network, and to a lesser extent, to the limited use of lateral inhibition.

#### *C.1 TISK units*

Because only one level is time-specific in TISK, the notion of alignment doesn't have course anymore. Therefore the number of time-specific phonemes is simply given by the number of phonemes multiplied by the number of time slices, or $PS$. With 14 phonemes and, 201 slices, this amounts to 2814 time-specific phoneme units. The nphone level hosts time-invariant phonemes and all possible diphones (even phonotactically illegal ones), and therefore uses $P + P^2$ units, which for $P = 14$ means 210 units. Finally the word level counts $W$ units, one for each word in the lexicon, and $W$ is set to 212 throughout most simulations. The total number of units in the model is therefore $PS + P + P^2 + W = P(P + S + 1) + W = 3236$ units. $W$ time-invariant word units (212). $P + P^2$ time-invariant n-phone units ($P$ 1-phones and $P^2$ diphones; = 210). Total units at basic parameters: 1360.

#### *C.2 TISK connections*

We only count non-zero connections throughout. We start by sizing connections in the symmetry network (**Figure 3**). A time-specific phoneme unit sends a connection to an nphone unit if and only if it is a constituent of this unit (for instance, $A_2$ sends a connection to A, AB, BA, and AA, but not to B). There are $2P - 1$ diphones that start or end with a given phoneme, and one time-invariant phoneme, so a given phoneme at time t will send $2P - 1 + 1 = 2P$ connections, and multiplying this by the number of time specific phonemes $PS$, we see that the total number of connections is $2P^2S$. From this, however, we must remove all zero connections: unit $A_1$ (resp. $A_T$) should not give evidence for diphone units that end with A (resp. that start with A), and therefore gradient coding assigns zero to these connections. We see that these cases only occur at the first and last time slices (implying that there are more than two time slices), and that for a given phoneme, $P - 1$ connections are concerned, resulting in $2P(P - 1)$ zero connections. There are therefore $2P^2S - 2P(P - 1)$, or $2P(SP - P + 1)$, phoneme-to-nphone connections in the symmetry network (with 14 phonemes and 201 time slices, this amounts to 78,428 connections).

We must now count the number of gating connections in the symmetry network. To prevent spurious activations at the nphone level, the symmetry network uses gating connections. These are hard-wired connections that originate from time specific phonemes, and inhibit some connections between time-specific phonemes and time-invariant nphones. Specifically, a given phoneme at a given time slice will inhibit all connections at later time slices that originate from the same phoneme and arrive to a diphone that begins with that phoneme (and does not repeat). Because there are $P - 1$ diphones that start with a given phoneme and do not repeat, and there are $P$ phonemes at a given time slice, $P(P - 1)$ connections must be gated at any time slice after the one considered, or for $S > 2$:

$$\begin{aligned} c_{\text{gating}} &= P(P-1)(S-1) + P(P-1)(S-2) + \ldots \\ &\quad + P(P-1)(1) \\ &= P(P-1)\sum_{s=1}^{S-1} s \\ &= \frac{P(P-1)S(S-1)}{2} \end{aligned} \quad (4)$$

With 14 phonemes and 201 time slices, this amounts to 3,658,200 gating units. The total in the time specific part of the network is therefore of 3,658,200 + 78,428 = 3,736,628 connections (Note that the formulas obtained here were verified empirically by direct inspection of the number of connections in the model for various number of time slices, and were found to be exact in all cases). We now proceed to count connections in the time invariant part of the network, first noticing that because lateral inhibition at the nphone level was set to zero, we only need to count the connections between the nphone and the word level, as well as the lateral connections within the word level. However, in TISK these numbers will depend not only on the size of the lexicon and the number of nphones, but critically also on the distribution of nphones in the particular lexicon being used, so that we are reduced to statistical approximations. Empirically, we find that an average word connects to 9.5 nphones in TISK, leading to an estimate of 9.5 W feedforward connections between the nphone and word level. Similarly, simulations show that the number of lateral inhibitory connections at the word level in TISK is $0.8W(W - 1)$. Therefore the number of connections in the time-invariant part of the model reaches $0.8W^2 - 0.8W + 9.6W = 0.8W^2 + 8.8W$. With a lexicon of 212 words, this amounts to 37,800 connections.

All in all, we arrive at the following expression for the number of connections in TISK for $S > 2$:

$$c_{\text{TISK}} = 2P^2S - 2P(P-1) + \frac{P(P-1)S(S-1)}{2} + W(0.8W + 8.8) \quad (5)$$

which amounts to 3,774,428 connections using our usual assumptions on $S$, $P$, and $W$. It can be seen when this expression is developed that it is quadratic in $S$, $P$, and $W$. This would seem to be a setback compared to the expression obtained for TRACE, which is only quadratic in $P$ and $W$ but linear in $S$. However, $S$ is *orders* of magnitudes smaller than $W$, and what we obtain in exchange of this quadratic dependence to $S$ is to decouple the $S$ and $W$ factors, reflecting the fact that in TISK the lexicon is not duplicated for every time slice anymore. Consequently there is a substantial gain in connections when switching from TRACE (45,573,266) to TISK (3,774,105) connections, the latter having ten times less connections, a gain of one order of magnitude which improves with lexicon size to reach an asymptota at three orders of magnitude.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 19 September 2013
doi: 10.3389/fpsyg.2013.00635

# Deep generative learning of location-invariant visual word recognition

***Maria Grazia Di Bono*** [1] ***and Marco Zorzi*** [1,2] ***

[1] *Computational Cognitive Neuroscience Lab, Department of General Psychology, University of Padova, Padova, Italy*
[2] *IRCCS San Camillo Neurorehabilitation Hospital, Venice-Lido, Italy*

***Edited by:***
*Pablo Gomez, De Paul University, USA*

***Reviewed by:***
*Colin Davis, Royal Holloway University of London, UK*
*Stéphanie Massol, Basque Center on Cognition, Brain and Language, Spain*

***Correspondence:**
*Marco Zorzi, Computational Cognitive Neuroscience Lab, Department of General Psychology, University of Padova, Via Venezia 12, Padova 35131, Italy*
*e-mail: marco.zorzi@unipd.it*

It is widely believed that orthographic processing implies an approximate, flexible coding of letter position, as shown by relative-position and transposition priming effects in visual word recognition. These findings have inspired alternative proposals about the representation of letter position, ranging from noisy coding across the ordinal positions to relative position coding based on open bigrams. This debate can be cast within the broader problem of learning location-invariant representations of written words, that is, a coding scheme abstracting the identity and position of letters (and combinations of letters) from their eye-centered (i.e., retinal) locations. We asked whether location-invariance would emerge from deep unsupervised learning on letter strings and what type of intermediate coding would emerge in the resulting hierarchical generative model. We trained a deep network with three hidden layers on an artificial dataset of letter strings presented at five possible retinal locations. Though word-level information (i.e., word identity) was never provided to the network during training, linear decoding from the activity of the deepest hidden layer yielded near-perfect accuracy in location-invariant word recognition. Conversely, decoding from lower layers yielded a large number of transposition errors. Analyses of emergent internal representations showed that word selectivity and location invariance increased as a function of layer depth. Word-tuning and location-invariance were found at the level of single neurons, but there was no evidence for bigram coding. Finally, the distributed internal representation of words at the deepest layer showed higher similarity to the representation elicited by the two exterior letters than by other combinations of two contiguous letters, in agreement with the hypothesis that word edges have special status. These results reveal that the efficient coding of written words—which was the model's learning objective—is largely based on letter-level information.

**Keywords: orthographic coding, open-bigrams, connectionist modeling, hierarchical generative models, deep unsupervised learning**

## INTRODUCTION

Visual word recognition and reading aloud is one of the cognitive domains where connectionist modeling has achieved its greatest success. Following seminal studies published in the 1980s (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982; Seidenberg and McClelland, 1989), recent modeling work has produced highly detailed simulations of skilled reading, reading development, and dyslexia (e.g., Plaut et al., 1996; Zorzi et al., 1998; Harm and Seidenberg, 1999; Coltheart et al., 2001; Perry et al., 2007, 2010, 2013; Zorzi, 2010; Ziegler et al., in press; see Zorzi, 2005, for a review). Nevertheless, despite an impressive upscaling of connectionist models of reading in recent years (e.g., Perry et al., 2010, 2013), most of these models remain largely underspecified with regard to the "visual front-end" of the reading system. That is, most models stipulate that the identity and position of individual letters is coded in a way that is abstracted from the retinal input both in terms of shape and spatial location with respect to eye fixation. In particular, the latter assumption implies a location-invariant word-centered representation, with letters aligned according to a fixed template (e.g., left-justified slot-based coding). The issue of how location-invariance might be computed from the native retinotopic (eye-centered) code has recently attracted much interest (Dehaene et al., 2005; Dandurand et al., 2010; Hannagan et al., 2011), because it is closely tied to a lively debate on the nature of orthographic coding and more specifically on the coding of letter position during visual word recognition (e.g., Whitney, 2001; Grainger and van Heuven, 2003; Davis and Bowers, 2006; Gomez et al., 2008; Davis, 2010; Grainger and Ziegler, 2011).

The theoretical debate on letter position coding was triggered by studies that reported relative-position and transposition priming effects in visual word recognition using the masked priming paradigm (e.g., Humphreys et al., 1990; Peressotti and Grainger, 1999; Perea and Lupker, 2003; Schoonbaert and Grainger, 2004; Grainger et al., 2006). The first phenomenon refers to the finding that word recognition is facilitated when primes are composed of a subset of letters constituting the target word, but only when relative positions are respected. Transposition priming, instead,

refers to the finding that when primes share all the constituent letters of the target words, priming still persists when small changes in letter order is performed (e.g., transposing two adjacent letters). It is widely believed that these priming effects stem from a level of orthographic processing where some form of approximate, flexible coding of letter positions operates (Grainger, 2008) and have inspired alternative models of letter position coding (Grainger and Whitney, 2004; Gomez et al., 2008; Davis, 2010). All models share the assumption that visual word recognition is built upon parallel processing of the constituent letters, in contrast to an holistic word-shape representation (see Pelli et al., 2003; Grainger, 2008; Grainger et al., 2012). From the computational point of view, holistic word-shape coding is extremely costly because it requires to solve shape invariance for each word rather than for each letter of the alphabet. However, the models differ in terms of how approximate letter position coding is achieved. For example, the Overlap model of Gomez et al. (2008) assumes a noisy coding of letter position within the classic slot-based coding scheme used in the interactive-activation model (IAM) of McClelland and Rumelhart (1981). In the IAM model, words are processed in parallel from a set of letter detectors that are length-dependent and position-specific. Uncertainty about letter positions is implemented in the Overlap model as a Gaussian distribution of activation across the ordinal positions in the word. Letter position uncertainty is also a central feature of Davis' (2010) spatial coding model.

A different theoretical perspective is that orthographic coding is based on combinations of contiguous and non-contiguous ordered letter pairs, in a way to code relative rather than absolute letter positions (Whitney, 2001; Grainger and van Heuven, 2003). For instance, the word WITH would be coded with the set of bigrams [WI, WT, WH, IT, IH, TH], a scheme known as Open-bigram coding (Grainger and Whitney, 2004). Open-bigrams are an intermediate coding between the representation of single letters and whole-words. Grainger and van Heuven (2003) propose the existence of a bank of letter detectors performing parallel letter identification, independently from the physical characteristics of the letters (i.e., shape and size) but not from the spatial location. Therefore, the activity of letter detectors is an abstract representation of letters conveying information about letter identity at a specific locations. In the next stage, a more abstract "relative position map" is formed, coding for the relative position of letter identities within the word, independently from their shape and size, and independently from the spatial location of the word (i.e., location invariance). According to the open-bigram model, this is possible through a bank of open-bigram units, receiving the input from the letter detectors: the open bigram for a specific ordered letter pair (e.g., A_C) is activated by all the possible location combinations in the letter detectors for the given letter order. Open-bigrams then send their activations to all compatible word representations. In this way a flexible relative-position code mediates the processing of reading words as a whole.

The idea that visual word recognition might involve a number of intermediate and progressively more abstract levels of orthographic coding is the key aspect of Dehaene et al.'s (2005) local combination detector (LCD) model. Though not implemented as a computer simulation, the LCD model is inspired by the neurophysiology of the primate visual object recognition system. Specifically, object recognition is based on hierarchical processing of basic local features that are gradually integrated into more complex and abstract features (through increasing size of receptive fields) to progressively achieve invariance for size, shape, and location (see Riesenhuber and Poggio, 1999, for a computational model). Given that reading is a recent cultural invention, it is unlikely that the brain contains a specific neural mechanism for visual word recognition. Thus, learning to recognize printed words independently from their location, font, size, etc. might be achieved by recycling the cortical machinery for object recognition (Dehaene and Cohen, 2011). According to the LCD model, part of the occipito-temporal "what" pathway is organized into a hierarchy of neuronal levels, each composed of local combination detectors that are gradually sensitive, through increasing complexity and size of their receptive fields, to larger fragments of words. Besides the well-known finding that the occipito-temporal cortex of skilled readers contains a "visual word form area" (Cohen et al., 2002; Cohen and Dehaene, 2004), recent functional neuroimaging studies support the LCD model by showing that perception of written words involves the sensitivity to increasingly larger visual units along a posterior-to-anterior gradient in the ventral visual stream (Vinckier et al., 2007). Notably, open bigrams are important intermediate-size units in the LCD model.

The problem of learning a location-invariant orthographic representation of printed words was recently tackled by Dandurand et al. (2010) with connectionist simulations. They used error backpropagation to train a feedforward neural network with one layer of hidden units on the mapping from location-specific letter identities to location-invariant localist word representations. The phenomena of transposed-letter and relative-position priming were investigated in the network by presenting stimuli obtained by transposing two letters or removing one letter from a trained target word. The transposed letter stimuli, compared to control stimuli in which the two letters were replaced by non-constituent ones, produced an activation pattern that was more similar to that produced by the target word. In the same vein, stronger similarity to the target word activation was obtained when the input stimuli maintained the letter order (e.g., ABC for ABCD) with respect to controls in which the letter order was reversed (e.g., CBA for ABCD). Moreover, when the order was maintained, stimuli composed of non-contiguous letters yielded a stronger similarity to the target word in comparison to stimuli containing only the contiguous letters (e.g., ABD vs. ABC for ABCD). These findings suggested that the network had learned a code for contiguous and non-contiguous letter combinations. Hannagan et al. (2011) further investigated the neural network model of Dandurand et al. (2010) by analyzing its hidden layer activity. They found that no knowledge about bigrams was learned by the network. Instead, the network learned letter identities almost independently from their locations (in a "semi-location-invariant" way). This information allowed to compute constituent bigrams and words without the explicit coding of letter combinations. These results are in line with the overlap model of Gomez et al. (2008).

While the connectionist studies of Dandurand et al. (2010) and Hannagan et al. (2011) represent a first important attempt

to understand how a location-specific letter-based code could be mapped onto location invariant word representations, the plausibility of the model is hindered by its network architecture and by the use of supervised learning by error backpropagation. Besides the well-known lack of biological plausibility of the backpropagation algorithm (O'Reilly, 1998), the supervised learning regimen is problematic because it implies that orthographic learning requires an external, explicit teaching signal at each word encounter. Moreover, the classic feedforward network with one layer of hidden units used by Dandurand and colleagues does not capture the hierarchical organization of the visual system, which is a key feature for achieving invariant object recognition in biologically inspired computational models of vision (Riesenhuber and Poggio, 1999).

In this article we present a connectionist model of location-invariant visual word recognition that can be cast within the broader theoretical framework of Dehaene et al.'s (2005) LCD model. The assumption that orthographic learning exploits the cortical machinery for object recognition leads to the prediction that perceptual invariance for visual words might emerge from unsupervised generative learning in a neural network with a hierarchical architecture, that is a "Deep Belief Network" (DBN; Hinton, 2007; Stoianov and Zorzi, 2012; Zorzi et al., 2013). DBNs are stochastic recurrent neural networks with many layers of hidden units that encode increasingly more complex features of the sensory input across layers (Hinton and Salakhutdinov, 2006; Hinton, 2007, 2013). In practice, a DBN is a stack of Restricted Boltzmann Machines (RBMs; Hinton, 2002) trained in a layer-wise fashion. RBMs are stochastic networks with one layer of visible neurons encoding the input patterns and one layer of hidden neurons connected through bidirectional symmetric links. Learning in RBMs is unsupervised and its objective is to build internal representations of the sensory input by fitting a generative model to the data. Therefore, after training all RBM layers in succession, the DBN is a hierarchical generative model in which the latent causes of the data are represented through distributed non-linear representations across hidden layers (HLs). DBNs represent the state-of-the-art in machine learning but they are also particularly appealing for connectionist modeling of cognition because they learn multiple levels of representation without any supervision or reward and they have a sound probabilistic formulation (see Zorzi et al., 2013, for a tutorial review). Crucially, the analyses of the internal representations can reveal an emergent coding strategy that closely mirrors single-cell recording data (e.g., Lee et al., 2008; De Filippo De Grazia et al., 2012a; Stoianov and Zorzi, 2012). In the present work we trained a DBN on an artificial dataset of letter strings presented at five possible retinal locations. We asked whether location-invariant word recognition would emerge from unsupervised deep learning and what type of intermediate coding would support the transition between location-specific (i.e., eye-centered) letter coding and location-invariant word representations.

## MATERIALS AND METHODS

We employed a DBN with three HLs for learning a generative model of written words. In the following subsections, we describe the training dataset and the network architecture. The code (Matlab/Octave) used for the simulation and the training set is available at http://ccnl.psy.unipd.it/deeplearning.

### TRAINING DATASET

We used an artificial dataset constructed *ad hoc* in order to investigate orthographic learning in a restricted but tightly controlled way (also see Dandurand et al., 2010). The dataset was composed of 120 3-letter words presented at 5 different eye-centered locations (one central and two locations on each side of the central one), for a total of 600 (120 words × 5 locations) input patterns. An artificial lexicon was generated by considering all simple permutations of three letters without repetitions from a partial alphabet composed of six letters (i.e., A B C D E F). In this way, it was possible to balance the frequency of each letter in the lexicon and to avoid letter repetition. Indeed, including letter repetition within the same word could introduce a possible confound in identifying the contribution of open bigrams to the orthographic coding for visual word recognition.

### INPUT CODING

We used a sparse coding (i.e., slot coding) for representing the training words (see also Dandurand et al., 2010). Input words were coded by the pattern of activity over 7 location-specific (i.e., eye-centered) letter slots (see **Figure 1**) and each word could be coded at 5 different locations. Each letter within a word was coded by the activation of a specific letter unit (over a set of 26 units, one for each letter of the alphabet[1]), at a specific eye-centered location.

[1]Note that using a more compact representation with only six letter units (for letters A–F) at each slot does not change the results presented here. Conversely, the use of a full set of letter units allows to readily extend the model's training set.

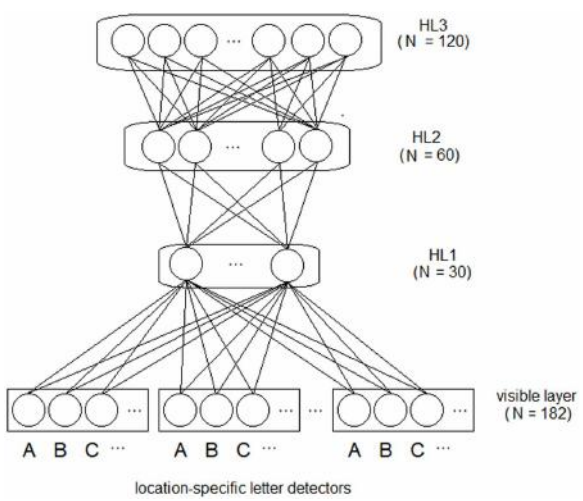



**FIGURE 1 | Architecture of the deep network model.** Letter strings are presented to the visible (i.e., input) layer using a bank of location-specific letter detectors within slots encoding 7 different spatial locations. Activity of the visible layer is fed to three hierarchically organized layers of feature detectors (hidden neurons). All connections are bidirectional and symmetric. Note that word-level information is not explicitly represented in the network and it is not supplied during training.

Blank locations were coded using zeros for all units of a slot. Thus, the input pattern was a vector of 182 binary values.

### NETWORK ARCHITECTURE

The deep network had one visible layer encoding the input data and three hierarchically organized HLs (see **Figure 1**). A characteristic of deep networks is that adding HLs generally increases the complexity of the features that can be detected during learning. There is a point, however, where adding further layers does not improve global performance (see Hinton, 2012 for a practical guide). We measured the global performance of the network by linearly decoding training words from the activity of the deepest layer (see Zorzi et al., 2013, for discussion). Thus, we empirically determined the number of layers, starting from a first hidden layer of 30 neurons (approximately corresponding to the square root of the total number of training patterns), then we increased the number of layers, doubling the number of hidden neurons with respect to the previous layer (i.e., 60 and 120 neurons for layer 2 and 3, respectively). Performance did not improve with more than three HLs.

Learning proceeded layer-wise (i.e., one layer at a time) for computational efficiency. For the first hidden layer, the input was the activity of the visible layer. For the other layers, the input was the activity of the previously trained layer. Each RBM (one for each layer), was trained with the Contrastive Divergence (CD) learning algorithm (Hinton and Salakhutdinov, 2006) to learn a generative model of the input data without supervision (i.e., maximizing the likelihood of reconstructing the data). Crucially, no word level information was provided to the network. For each RBM, learning involves two phases: a "positive" and a "negative" phase. During the positive phase the visible units are clamped to the data pattern and their activity ($v_i^+$) spreads to the hidden neurons ($h_j^+$). In the negative phase, a stochastically selected binary state of the hidden neurons (using their state $h_j^+$ as probability to turn them on) feeds back to the visible units ($v_i^-$) through the top-down weights (i.e., reconstruction of the input vector) and then feeds forward again to the hidden neurons ($h_j^-$) (see Zorzi et al., 2013, for a more detailed discussion). The weights $w_{ij}$ are updated with a small learning fraction ($\eta$) of the difference between pairwise correlations measured in the positive and negative phases:

$$\Delta W = \eta(v^+h^+ - v^-h^-)$$

We trained the deep network for a maximum number of 1000 epochs, using a learning rate of 0.1, and an increasing momentum ranging between 0.5 and 0.9. To ensure robustness of the results, we trained 10 versions of the same network using different initial random weights.

Unsupervised deep learning was carried out on a multi-core high-performance cluster using an Octave/Open-MPI parallelization (De Filippo De Grazia et al., 2012b; Testolin et al., 2013). Note that Testolin et al. (2013) provide code for a variety of parallel solutions and show that learning time can be further reduced by exploiting the GPUs of low-cost graphic cards on a desktop PC.

## RESULTS

### DECODING FROM ACTIVITY OF HIDDEN LAYERS

After training, we investigated the quality of the representation generated within each HL. We used a linear classifier for decoding the input words from each of the three HLs; the classifier weights were computed using the pseudo-inverse method (Hertz et al., 1991), which is equivalent to using the delta rule but more efficient and parameter-free (see Zorzi et al., 2013). Only at this level of analysis we introduced word-level information for learning a linear association between the activity of each hidden layer, computed presenting an input word on the visible layer, and the same word used as target. Each target word was coded into a binary output unit, independently from the location at which it was presented. The presence of a corresponding word was marked by a value of 1, its absence by a value of 0. There were 120 output units, each corresponding to a training word, independently from its location. For instance, with 4 target words (e.g., ABC, ABD, ABE, ABF) the input word ##ABC## (as well as ####ABC) would be coded as 1 0 0 0, whereas the word ###ABE# (as well as ABE####) would be coded as 0 0 1 0. Recognition performance was expressed in terms of the percentage of input words correctly decoded, independently from the location.

We hypothesized that decoding accuracy would increase across layers, thereby indexing that internal representations become more abstract with the increasing of the network's depth. The percentage of correctly decoded words is shown in panel **(A)** of **Figure 2** as a function of the layer used as input to the classifier. Indeed, decoding accuracy significantly increased with layer depth [$F_{(1.19,\ 10.73)} = 1872.01$, $p < 0.0001$, $\eta_p^2 = 0.99$, Greenhouse-Geisser corrected for sphericity] and it reached near-perfect accuracy ($M = 99.43 \pm 0.14$ s.e.m.) at the deepest hidden layer (HL3). Panel **(B)** of **Figure 2** shows decoding accuracy as a function of location of the input words. Notably, location-invariance increased as a function of layer depth: that is, decoding accuracy in HL1 and HL2 varied among locations (with a tendency for higher accuracy at the two outer locations), whereas accuracy in HL3 was near-perfect across all locations.

The distribution of decoding errors can provide insights about how orthographic information is encoded within the different layers of the deep network. We therefore analyzed the decoding error distribution as a function of the orthographic distance between the input word and the incorrectly decoded word, indexed by the Levenshtein Distance (LD) (Yarkoni et al., 2008). For example, the word ABC has a distance of 1 from the words *BC, A*C, AB* (where the symbol * means a letter that does not belong to it), a distance of 2 from the words with transposed letters (i.e., ACB, BAC, CBA, BCA, and CAB) as well as from the words A**, *B*, **C, and a distance of 3 from all words containing letters that do not belong to it (e.g., DEF, EFD). Note that the LD measure was computed independently from the location of the input word. The error distribution is shown in panel **(C)** of **Figure 2** as function of LD and layer depth. Note that the majority of errors consisted in producing words at a distance of 2. The finding that a large proportion of decoding errors do not involve words at the closest orthographic distance ($LD = 1$) but are concentrated on a distance of 2 suggests that most errors might be in fact transposition errors. Splitting the error distribution for

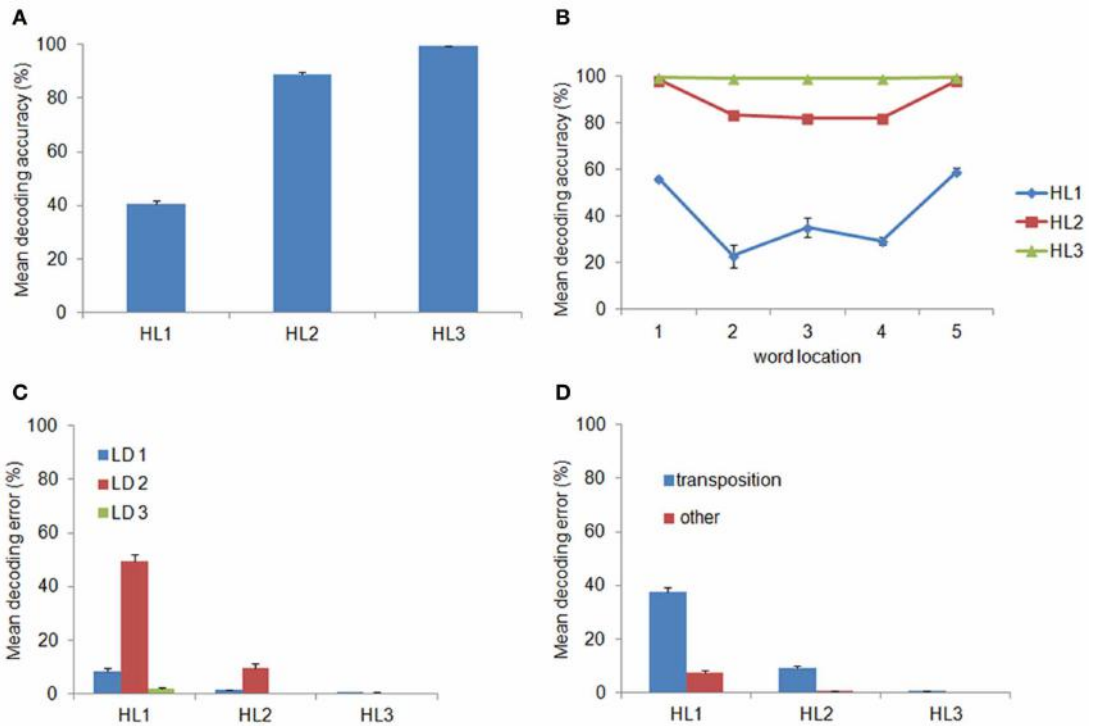



**FIGURE 2 | Word decoding from the Hidden Layers (HLs). (A)** Mean decoding accuracy, expressed in terms of the percentage of correctly recognized words, as a function of layer depth. Decoding accuracy significantly increased across layers and was near-perfect at the deepest layer (i.e., HL3). **(B)** Decoding accuracy as a function of word location. Decoding accuracy in HL1 and HL2 varied among locations, with higher accuracy at the two outer locations, whereas accuracy in HL3 was near-perfect across all locations. **(C)** Decoding error as a function of the Levenshtein Distance (LD) from the correct word and layer depth. Most of the decoding errors were not close neighbors but were at a distance of 2, which include transposition errors. **(D)** Error distribution for $LD = 2$, after splitting it between transposition and other errors, as a function of layer depth. Transposition errors were predominant and they accounted for virtually all errors in HL2 and HL3. Error bars in all graphs indicate standard error across ten simulations using networks with different initial random weights.

$LD = 2$ between words with transposed letters and other words showed that this was indeed the case (see **Figure 2D**). Finally, we also assessed whether the distribution of transposition errors varied across locations. The results are shown in **Figure 3**. For HL1 and HL2, transposition errors were mainly and almost similarly distributed across the three inner locations. This result is complementary to the distribution of decoding accuracy across locations (see **Figure 2B**), which was higher at the two outer locations. This advantage can be readily explained by the fact that training implies less position uncertainty for letters in the outer slots. For example, during training of the word ABC the only letter presented in slot 1 is A, whereas slot 3 can contain the letters A, B, or C. Thus, letter A in the leftmost (or rightmost) slot provides unambiguous evidence for words starting (or ending) with A, whereas letter A in slot 3 may belong to any word that contains A.

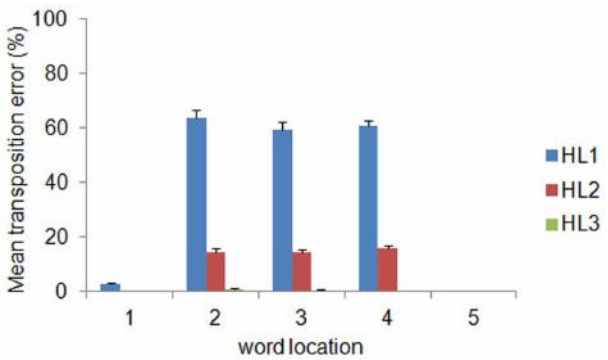



**FIGURE 3 | Mean transposition errors as a function of layer depth and word location.** For HL1 and HL2, transposition errors were mainly and almost similarly distributed across the three inner locations. Error bars indicate standard error across ten simulations.

## ANALYSIS ON SINGLE NEURONS

To further characterize the information encoded into the trained network, we analyzed the activity of each neuron within each HL. This analysis was performed on a single network (i.e., the first network trained). Borrowing the classic method used in single-cell recording studies (also see Stoianov and Zorzi, 2012; Zorzi et al., 2013), we sought to establish whether and how the activity of each neuron is modulated by two key factors: (i) word selectivity; (ii) location invariance. Finally, a further analysis on single neurons allowed us to assess whether knowledge about bigrams had emerged in the network.

### WORD SELECTIVITY

We fixed the preferred location of each neuron, choosing the location that maximized its activity across all trained words. We then fixed its preferred word, on the basis of its maximum activity at the preferred location. Finally, we performed a linear regression on its normalized activity in response to the training words (presented at the preferred location) using LD as predictor (3 levels: 0, 1, and 2; note that $LD = 0$ indexes the preferred word). We discarded the words at an orthographic distance $LD = 3$ from the input word, which are those words composed of all letters that did not belong to the input word. After False Discovery Rate (FDR) correction for multiple comparisons, we selected all the neurons for which the regression was significant. No word selectivity was found in HL1. In contrast, word selectivity emerged in 95% of HL2 neurons (FDR $p = 0.037$) and in 97.5% of HL3 neurons (FDR $p = 0.037$). Activity of these neurons was modulated in a monotonically decreasing way by the orthographic distance of the input words from the preferred word.

### WORD LOCATION INVARIANCE

We fixed the preferred word of each neuron, choosing the training word that maximized its activity. Then, we used a pattern matching procedure for assessing the degree of invariance to the spatial location of the preferred word. In particular, we defined a set of binary location vectors, each encoding the preference for one or more specific locations (e.g., 0 0 1 0 0, coding the preference for the central location; 1 1 1 1 1, coding an equal preference for all the available locations). For instance, for a neuron with a preferred word ABC, we collected its activity as a function of the location at which the input word ABC was presented. Then, we selected the more similar location vector using the Euclidean Distance as similarity index. This procedure revealed the number of locations for which the neuron activity was highly similar, that is, the number of neuron's preferred locations. A single preferred location indexes location-specific word coding, whereas 5 preferred locations (i.e., equal preference across locations) indexes location-invariant word coding. **Figure 4** shows the distribution of neurons as a function of the number of preferred locations and the word selectivity index (β coefficient of the LD predictor). Notably, location invariance emerged only in the deepest layer (30.83% of HL3 neurons), where word recognition was also near-perfect.

### BIGRAM CODING

To assess whether tuning to bigrams had emerged in the network, we presented all possible sub-word units (i.e., letters and bigrams) to the network and recorded the activation of each neuron within each layer. Letters were presented at all the 7 possible locations, whereas bigrams were divided between contiguous (e.g., AB for ABC) and non-contiguous (e.g., A_C for ABC) and were presented at the 6 and 5 possible locations, respectively. The activity of each neuron across sub-word units was normalized, so that it had a maximum of 1 for its preferred stimulus. We then determined the preferred bigram for each neuron, choosing the bigram that maximized its activity, and performed three diagnostic tests. First, we assessed whether the neuron's responses to both constituent letters were smaller than the response to the bigram by at least 10% (i.e., the neuron's response to AB should be larger than the response to A and to B presented in isolation). Note that we chose a lenient criterion because assuming additivity of the response to the constituent letters (i.e., response to AB as sum of the responses to A and B) is unwarranted for non-linear neurons. Second, we assessed whether it was maximally active for all the words containing the preferred bigram, in order to exclude that the neuron was tuned to specific words. Finally, the candidate bigram neuron was assessed for its sensitivity to letter order. Indeed, a neuron might respond to the co-occurrence of two letters (e.g., A and B), but to be qualified as bigram detector, its response to the transposed letter pair should be smaller (i.e., response should be stronger for AB than for BA). Thus, we computed an index of order sensitivity as the difference between the response to the preferred bigram and to its transposed version presented at the same location. Values close to

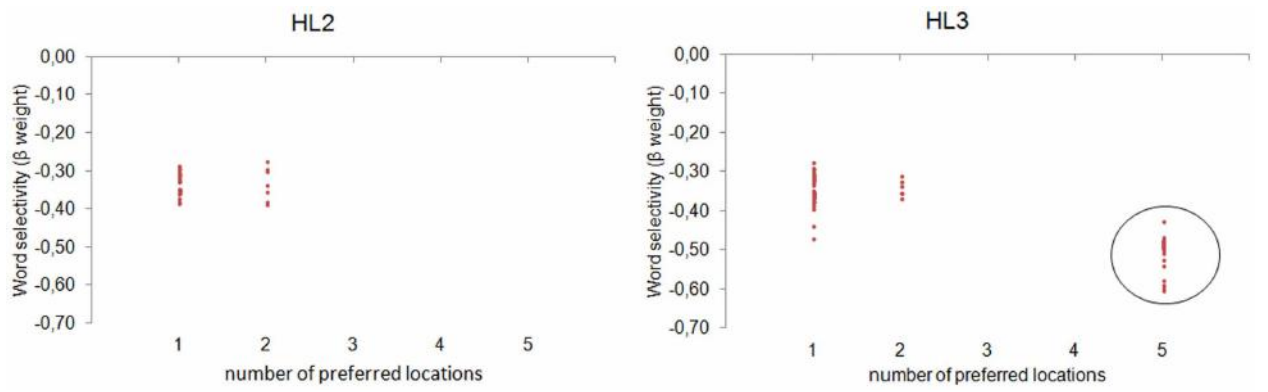



**FIGURE 4 | Analysis on single neurons.** The distribution of neurons in HL2 (left panel) and HL3 (right panel) is shown as a function of the number of preferred locations and the word selectivity index (regression weight of the LD predictor). Only neurons showing a significant modulation of activity in the regression analysis are plotted. Location invariance emerged only in 30.83% of HL3 neurons (circled in black), where word recognition was also near-perfect.

zero would index lack of order sensitivity, whereas values close to one would show that the neuron does not respond at all to the bigram with the opposite letter order. A neuron passing the first two tests and showing high sensitivity to order would be classified as bigram detector, thereby providing evidence that sensitivity to bigrams has emerged as intermediate coding strategy in the network. This analysis showed that there were no neurons, across the three layers, that could be classified as bigram detectors—indeed, no neuron passed the first two tests.

### ANALYSIS ON ACTIVATION PATTERNS

The contribution of sub-word orthographic units to the representation of words can also be assessed at the level of distributed representations over the hidden neurons of the deepest layer (where word recognition is near-perfect). We therefore, analyzed the similarity between activation patterns produced by training words and those produced by the different types of sub-word units. This analysis was performed on the same network selected for the single neuron analysis (i.e., the first network trained). More specifically, we presented letters and bigrams to the trained network and recorded the pattern of activation of the deepest layer (HL3). Letters and bigrams were presented at all the possible input locations; bigrams were divided between contiguous (e.g., AB and BC for ABC) and non-contiguous (e.g., A_C for ABC). Moreover, letters and bigrams (both contiguous and non-contiguous) could be constituent (e.g., A, B, C, AC, etc. for ABC) or non-constituent (e.g., D, E, F, DE, etc. for ABC). We computed the cosine distance [2] between the activation pattern produced by each word presented at a randomly selected location and those produced by open bigrams and letters. Note that after fixing the position of the training word, letters and bigrams were presented in the corresponding locations within the word. We then performed a repeated measure analysis of variance on the mean cosine distance, with Unit (3 levels: letters, contiguous bigrams, and non-contiguous bigrams) and Type (2 levels: constituent vs. non-constituent) as factors. Results (see **Figure 5**) showed significant main effects of Unit, $F_{(2,\ 238)} = 78.09$, $p < 0.0001$, $\eta_p^2 = 0.4$, and Type, $F_{(1,\ 119)} = 23271.38$, $p < 0.0001$, $\eta_p^2 = 0.99$. The interaction was also significant, $F_{(1.92,\ 228.62)} = 395.31$, $p < 0.0001$, $\eta_p^2 = 0.77$ (Huynh-Feldt corrected for sphericity). Paired $t$-tests (Bonferroni corrected) showed that, for each of the sub-word units, there was a higher similarity with constituent than non-constituent units [letters: $t_{(119)} = -52.39$, $p < 0.0001$; contiguous bigrams: $t_{(119)} = -118.86$, $p < 0.0001$; non-contiguous bigrams, $t_{(119)} = -95.93$, $p < 0.0001$]. For constituent units, non-contiguous bigrams had higher similarity to the target words with respect to both single letters [$t_{(119)} = 39.49$, $p < 0.0001$] and contiguous bigrams [$t_{(119)} = 3.86$, $p < 0.0001$]. Contiguous bigrams had higher similarity than letters [$t_{(119)} = 42.109$, $p < 0.0001$]. For non-constituent units, only single letters showed a significant difference from the other sub-word units [contiguous bigrams, $t_{(119)} = -6.64$, $p < 0.0001$; non-contiguous bigrams, $t_{(119)} = -4.28$, $p < 0.0001$].

[2]Cosine distance between patterns X and Y is calculated as 1-cosine (X, Y).

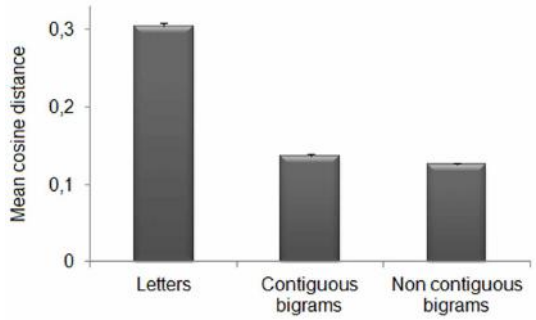



**FIGURE 5 | Pattern analysis on HL3.** Mean cosine distance between internal representations for each word and sub-word units (i.e., letters, contiguous bigrams, and non-contiguous bigrams). Note that smaller values index higher similarity between activation patterns. The sub-word unit showing the highest similarity to the corresponding word was the non-contiguous bigram, that is the combination of exterior letters (word edges).

In summary, the activation pattern of each word was more similar to those of the constituent rather than those of non-constituent sub-word units. Importantly, among constituent units, non-contiguous bigrams (i.e., those formed by the first and last letter) produced an activation pattern that was more similar to that of the corresponding word in comparison to both letters and contiguous bigrams. For constituent letters and contiguous-bigrams we performed a further analysis (i.e., two-tailed paired $t$-tests, Bonferroni corrected), in order to establish whether the position of the constituent stimuli within the word was important. Results revealed no significant differences among positions for both letters (first letter: $M = 0.31 \pm 0.005$ s.e.m., second letter: $M = 0.30 \pm 0.005$ s.e.m., third letter: $M = 0.30 \pm 0.004$ s.e.m.; all $t$s $< 2.01$) and contiguous bigrams (first bigram: $M = 0.14 \pm 0.003$ s.e.m., second bigram: $M = 0.13 \pm 0.002$ s.e.m.; $t = 1.55$). We also assessed whether the higher similarity of the non-contiguous bigram pattern to the word pattern with respect to the continuous bigram patterns would persist when the leftmost and rightmost word locations (1 and 5) were excluded from the analysis. The results did not change [non-contiguous vs. contiguous bigrams: $t_{(119)} = 3.71$, $p < 0.0001$].

## DISCUSSION

Models of orthographic coding (e.g., Grainger and van Heuven, 2003; Grainger and Whitney, 2004; Gomez et al., 2008; Davis, 2010) share the assumption that visual word recognition is performed through the processing of constituent letters but differ on how letter position information is coded and whether the mapping between location-specific (eye-centered) letter coding and location-invariant word representations requires the computation of an intermediate orthographic code, such as open bigrams. A prior attempt to tackle these issues through a connectionist approach has led to contrasting results. After training a feedforward neural network with one hidden layer (using error back-propagation) on the mapping between a location-specific letter code and location-invariant localist word

representations, Dandurand et al. (2010) showed computational evidence supporting the bigram coding hypothesis. However, subsequent analyses of the hidden layer representations carried out by Hannagan et al. (2011) suggested that in this network model the mapping does not imply the extraction of information about letter combinations but it is based on semi-location-invariant letter representations that are broadly consistent with the overlap model of Gomez et al. (2008).

Our current attempt to use connectionist simulations for cracking the orthographic code is tied to a more general framework suggesting that perceptual invariance can emerge from unsupervised learning in a hierarchical processing architecture that extracts increasingly more complex and abstract features (Hinton, 2007, 2013; Stoianov and Zorzi, 2012; Zorzi et al., 2013), as well as to the hypothesis that visual word recognition recycles the cortical machinery used for visual object recognition (Dehaene and Cohen, 2011; Dehaene et al., 2005). Accordingly, we exploited deep neural networks (Hinton and Salakhutdinov, 2006) to investigate whether location-invariant word recognition might emerge from unsupervised learning of a hierarchical generative model of location-specific letter patterns. Although word-level information (i.e., word identity) was never provided to the network during training, linear decoding from the activity of the deepest hidden layer yielded near-perfect accuracy in location-invariant word recognition. In contrast, decoding accuracy from lower HLs showed a sharp and progressive decrease, with a pattern of errors suggesting that letter position information was not coded in a location-invariant way. Indeed, the majority of the word decoding errors, especially at the second hidden layer, consisted of transposition errors. This finding is consistent with the transposition priming effect, as predicted by both letter-based (e.g., Gomez et al., 2008; Davis, 2010) and open bigram (e.g., Grainger and van Heuven, 2003; Grainger and Whitney, 2004) models of orthographic coding.

We then carried out a series of analyses to investigate the nature of the orthographic representations emerged at the different HLs. Analysis on single neurons showed that only the deepest layer of the deep network contained neurons that were both word selective and location-invariant. Interestingly, some word-selective neurons found at the second hidden layer were tuned to specific word locations. These results are in line with those provided by the decoding analysis and confirm that linear decoding of hidden layer activity is an helpful method for investigating the internal representations emerged in a deep network model. Notably, the single neuron analysis showed that bigrams did not emerge as unit of representation in the network. This finding fits well the results of Hannagan et al. (2011) in their re-analysis of the Dandurand et al. (2010) model and it is broadly consistent with letter-based models of orthographic coding (Gomez et al., 2008; Davis, 2010). It is worth noting that learning a generative model is equivalent to discovering efficient ways of coding the input data (Ghahramani et al., 1999); this suggests that the information carried by bigrams is not necessary for efficient orthographic coding, at least in the context of the highly constrained training set employed in the present study (see further discussion below).

In a final set of analyses, we recorded the activation patterns over the deepest hidden layer produced by each word and compared them to those produced by letters and bigrams. This analysis provides a measure of similarity between internal representations that can be readily interpreted in terms of priming effect. Not surprisingly, constituent letters and bigrams (including non-contiguous bigrams) had an advantage over non-constituent ones. Moreover, constituent bigrams had an advantage over constituent letters, which is also expected due the increasing orthographic overlap (i.e., two letters vs. one letter). Interestingly, we also found a significant greater similarity for non-contiguous bigrams (i.e., those formed by the first and last letter) over contiguous bigrams. The advantage for non-contiguous bigrams persisted when the extreme word locations (1 and 5) were excluded from the analysis. The superiority of non-contiguous bigrams with respect to the other constituent stimuli might be interpreted as an index of the edge effect (Fischer-Baum et al., 2011), that is the superiority of the first and last letters for coding words as a sequence of ordered letters, observed using the illusory word paradigm. Fischer-Baum and colleagues argued that the edge effect supports an orthographic coding scheme in which the beginning and the end letters of a word act as anchoring points. Though several models assume that the exterior letters have special status in orthographic coding (e.g., Gomez et al., 2008; Davis, 2010), our model shows that this aspect is an emergent property that does not require additional mechanisms or specific parameters.

In conclusion, our study shows that location-invariant visual word recognition can emerge from unsupervised learning in a neural network with a deep (hierarchical) architecture. Our deep network model extracted increasingly more complex and abstract orthographic features across layers. Moreover, our analyses show that the emergent orthographic code is not based on bigrams and it assigns special status to the exterior letters (word edges). Although restricting our simulations to an artificial dataset of 3-letter strings is indeed an important limit of the current study, this allowed us to investigate orthographic coding in a simplified and tightly controlled way. Future extensions of this work will therefore focus on scaling-up the training dataset and on testing the model on a corpus of real words. For example, it cannot be excluded that the distributional statistics of letters in real words, whereby some letter combinations have higher frequency than others, might lead to the emergence of sub-word units like bigrams (see Dandurand et al., 2011, for analyses of English, French and Spanish word corpora). Nevertheless, we believe that our preliminary findings pave the way for a better understanding of how orthographic representations can emerge through unsupervised learning within a sound probabilistic framework.

## ACKNOWLEDGMENTS

This study was supported by the European Research Council (grant no. 210922 to Marco Zorzi).

## REFERENCES


Cohen, L., and Dehaene, S. (2004). Specialization within the ventral stream: the case for the visual word form area. *Neuroimage* 22, 466–476. doi: 10.1016/j.neuroimage.2003.12.049

Cohen, L., Lehéricy, S., Chochon, F., Lemer, C., Rivaud, S., and Dehaene, S. (2002). Language-specific tuning of visual cortex. Functional properties of the visual word form area. *Brain* 125, 1054–1069. doi: 10.1093/brain/awf094

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204. doi: 10.1037/0033–295X.108.1.204

Dandurand, F., Grainger, J., and Dufau, S. (2010). Learning location-invariant orthographic representations for printed words. *Connect. Sci.* 22, 25–42. doi: 10.1080/09540090903085768

Dandurand, F., Grainger, J., Duñabeitia, J. A., and Granier, J. P. (2011). On coding non-contiguous letter combinations. *Front. Psychol.* 2:136. doi: 10.3389/fpsyg.2011.00136

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychol. Rev.* 117, 713–758. doi: 10.1037/a0019738

Davis, C. J., and Bowers, J. S. (2006). Contrasting five theories of letter position coding: evidence from orthographic similarity effects. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 535–557. doi: 10.1037/0096-1523.32.3.535

De Filippo De Grazia, M., Cutini, S., Lisi, M., and Zorzi, M. (2012a). Space coding for sensorimotor transformations can emerge through unsupervised learning. *Cogn. Process.* 13, S141–146. doi: 10.1007/s10339-012-0478-4

De Filippo De Grazia, M., Stoianov, I., and Zorzi, M. (2012b). "Parallelization of deep networks," in *Proceedings of the 2012 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning—ESANN* (Bruges), 621–626.

Dehaene, S., and Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends Cogn. Sci.* 15, 254–262. doi: 10.1016/j.tics.2011.04.003

Dehaene, S., Cohen, L., Sigman, M., and Vinckier, F. (2005). The neural code for written words: a proposal. *Trends Cogn. Sci.* 9, 335–341. doi: 10.1016/j.tics.2005.05.004

Fischer-Baum, S., Charny, J., and McCloskey, M. (2011). Both-edges representation of letter position in reading. *Psychon. Bul. Rev.* 18, 1083–1089. doi: 10.3758/s13423-011-0160-3

Ghahramani, Z., Korenberg, A. T., and Hinton, G. E. (1999). "Scaling in a hierarchical unsupervised network," in *Proceedings of the 1999 Ninth International Conference on Artificial Neural Networks—ICANN* (Edinburgh), 13–18.

Gomez, P., Ratcliff, R., and Perea, M. (2008). The overlap model: a model of letter position coding. *Psychol. Rev.* 115, 577–600. doi: 10.1037/a0012667

Grainger, J. (2008). Cracking the orthographic code: an introduction. *Lang. Cogn. Proc.* 23, 1–35. doi: 10.1080/01690960701578013

Grainger, J., Dufau, S., Montant, M., Ziegler, J. C., and Fagot, J. (2012). Orthographic processing in baboons (Papio papio). *Science* 336, 245–248. doi: 10.1126/science.1218152

Grainger, J., Granier, J. P., Farioli, F., Van Assche, E., and van Heuven, W. (2006). Letter position information and printed word perception: the relative-position priming constraint. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 865–884. doi: 10.1037/0096-1523.32.4.865

Grainger, J., and van Heuven, W. J. B. (2003). "Modeling letter position coding in printed word perception," in *The Mental lexicon*, ed P. Bonin (NewYork, NY: Nova Science Publishers), 1–23.

Grainger, J., and Whitney, C. (2004). Does the huamn mnid raed wrods as awlohe. *Trends Cogn. Sci.* 8, 58–59. doi: 10.1016/j.tics.2003.11.006

Grainger, J., and Ziegler, J. C. (2011). A dual-route approach to orthographic processing. *Front. Psychol.* 2:54. doi: 10.3389/fpsyg.2011.00054

Hannagan, T., Dandurand, F., and Grainger, J. (2011). Broken symmetries in a location-invariant word recognition network. *Neural Comput.* 23, 251–283. doi: 10.1162/NECO_a_00064

Harm, M. W., and Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychol. Rev.* 106, 491–528. doi: 10.1037/0033-295X.106.3.491

Hertz, J. A., Krogh, A. S., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Weasley.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* 14, 1771–1800. doi: 10.1162/089976602760128018

Hinton, G. (2013). Where do features come from. *Cogn. Sci.* doi: 10.1111/cogs.12049. [Epub ahead of print].

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends Cogn. Sci.* 11, 428–434. doi: 10.1016/j.tics.2007.09.004

Hinton, G. E. (2012). "A practical guide to training Restricted Boltzmann Machines," in *Neural Networks: Tricks of the Trade*, eds G. Montavon, G. Orr and K. R. Müller (Berlin; Heidelberg: Springer), 599–619.

Hinton, G. E., and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Humphreys, G. W., Evett, L. J., and Quinlan, P. T. (1990). Orthographic processing in visual word identification. *Cogn. Psychol.* 22, 517–560. doi: 10.1016/0010-028590012-S

Lee, H., Ekanadham, C., and Ng, A. Y. (2008). Sparse deep belief net models for visual area V2. *Adv. Neural Inf. Process. Syst.* 20, 873–880.

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: part 1. An account of basic findings. *Psychol. Rev.* 88, 375–407. doi: 10.1037/0033-295X.88.5.375

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends Cogn. Sci.* 2, 455–462. doi: 10.1016/S1364-6613(98)01241-8

Pelli, D. G., Farell, B., and Moore, D. C. (2003). The remarkable inefficiency of word recognition. *Nature* 423, 752–756. doi: 10.1038/nature01516

Perea, M., and Lupker, S. J. (2003). "Transposed-letter confusability effects in masked form priming," in *Masked Priming: State of the Art*, eds S. Kinoshita and S. J. Lupker (Hove: Psychology Press), 97–120.

Peressotti, F., and Grainger, J. (1999). The role of letter identity and letter position in orthographic priming. *Percept. Psychophys.* 61, 691–706. doi: 10.3758/BF03205539

Perry, C., Ziegler, J. C., and Zorzi, M. (2007). Nested incremental modelling in the development of computational theories: the CDP+ model of reading aloud. *Psychol. Rev.* 114, 273. doi: 10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J. C., and Zorzi, M. (2010). Beyond single syllables: large-scale modelling of reading aloud with the connectionist dual process (CDP++) model. *Cogn. Psychol.* 61, 106–151. doi: 10.1016/j.cogpsych.2010.04.001

Perry, C., Ziegler, J. C., and Zorzi, M. (2013). A computational and empirical investigation of graphemes in reading. *Cogn. Sci.* 37, 800–828. doi: 10.1111/cogs.12030

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56. doi: 10.1037/0033-295X.103.1.56

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rumelhart, D. E., and McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychol. Rev.* 89, 60–94. doi: 10.1037/0033-295X.89.1.60

Schoonbaert, S., and Grainger, J. (2004). Letter position coding in printed word perception: effects of repeated and transposed letters. *Lang. Cogn. Process.* 19, 333–367. doi: 10.1080/01690960344000198

Seidenberg, M. S., and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychol. Rev.* 96, 523. doi: 10.1037/0033–295X.96.4.523

Stoianov, I., and Zorzi, M. (2012). Emergence of a "visual number sense" in hierarchical generative models. *Nat. Neurosci.* 15, 194–196. doi: 10.1038/nn.2996

Testolin, A., Stoianov, I., De Filippo De Grazia, M., and Zorzi, M. (2013). Deep unsupervised learning on a desktop PC: a primer for cognitive scientists. *Front. Psychol.* 4:251. doi: 10.3389/fpsyg.2013.00251

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., and Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: dissecting the inner organization of the visual word-form system. *Neuron* 55, 143–156. doi: 10.1016/j.neuron.2007.05.031

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: the SERIOL model

and selective literature review. *Psychon. Bull. Rev.* 8, 221–243. doi: 10.3758/BF03196158

Yarkoni, T., Balota, D., and Yap, M. (2008). Moving beyond Coltheart's N: a new measure of orthographic similarity. *Psychon. Bull. Rev.* 15, 971–979. doi: 10.3758/PBR.15.5.971

Ziegler, J. C., Perry, C., and Zorzi, M. (in press). Modelling reading development through phonological decoding and self-teaching: implications for dyslexia. *Proc. R. Soc. Lond. B.*

Zorzi, M. (2005). "Computational models of reading," in *Connectionist Models in Cognitive Psychology*, ed G. Houghton (Hove, UK: Psychology Press), 403–444.

Zorzi, M. (2010). The connectionist dual process (CDP) approach to modelling reading aloud. *Eur. J. Cogn. Psychol.* 22, 836–860. doi: 10.1080/09541440903435621

Zorzi, M., Houghton, G., and Butterworth, B. (1998). Two routes or one in reading aloud. A connectionist dual-process model. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 1131. doi: 10.1037/0096-1523.24.4.1131

Zorzi, M., Testolin, A., and Stoianov, I. (2013). Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Front. Psychol.* 4:515. doi: 10.3389/fpsyg.2013.00515

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 16 May 2013; accepted: 27 August 2013; published online: 19 September 2013.*

*Citation: Di Bono MG and Zorzi M (2013) Deep generative learning of location-invariant visual word recognition. Front. Psychol.* ***4****:635. doi: 10.3389/fpsyg.2013.00635*

*This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 04 October 2013
doi: 10.3389/fpsyg.2013.00676

# A computational model to investigate assumptions in the headturn preference procedure

**Christina Bergmann[1,2]*, Louis ten Bosch[1], Paula Fikkert[1] and Lou Boves[3]**

[1] Centre for Language Studies, Radboud University Nijmegen, Nijmegen, Netherlands
[2] International Max Planck Research School for Language Sciences, Max Planck Institute for Psycholinguistics, Radboud University Nijmegen, Nijmegen, Netherlands
[3] Centre for Language and Speech Technology, Radboud University Nijmegen, Nijmegen, Netherlands

**Edited by:**
*Julien Mayor, University of Geneva, Switzerland*

**Reviewed by:**
*Padraic Monaghan, Lancaster University, UK*
*Caspar Addyman, University of London, UK*

***Correspondence:**
*Christina Bergmann, Centre for Language Studies, Radboud University Nijmegen, Erasmusplein 1, 6525 HT Nijmegen, Netherlands*
*e-mail: cbergmann@science.ru.nl*

In this paper we use a computational model to investigate four assumptions that are tacitly present in interpreting the results of studies on infants' speech processing abilities using the Headturn Preference Procedure (HPP): (1) behavioral differences originate in different processing; (2) processing involves some form of recognition; (3) words are segmented from connected speech; and (4) differences between infants should not affect overall results. In addition, we investigate the impact of two potentially important aspects in the design and execution of the experiments: (a) the specific voices used in the two parts on HPP experiments (familiarization and test) and (b) the experimenter's criterion for what is a sufficient headturn angle. The model is designed to be maximize cognitive plausibility. It takes real speech as input, and it contains a module that converts the output of internal speech processing and recognition into headturns that can yield real-time listening preference measurements. Internal processing is based on distributed episodic representations in combination with a matching procedure based on the assumptions that complex episodes can be decomposed as positive weighted sums of simpler constituents. Model simulations show that the first assumptions hold under two different definitions of recognition. However, explicit segmentation is not necessary to simulate the behaviors observed in infant studies. Differences in attention span between infants can affect the outcomes of an experiment. The same holds for the experimenter's decision criterion. The speakers used in experiments affect outcomes in complex ways that require further investigation. The paper ends with recommendations for future studies using the HPP.

**Keywords: headturn preference procedure, language acquisition, segmentation, attention, speech processing**

## 1. INTRODUCTION

Infants begin to acquire what will become their native language long before they produce meaningful speech themselves. The last decades have seen a substantial growth in experimental studies that explore this pre-verbal phase of language acquisition, with a particular focus on how infants process speech input. The advent of behavioral research paradigms that tap into infants' underlying cognitive abilities made this research line possible. The paradigms recruit actions infants can readily perform in their daily lives. The prime example of such a paradigm is the Headturn Preference Procedure (HPP), which uses the eponymous headturns to investigate speech processing.

The HPP is based on the observation that infants tend to turn their heads toward interesting events. The time this headturn in maintained is interpreted as infants' amount of interest. Jusczyk and Aslin (1995) demonstrated how the HPP can be used to investigate infants' ability to memorize and recognize speech[1]. A common version of the HPP, as used by Jusczyk and Aslin, typically has two phases. In an initial familiarization phase, infants are exposed to words spoken in isolation. In the test phase that immediately follows familiarization, infants listen to sentences that contain either one of the previously heard words or an unfamiliar word. Differences in the time the head is turned toward each of the two types of test stimuli indicate that infants process test stimuli with and without familiar words differently. Jusczyk and Aslin interpreted such listening time differences as the ability of the infants to discover that the familiarized words are present in some of the test sentences.

Following the seminal work by Jusczyk and Aslin (1995), numerous studies have utilized the HPP to investigate infants' emerging speech processing abilities. Almost invariably, HPP studies use the familiarization-followed-by-test design briefly outlined above, where listening time during the test phase is the behavioral measure (c.f., Section 2 for further details). Subsequent studies have replicated the original finding with infants learning French (Nazzi et al., 2013), Spanish (Bosch et al., 2013), and many other languages. Others have used the HPP to shed light on the influence of various extra-linguistic factors in the processing of speech signals. A number of studies showed that infants cannot readily detect the familiarized words in the test sentences if there are large acoustic differences between familiarization and test phase, for example, when they differ in mood,

[1] For a detailed description of the HPP, see Section 2.

accent, and gender of the speaker (Houston and Jusczyk, 2000, 2003; Singh et al., 2004; Schmale and Seidl, 2009; Schmale et al., 2010)[2].

Although there are few published reports of null-results, failures to replicate the outcome of published HPP experiments are not uncommon (see Ferguson and Heene, 2012; for the bias against publishing papers that report failures to replicate). Furthermore, seemingly comparable studies can yield results that support contradicting interpretations. For example, Houston and Jusczyk (2000) tested infants' ability to detect words spoken by one speaker during familiarization in test passages that were spoken by a different speaker. Thereby the authors investigated is whether infants are able to generalize across speakers. The results showed that infants only listened longer to test stimuli containing familiarized words than to test stimuli with novel words if the speakers' gender matched between familiarization and test phase. In a seemingly comparable study, van Heugten and Johnson (2012) found that gender differences do not seem to matter for infants of the same age as tested by Houston and Jusczyk. In addition, the infants in the study by van Heugten and Johnson showed a novelty preference, where infants listened longer to test stimuli without the familiarized words, while Houston and Jusczyk found a familiarity preference.

It is not yet entirely clear which factors exactly determine the behavior of infants in HPP studies (Houston-Price and Nakai, 2004; Aslin, 2007; van Heugten and Johnson, 2012; Nazzi et al., 2013). Studies using the HPP vary in several aspects, including the stimulus material and implementation details. For example, different speakers are used to record stimuli across experiments, and potentially relevant properties of the stimuli (such as voice characteristics) are difficult to report in a meaningful way. Sharing stimulus material among research groups would be an improvement, but is often not feasible unless infants are acquiring the same language (c.f., Nazzi et al.). Differences in implementation are exemplified by seemingly varying criteria for a sufficient headturn, ranging from "at least 30° in the direction of the loudspeaker" (Jusczyk and Aslin, 1995, p. 8) to "at least 70° toward the flashing light" (Hollich, 2006, p. 7). It is possible that such differences in assessment criteria, even if used systematically and accurately, can cause conflicting results.

In addition to these practical issues with HPP studies, there is a more fundamental question that urgently needs attention. In behavioral paradigms, including the HPP, the cognitive processes of interest must be inferred from observable behavior, and these inferences rely on numerous assumptions about the link between overt behavior and cognitive processes. Most behavioral data are compatible with different, perhaps conflicting, assumptions and interpretations (Frank and Tenenbaum, 2011). The present paper addresses these practical and fundamental issues by using a computational model that simulates the test situation of the HPP. The use of a computational model allows for the investigation of fundamental issues, because the implementation of the procedure makes crucial assumptions explicit, and model simulations make it possible to assess whether these assumptions are necessary to simulate infant behavior. At the same time simulations allow us to study the impact of differences in stimulus material and in the practical implementation of the HPP. Although the model is — by necessity — a simplified analogue of an infant (or a group of infants) in an HPP experiment, we aim for its operations and representations to be as cognitively plausible as possible. In consequence, the model simulations can help to better understand the outcome of HPP experiments.

The remainder of this paper is organized as follows: In Section 2 we first describe the HPP in detail along with the assumptions that are commonly made in the interpretation of the results in infant studies before we introduce our computational model in Section 3. We explain how the model makes it possible to test the assumptions discussed in Section 2.1. In addition, we outline how the model is built to maximize cognitive plausibility. The design of the experiments that allow us to investigate the impact of the stimulus material and details of how HPP experiments are conducted is further elaborated on in Section 4. Section 5 presents the results of our experiments. The paper concludes with a general discussion and outlines the implications of the modeling results for the interpretation of results reported in infant studies.

## 2. THE HEADTURN PREFERENCE PROCEDURE

HPP experiments typically consist of two consecutive phases, as **Figure 1** illustrates using an example from the experiments by Jusczyk and Aslin (1995). In the first phase an infant is familiarized with a specific audio(-visual) phenomenon (here: spoken words and the accompanying flashing lamp). The criterion for familiarization is usually a cumulative listening time of at least 30 s for each word. When the familiarization criterion is met the second phase immediately commences. In this phase the infant's

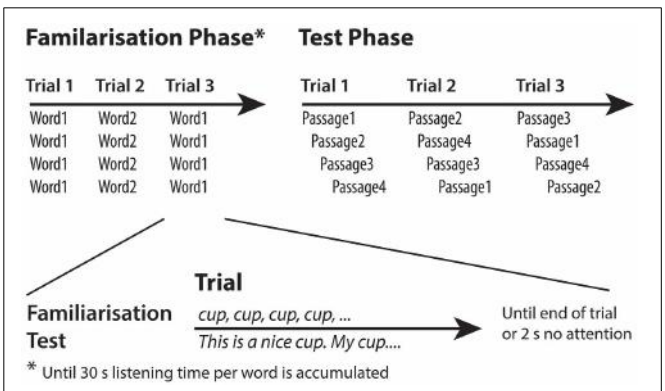



**FIGURE 1 | Exemplary outline of a two-phase headturn experiment.** Infants first hear words spoken in isolation and then listen to sentences that do or do not contain these words.

[2]The HPP has also been used to investigate infants' ability to discover regularities in auditory input (see Frank and Tenenbaum, 2011; for a summary of studies in that field). However, these studies generally use artificial speech and require monitoring of a continuous monotone speech stream, arguably a different task from the segmentation studies conducted following the work of Jusczyk and Aslin (1995).

reaction to test stimuli is measured that either contain the two familiarized words or two novel words[3].

In the study of Jusczyk and Aslin (1995), infants were familiarized with two words spoken in isolation (either *cup* and *dog*, or *feet* and *bike*). In the test phase passages of six sentences containing one of the four words were presented in each trial[4]. The infants listened longer to passages containing words with which they were familiarized, as indicated by their maintained headturns (see below for details). Hence, infants showed sufficient memory and processing abilities to store and detect words and to overcome an acoustic difference between embedded and isolated words. Based on their results Jusczyk and Aslin concluded that infants have segmented the passages into smaller chunks and detected the embedded words.

The rationale behind the HPP is that the time an infant spends with the head turned toward a side lamp while presumably listening to speech stimuli coming from that same side indicates the infant's interest in the stimuli. The experimental set-up based on this rationale is depicted in **Figure 2**. Infants are placed in a three-sided booth with lamps on each wall, one in front of the infant and one on each side. A loudspeaker is mounted beneath each side lamp. Through a video camera facing the infant, the experimenter observes the infant's movements and controls the experiment. A trial starts with the center lamp flashing. As soon as the infant attends to that lamp by turning toward it, one of the side lamps begins to flash, and the central lamp turns off. When the infant turns her head to the side lamp by a pre-determined angle off-center, speech stimuli begin to play from the loudspeaker beneath the flashing side lamp. As long as the head is turned toward the side lamp, the trial continues. Turning the head away for more than 2 consecutive seconds ends the trial prematurely. If the infant turns her head back toward the lamp before 2 s have elapsed the trial is not ended. The time during which the head was turned away is not measured as listening time. Importantly, while headturn angle is a continuous variable, it is converted into a binary criterion by the experimenter: the head is, or is not, turned sufficiently toward the side lamp and the loudspeaker at any moment throughout the trial. The side of the flashing lamp and of presenting the speech stimuli is counterbalanced and bears no relation to the type of trial.

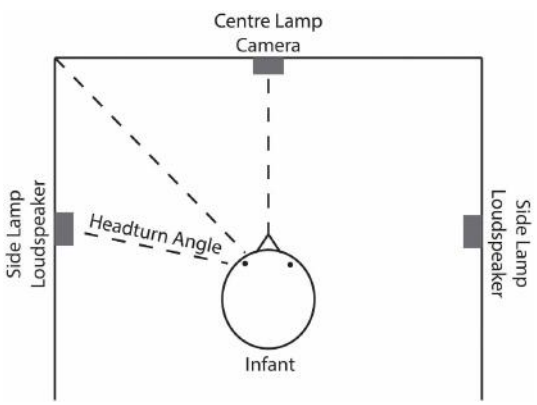



**FIGURE 2 | Schematic outline of the experimental set-up in headturn studies.** The infant is placed in a three-sided booth with lamps on each side and loudspeakers to the left and right. Through a frontal camera, the headturns are observed by the experimenter.

[3]Some HPP studies familiarize with paragraphs of continuous sentences and test with words in isolation, but in the present paper we focus on the predominant set-up.

[4]Jusczyk and Aslin (1995), Experiments 1–3 of 4.

### 2.1. ASSUMPTIONS IN THE HEADTURN PREFERENCE PROCEDURE

The HPP aims to tap into infants' linguistic abilities by inferring cognitive processes (in particular speech processing) from observable behavior. Linking overt behavior in HPP experiments to infants' underlying cognitive processes is based on at least four main (implicit) assumptions, which are not straightforward to test experimentally.

First, a listening preference for one type of test stimulus stems from some form of underlying *recognition* of recently heard words. In their seminal work, Jusczyk and Aslin (1995) equate recognition with the detection of a sufficiently high degree of similarity between perceived sound patterns. In a two-phase HPP experiment, presumably unknown words are presented to the infant during familiarization, and then two sets of previously unknown words are compared in testing (one familiarized and one novel). It is thus measured how infants react to words that were recently presented in comparison to entirely novel words.

Second, systematic differences in listening time to passages containing familiar or novel words are due to systematic internal processing differences. Infants' behavior in HPP studies is assumed to be resulting from several processing steps: infants have to internally process speech input and match it to representations stored in internal memory. The memory contains representations of experience before the lab visit as well as representations stored during the familiarization phase, whereas the focus lies on the memorization of familiarized items.

Third, recognition of words in passages, while those words were presented in isolation during familiarization, requires infants to be able to segment words from continuous speech prior to matching. Segmentation entails the chunking of speech into smaller parts and representing those constituents independently.

Fourth, differences between individual infants do not affect the outcome of an experiment, as the main comparison (listening to novel or familiar test stimuli) takes place within participants. This assumption mainly concerns infant-specific factors independent of their linguistic abilities.

## 3. MODELING THE HEADTURN PREFERENCE PROCEDURE

First we outline how the model architecture and the simulations aim to address the assumptions discussed in Section 2.1. The model subscribes to the first two assumptions. Following the first assumption, recognition is implemented in the model in the form of a matching process which compares test items to the familiarized stimuli along with a form of past experience. The contents of the memory that the matching process works on are described in Section 3.3, the matching process that operates on the memory is explained in detail in Section 3.4. Section 3.5 lays out how recognition can be implemented. In accordance

with the second assumption, the matching procedure should yield systematically different outcomes that signify the model's internal ability to distinguish novel and familiar test items. Based on the outcome of the matching procedure, headturns are simulated. The conversion of internal recognition into overt behavior is discussed in Section 3.6. The third assumption will be assessed by our model. The claim that infants are able to segment words from continuous speech utterances seems unnecessarily strong. A strong segmentation procedure is difficult to implement without assuming that the model decodes and memorizes speech in the form of sequences of discrete linguistic units (such as syllables and phonemes), an ability that infants are still in the process of acquiring (Kuhl, 2004; Newman, 2008). Therefore, we follow the proposal that infants are able to divide a passage consisting of a sequence of six naturally spoken utterances, separated by clear pauses, into the constituting sentences (Hirsh-Pasek et al., 1987; Jusczyk, 1998). The model thus receives its test input in the form of complete sentences, as Sections 3.2 and 3.3 describe. If the model is able to distinguish familiar from novel test items, we show that segmentation is not necessary in the two-phase HPP studies simulated in the present paper. We will investigate the fourth assumption that differences between individual infants do not affect the outcome of an experiment. The role of an infant-dependent parameter that transforms internal recognition into overt headturns will be investigated to this end (see Section 3.6 for further detail).

Simulations with varying criteria for a sufficient degree of headturn assess the impact of implementation details. Furthermore, we use speech produced by four speakers to address the role of the stimulus material in HPP experiments and the model's ability to generalize across speakers. These issues will be explained in more detail in Sections 3.7 and 4.

### 3.1. THE MODEL ARCHITECTURE

We developed a computational model that, despite the necessary simplifications, is as cognitively plausible as possible. The model contains general purpose processing skills which infants would also need for other tasks. The architecture of the model during the familiarization phase is shown in **Figure 3**. All input consists of real speech that proceeds through a sequence of processing steps, which are explained in detail in the following sections. In the model, the familiarization phase is simulated by storing the stimuli in an internal model memory that is already populated by episodic representations of speech (and sounds) that the modeled infant heard before the lab visit (Goldinger, 1998). The details of the model memory are described in Section 3.3.

The focus in the present paper lies on applying the model to the test situation, as depicted in **Figure 4**. During the test, the model hears test sentences, which are processed and encoded in the same way as the contents of the internal memory (c.f., Section 3.2). Using the matching procedure described in Section 3.4, weights for the complete memory content are generated, which correspond to the strength of the contribution of every episode stored in the memory to processing a test stimulus. Based on the weights of the familiarization episodes and the past experience (c.f., **Figure 3**), a measure of recognition is computed (c.f., Section 3.5). An independent process transforms the internal familiarity score into overt behavior, as explained in Section 3.6. This allows for a direct comparison of the model output to the results of infant experiments. In the following sections we describe the model in detail.

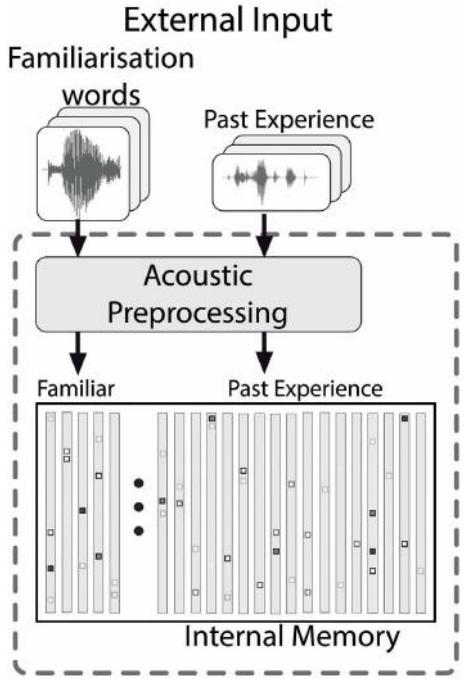



**FIGURE 3 | The memory structure of the model, which contains both the familiarized items and past experience.** Acoustic preprocessing is applied to all contents of the memory.

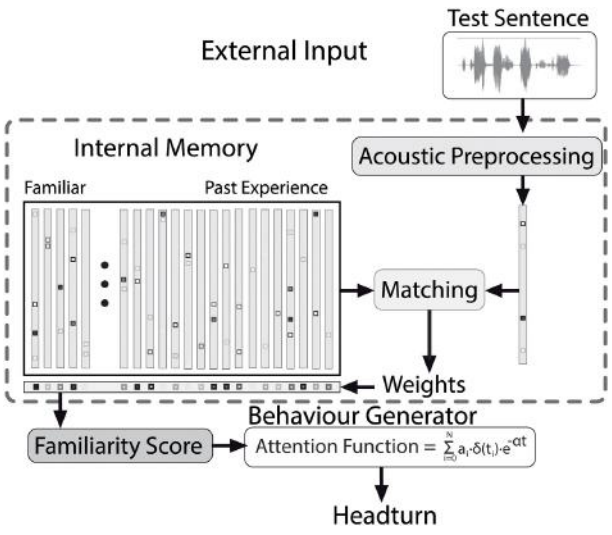



**FIGURE 4 | The Headturn Preference Procedure model during the test phase, with processing stages and flow of information from external input (top) to overt behavior (bottom).**

### 3.2. ACOUSTIC PREPROCESSING

The processing of the acoustic speech signals starts with representing the continuous wave form in terms of its frequency and power at a given moment and the change of these properties of the speech signal over time. From the literature it appears that infant auditory processing is compatible with this form of signal processing (Saffran et al., 2007). The continuous speech signal is divided into windows with a duration of 20 ms, and for each such window a short-time spectrum is computed (Coleman, 2005). Adjacent windows overlap by 10 ms, we thus obtain 100 short-time spectra per second. The short-time spectra are converted to vectors of 13 real numbers, the Mel-Frequency Cepstral Coefficients (MFCCs), a representation that is based on knowledge about human auditory processing (Gold and Morgan, 2000). Because the auditory system is more sensitive to the rate of change in the spectrum than to static spectral features, we add the difference between adjacent MFCC vectors (known as $\Delta$ coefficients in the automatic speech processing literature) as well as the differences between adjacent $\Delta$s (known as $\Delta\Delta$s). $\Delta$s and $\Delta\Delta$s are vectors comprising 13 real numbers. The resulting MFCC, $\Delta$ and $\Delta\Delta$ vectors corresponding to successive windows of a speech signal, are used to learn a limited number of acoustic phenomena, or prototypes. In our model we use 150 prototypes for static MFCC vectors, 150 prototypes for the $\Delta$ vectors, and 100 prototypes for the $\Delta\Delta$ vectors[5]. These prototypes are used to condense the information in the MFCC, $\Delta$ and $\Delta\Delta$ vectors, by representing each MFCC vector by its best matching prototype (and doing the same for all $\Delta$ and $\Delta\Delta$ vectors). This converts a representation in the form of $3 * 13 = 39$ real numbers to a set of three labels from a set of $150 + 150 + 100$ prototypes. The conversion of the infinite number of possible MFCC, $\Delta$ and $\Delta\Delta$ vectors to sets of three labels corresponds to the—admittedly unproven but plausible—assumption that audio signals are represented in the brain as sequences of acoustic prototypes.

Variable-length sequences of prototypes corresponding to an utterance must be converted to a fixed-length representation to be used in a matching procedure. For this purpose we count the number of occurrences and co-occurrences of prototypes. This results in a so called Histogram of Acoustic Co-occurrences (HAC, Van hamme, 2008). The histogram keeps a count of the number of times each of the $150 + 150 + 100$ acoustic prototypes co-occurs with any prototype in its own class (including itself) at distances of 20 and 50 ms. Including co-occurrences at lags of 20 and 50 ms allows HAC vectors to capture some information about the temporal structure of an utterance. In total, a HAC vector has slightly more than 100,000 entries for all possible prototype co-occurrences. As a result, an utterance of arbitrary length, be it a single word or a complete sentence, is represented by a HAC vector of a fixed dimension. The fixed dimensionality is a requirement for most matching procedures.

### 3.3. INTERNAL MEMORY

Infants in HPP experiments have been exposed to speech prior to their lab visit. Therefore, the model's memory should contain some acoustic representations of past experience. Specifically, the memory contains HAC representations of a number of previously heard utterances. During the familiarization phase the acoustic HAC representations of the familiarization words are added to the memory. Therefore, the collection of HAC vectors in the memory during the test phase comprises two types of entries: the experience before the start of the HPP experiment, and the episodes the infant has stored during the familiarization phase.

The infant's experience with speech input before the lab visit is modeled by randomly selecting utterances from a corpus of infant-directed speech (Altosaar et al., 2010). Familiarization consists of adding HAC representations of tokens of two words to the memory. Although technically the model uses one single homogeneous memory, we assume that infants are able to distinguish the familiarization entries in the test from the entries from previous experience. A compelling justification for this distinction would be to assume that the familiarization utterances are stored as episodes in the hippocampus, while the previous experience is stored in the cortex (Kumaran and McClelland, 2012).

### 3.4. MATCHING PROCEDURE: NON-NEGATIVE MATRIX FACTORIZATION

In the test phase, depicted in **Figure 4**, a matching procedure is necessary to compare an input stimulus to the contents of the model's memory. This matching procedure should yield scores that can be transformed into a score that corresponds to how well the representations in the memory match any particular unknown input. Episodic representations of a small number of stimuli, such as the ones the model stored during familiarization, are not straightforwardly compatible with conventional Neural Networks and similar types of Parallel Distributed Processing. Therefore, the model contains a matching procedure that is based on the assumption that the brain processes complex inputs as a non-negative weighted sum of a limited number of simpler units stored in memory. This assumption is inspired by studies on visual processing, which found that complex visual patterns are represented in primary visual cortex in the form of lines, directions, colors, and so forth (c.f., Lee and Seung, 1999; and citations therein).

Non-negative Matrix Factorization (NMF, Lee and Seung, 1999) approximates a given input (in the present simulations a HAC vector) as a weighted sum of all stored representations (here also HAC vectors) in the internal memory. Usually, NMF learns the primitives from a set of stimuli before it can be used for 'recognizing' unknown input, but in simulating HPP experiments we skip the NMF learning phase, and use only the decomposition mechanism. NMF can be phrased in the same terms as activation and inhibition in neural networks (Van hamme, 2011). This makes NMF , especially in the implementation that enables incremental learning (Driesen et al., 2009), a potentially interesting alternative to conventional Artificial Neural Net and Parallel Distributed Processing techniques for simulating language acquisition.

The variant of NMF used in the present paper minimizes the Kullback–Leibler divergence between a HAC-encoded test stimulus and its approximation as a positive weighted sum of all representations stored in the memory. Decoding of an unknown

[5]We used about 30 min of speech produced by two female and two male speakers of Dutch to learn the prototypes.

utterance results in a set of non-negative weights for each representation stored in the memory. The higher the weight assigned to a representation, the larger its contribution to explaining the unknown input. These weights become available immediately after the end of a test utterance[6].

### 3.5. RECOGNITION AND FAMILIARITY SCORES

The matching procedure described in the previous section yields weights for all entries of the memory. The model converts these weights into a *familiarity score* that describes how well the test stimulus was recognized. The familiarity scores drive observable behavior (see the next sections). We compare two possible ways to compute familiarity scores and thereby simulate recognition.

In the first method, the familiarity score represents how much the single best-matching episode stored in memory during the familiarization phase contributes to approximating an unknown utterance in the test phase (in the presence of all other entries in the memory). This form of recognition will therefore be called *single episode activation*. In cognitive terms, single episode activation corresponds to the proposal that an infant treats the tokens of the familiarization stimuli as independent episodes that are not related to each other. This is motivated by the large acoustic differences between familiarization tokens of the same word that can be observed in the stimuli used in some HPP experiments. The second method, in which the familiarity score accumulates the weights of all familiarization entries, corresponds to the idea that the infant treats all episodes stored during familiarization as a cluster of tokens that all relate to one type of experience. This implementation of recognition will be termed *cluster activation* throughout the paper.

The scores are computed as follows: In the first implementation, the familiarity score is set equal to the *maximum* of the weights of all familiarization entries, while in the second method the familiarization score is defined by the *sum* of the weights of the familiarization entries. Both implementations of recognition yield familiarity scores that can be considered as a measure of the activation of memory representations resulting from the acoustic processing and matching procedures in the model. The familiarity score is computed independently for each test sentence. In the model we have access to the familiarity scores of each test utterance, which is evidently not possible in infants. To investigate whether familiarity scores corresponding to sentences containing a familiarized word are treated systematically differently from sentences without a familiarized word we subject the scores to independent statistical tests.

### 3.6. BEHAVIOR GENERATION

In HPP studies, the time an infant maintains a headturn toward a flashing side lamp is measured as an overt sign of underlying attention to the speech stimuli presented via a loudspeaker on the same side. Attention is in turn driven by internal recognition. Familiarity scores, which represent cognitive processing, cannot be observed directly in infant experiments. To convert a sequence of familiarity scores to a headturn angle that varies continuously over time, our model transforms the discrete-time familiarity scores that become available at the end of each sentence in a test passage into a continuous attention function which directly drives headturns. The attention function's value at a particular time point can be interpreted as the degree to which the head is turned toward the flashing lamp and the loudspeaker. While the function value is high, the infant's head is completely turned toward the flashing lamp. As the attention value decreases, the head is more likely to be turned away from the lamp.

In the module that converts familiarity scores into the continuous attention function, we assume that attention is renewed whenever a new familiarity score is computed (at the end of a test sentence) and that attention wanes exponentially during the course of the next sentence. The discrete-time familiarity scores are converted to discrete pulses $a_i \cdot \delta(t_i)$ with an amplitude $a_i$ equal to the familiarity score of the $i$th test utterance, separated by the duration of the utterances (see **Figure 5**, top panel, for an illustration). The sequence of pulses $a_i \,.\, \delta(t_i)$ is converted into a continuous function by applying an exponential decay. The resulting attention function for a passage with $N$ sentences is defined as $\sum_{i=0}^{N} a_i \cdot \delta(t_i) \cdot e^{-\alpha t}$. In this function $\alpha$ is a (positive) parameter specifying the decay rate, and $t$ denotes time. The value of $a_0$, the value of the attention function at the moment that the test passage starts playing depends on the value of a separate parameter $\rho$ (see Section 3.7 for details). **Figure 5** illustrates the link between pulses $a_i \,.\delta(t)$ based on the familiarity scores (top panel) and the corresponding attention function with different values for $\alpha$ (bottom panel).

The decay rate $\alpha$ can be interpreted as the attention span of an infant. Small values of $\alpha$ correspond to a long attention span, while larger values of $\alpha$ cause the attention function to decrease more rapidly, which leads to shorter attention spans. A fixed

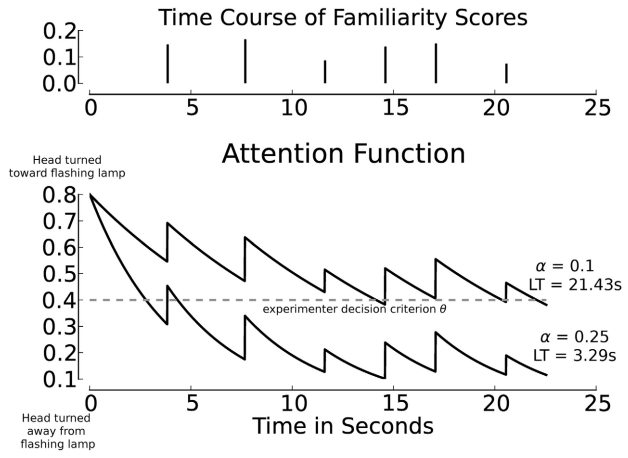



**FIGURE 5 | Familiarity scores, separated by sentence duration (top panel), and exemplary corresponding attention functions (bottom panel) using grouped activations.** All material was spoken by Speaker M1. The threshold $\theta$ is set to 0.4 (dashed line), resulting listening times (LT) across exemplary values for $\alpha$ are annotated. In all cases the initial attention level is 0.8, which exceeds the threshold $\theta$. The decay parameter $\alpha$ is independent of the familiarity scores.

[6]To allow for comparisons between the decoding of different utterances, the weights obtained after each stimulus are normalized to sum to one.

exponential decay rate, which corresponds to an attention span that is constant for the complete duration of an experiment, is undoubtedly a strong simplification of the cognitive processes involved in converting the results of perceptual processing into observable behavior. However, there are no behavioral data that can be used to implement more complex procedures. The parameter $\alpha$ makes it possible to investigate whether differences in attention span between individual infants can affect the outcomes of an HPP experiment.

It should be noted that restricting a possible impact of attention span to the test phase implies that we do not model differences between infants during the familiarization phase of an HPP experiment. Effectively, the way in which we construct the memory after familiarization corresponds to the assumption that an infant pays full attention and that there are no errors in the perceptual processing. Again, this is a simplification that can only be justified by quoting a complete absence of behavioral data that would allow creating a more realistic model.

### 3.7. SIMULATING THE TEST SITUATION

In simulating the test situation, an experimenter's evaluation of infants' responses to a sequence of sentences in a test passage has to be modeled. To this end, the attention function for a passage consisting of several test sentences is assessed in a way comparable to HPP studies. In an infant study, the experimenter interprets the angle of the head relative to the center and side lamps in terms of discrete states throughout a test trial (c.f., **Figure 2**). The criterion that an experimenter uses to determine whether the head is turned sufficiently toward a side lamp is modeled by a threshold $\theta$ that is applied to the attention function. As long as attention exceeds $\theta$, the head is considered to be turned sufficiently in the direction of the flashing lamp. As soon as the attention level drops below $\theta$, the experimenter decides that the head is turned away from the lamp to such a degree that presumably the infant is no longer listening to the speech stimuli. If the value of the attention function stays below $\theta$ for more than 2 consecutive seconds, the trial is terminated (as in HPP studies).

The parameter $\rho > 0$ models the initial attention level above the threshold $\theta$ at the start of a test trial. It can be conceptualized as the initial degree of interest in the flashing lamp at trial onset. The value of $a_0$, the value of the attention function at trial onset (time $t = 0$), is defined as $\theta + \rho$, which guarantees that the infant's head is turned toward the flashing lamp sufficiently to be considered interested. In the simulations presented below, this parameter (interest at trial onset beyond threshold) was kept constant. Previous research showed that the parameter $\rho$ does not affect the simulation results in a cognitively interesting manner (Bergmann et al., 2012). It appeared that a fixed value $\rho = 0.4$ was representative for the explored range of values and consequently was chosen for the present paper[7]. In **Figure 5**, $\theta$ and the resulting listening times obtained with two exemplary attention functions are shown. The functions are derived from the same sequence of familiarity scores (top panel); the difference between depicted attention functions and resulting listening times is due to changes in the value of $\alpha$. The attention function for $\alpha = 0.25$ is shown for the total duration of a test six-sentence passage. In a HPP experiment the trial would be aborted during the third sentence, because the head was turned away from the loudspeaker for more than 2 consecutive seconds.

[7]Increasing or decreasing the initial interest modeled in $\rho$ shifts the overall outcome within the parameter space of $\{\alpha, \theta\}$ but does not impact the general outcome.

## 4. EXPERIMENTS

In the present paper, we test assumptions underlying the interpretation of HPP studies (c.f., Section 2.1), as well as two practical issues using a computational model. We briefly recall the four assumptions and explain how these are addressed in the experiments. Subsequently, we explain how the simulations address the implementation issues.

Initially, we test whether the model conforms to the assumption that test passages containing familiar and novel words yields systematic differences in internal processing and resulting listening times in two stages. In the first stage we investigate whether familiar passages yield significantly higher familiarization scores than unfamiliar passages. Thereby, we assess the model's internal ability to discriminate the two types of test stimuli. In the second stage it is tested whether the procedure that converts internal familiarization scores into overt headturns and listening times can enhance or obscure significantly different familiarization scores.

We investigate the relation between listening preference and internal recognition of the test passages by comparing two definitions of recognition (c.f., Section 3.5). In *single episode activation* the familiarity scores are based on the familiarized token in the model's memory that receives the highest weight. In *cluster activation* the familiarity scores are based on the sum of the weights of the 10 familiarization tokens in the memory. From the explanation of the model in Section 3 it will be clear that neither definition of recognition involves explicit word segmentation. If the simulations yield significant differences between test passages with familiar and with novel words, it would seem to call into question the claim that word segmentation is necessary for infants to show the observed behavior in HPP experiments. The fourth assumption that differences between individual infants do not affect the outcome of an HPP experiment will be investigated by running simulations with different values of the attention span parameter $\alpha$ (c.f., Section 3.6.)

In addition to the fundamental assumptions in interpreting the outcomes of HPP experiments our simulations address two implementation issues: the effects of stimulus materials and the impact of varying criteria for a sufficient degree of headturn. We run simulations with four speakers, and we will investigate familiarity scores and listening times for all combinations of these speakers in familiarization and test. By doing so, we aim to contribute to clarifying the seemingly contradicting results of previous HPP experiments on infants' generalization abilities (e.g., Houston and Jusczyk, 2000; van Heugten and Johnson, 2012). The effect of the experimenter decision criterion for a sufficient degree of headturn will be investigated by simulations with a range of values for the parameter $\theta$ (c.f., Section 3.7).

From simulations with previous versions of the computational model it became clear that many of the issues addressed above are not independent (e.g., Bergmann et al., 2012). That makes it

impossible to design experiments that address one single issue in isolation. We will mitigate this problem by coming back to the individual issues in the general discussion.

### 4.1. SPEECH MATERIAL

Our computational model requires three types of acoustic stimuli to simulate HPP studies: words spoken in isolation for familiarization, the same words embedded in continuous sentences for creating test passages, and utterances that do not contain the target words to model past language experience. All speech material in the present paper stems from a corpus of words and sentences spoken by native speakers of British English (Altosaar et al., 2010)[8]. The recordings were made in a virtually noise-free environment. Four adult speakers were available for the present study, two of whom were female.

The target words in our study were *frog* and *doll* or *duck* and *ball*. These were the words in the corpus that were most similar to the original stimuli of Jusczyk and Aslin (1995) who used monosyllabic words containing various vowels and at least one stop consonant. For each target word, five tokens spoken in isolation were available. To build the corresponding test passages, we randomly selected 24 short sentences for each of the four words. These sentences were identical for all four speakers. With these sentences a large number of distinct six-sentence test passages can be constructed by random selection.

Duration differences must be caused by different speech rates between speakers, as the sentences were identical. The mean sentence durations are between 2.69 s (standard deviation 0.33 s) for Speaker F1 and 3.0 s (standard deviation 0.39 s) for Speaker F2. The two male speakers show intermediate speech rates with 2.88 s (standard deviation 0.42 s) for Speaker M1 and 2.79 s (standard deviation 0.33 s) for Speaker M2. The range of speech rates indicates that the four speakers pronounce the same sentences at a different pace. Through the fixed time lags used to encode the acoustic input (see Section 3.2), each speaker will yield different HAC encoded vectors based on the diverging speech rates alone. We do not compensate for this source of speaker differences since there is little evidence that infants before their first birthday apply such speaker normalization (Houston and Jusczyk, 2000).

In all simulations, the internal memory consisted of 111 HAC vectors, 10 containing the two familiarized words (5 tokens for each) and 100 sentences comprising the past experience spoken by the same speaker. One additional HAC vector contained background noise (silence obtained during the recording session). The choice of 100 HAC vectors to model previous experience was motivated by exploratory simulations in which we investigated familiarity scores with memory sizes ranging from 50 to 1000 utterances to represent previous experience. Although the weights assigned to the familiarization tokens may decrease as the number of previous experience tokens increases, the relative difference between the weights of the familiarization tokens for familiar and novel test sentences is hardly affected. The NMF approximation of a test sentence will use the complete memory contents. If a familiarization token in memory is a good match for a test sentence, this is hardly changed by the number of other tokens in memory. The decision to use 100 entries for previous experience is in a sense arbitrary, but it does not crucially affect the results.

[8]The speech material will be made available through The Language Archive at tla.mpi.nl.

## 5. RESULTS

The description of the results is split into two parts: First we describe the outcome of internal speech processing in the model in terms of familiarity scores. Thereby we assess the model's underlying ability to recognize familiar words in the test sentences. Subsequently, we simulate listening times and assess how the transformation of familiarity scores into overt behavior affects our results.

### 5.1. FAMILIARITY SCORES

We first assess whether internal speech processing outcomes in the model can distinguish test sentences that contain familiarized words from sentences with novel words. To this end we investigate whether the familiarity scores for all 96 test sentences per speaker, used once as familiar and once as novel test item, are significantly different. For this purpose we apply the non-parametric Mann–Whitney *U*-Test. We chose this test because its efficiency is comparable to the *t*-Test with normally distributed data, while it is more robust when the data contain unequal variances or outliers.

All test sentences were recognized twice by models that were familiarized with speech from each of the four speakers. In the first recognition run the keyword in the sentence was familiar, in the second run it was novel. The whole experiment is conducted twice, once with the *single episode activation* and once with the *cluster activation* definition of recognition. Familiarity scores are computed in the manner described in Section 3.5 and are reported in percent for clarity.

#### *5.1.1. Single episode activation*

Computing familiarity scores based on the single episode that receives the maximum activation yields a mixed pattern of results. The descriptive values for familiarity scores corresponding to familiar and novel test sentences can be found

**Table 1 | Mean (and standard deviation) of the familiarity scores for familiar and novel sentences across speaker combinations in % with single episode activation.**

| | | | Test speaker | | | |
|---|---|---|---|---|---|---|
| | | | **M1** | **F1** | **M2** | **F2** |
| Fam. speaker | M1 | fam | **5.03 (2.73)**** | **6.26 (3.23)**** | 8.43 (6.28) | **6.35 (3.99)*** |
| | | nov | **4.11 (2.70)**** | **5.06 (2.32)**** | 8.19 (5.87) | **5.70 (4.49)*** |
| | F1 | fam | 8.61 (3.73) | **5.76 (2.90)**** | 16.60 (7.30) | 15.20 (8.50) |
| | | nov | 9.21 (4.34) | **4.75 (3.06)**** | 15.87 (5.48) | 15.10 (8.00) |
| | M2 | fam | 7.40 (3.94) | 11.67 (5.80) | **8.60 (4.26)*** | 16.58 (6.43) |
| | | nov | 7.75 (4.34) | 12.14 (6.32) | **7.41 (3.43)*** | 15.57 (5.62) |
| | F2 | fam | 8.15 (4.29) | 6.89 (4.83) | 9.62 (4.94) | 8.32 (5.26) |
| | | nov | 7.91 (4.42) | 6.18 (4.27) | 10.00 (4.76) | 7.46 (4.44) |

*Values that differ significantly across test stimulus types are marked in bold. Significance level markers are: $^{*}p < 0.05$, $^{**}p < 0.01$.*

in **Table 1**. The table shows the average (and standard deviation) of the familiarity scores for all speaker pairs. Each cell contains data for the sentences in the familiar ("fam") and novel ("nov") condition. It can be seen that the mean values and standard deviations differ between speaker pairs. The familiarity scores are expressed in terms of the percentage of the weights of the 111 memory entries assigned to the single highest-scoring familiarization token stored in the model's memory.

We find statistically significant higher scores for familiar than for novel test items in five of 16 speaker pairs. Except for Speaker F2, the distinction between test conditions is statistically significant when the speaker does not change between familiarization and test. The lack of a significant difference between familiar and novel stimuli for Speaker F2 may be due to the standard deviations that are relatively large compared to the mean.

Next to the cases where the speaker did not change between familiarization and test, we see two pairs in which the test speaker was different from the familiarization speaker that yield statistically significant distinctions of familiar and novel test items. When the model has stored familiarization words spoken by Speaker M1 in memory, test sentences spoken by Speaker F1 and Speaker F2 yield significantly different familiarity scores. Interestingly, the results do not show an advantage of same-sex pairs over mixed-sex pairs.

#### *5.1.2. Cluster activation*

Taking the sum of the weights for all familiarized items in memory yields statistically significant differences between familiar and novel test sentences for the four cases where familiarization and test speaker are identical, as shown in **Table 2**. The table is formatted in the same way as **Table 1**, and the values displayed refer to the percentage assigned to all 10 memory representations of the familiarized tokens. The mixed-gender speaker pairs {M1, F1} and {M1, F2} show significant differences between familiar and novel test sentences (as was the case with single episode activation). Again, we do not observe a clear advantage of same-sex pairs over mixed-sex pairs.

**Table 2 | Mean (and standard deviation) for the familiarity scores for familiar and novel sentences across speaker combinations in % with cluster activation.**

| | | | Test speaker | | | |
|---|---|---|---|---|---|---|
| | | | **M1** | **F1** | **M2** | **F2** |
| Fam. speaker | M1 | fam | **15.30 (7.97)**** | **14.56 (4.51)**** | 22.31 (11.90) | **16.23 (7.44)*** |
| | | nov | **12.69 (7.96)**** | **12.63 (4.89)**** | 21.11 (11.98) | **14.39 (7.27)*** |
| | F1 | fam | 24.72 (7.92) | **15.93 (5.18)***** | 37.37 (9.96) | 27.40 (10.78)† |
| | | nov | 24.05 (8.68) | **12.08 (5.23)***** | 36.10 (9.61) | 24.92 (10.13)† |
| | M2 | fam | 21.19 (8.52) | 24.46 (7.80) | **23.10 (7.25)***** | 35.88 (8.02)† |
| | | nov | 20.74 (8.06) | 23.64 (8.52) | **19.20 (6.77)***** | 34.00 (7.72)† |
| | F2 | fam | 21.80 (8.99) | 16.54 (8.66)† | 29.14 (10.84) | **21.96 (9.51)***** |
| | | nov | 20.52 (7.90) | 14.51 (8.03)† | 27.52 (9.10) | **17.93 (9.27)***** |

*Values that differ significantly across test stimulus types are marked in bold. Significance level markers are: †p < 0.1, *p < 0.05, **p < 0.01, ***p < 0.001.*

### 5.2. DISCUSSION

Overall, the model implements the assumption that processing sentences with familiar words yields higher familiarity scores than sentences with novel words, which is confirmed by the results of the simulations. The differences between familiarity scores for familiar and novel test items are larger if the speakers in familiarization and test are identical, but there is no clear effect of the sex of the speaker. The differences between the absolute values of the familiarization scores in the single episode and cluster activation runs were to be expected: sums of a set of positive numbers will always be larger than the largest individual member of a set. Perhaps the most intriguing difference between single episode and cluster activation is present when Speaker F2 utters all speech material: in the single episode activation, familiar sentences yielded no statistically significant higher familiarity scores than novel sentences, while the difference is highly significant with cluster activation.

### 5.3. SIMULATED LISTENING TIMES

In the previous section we found that our model tends to assign higher internal familiarity scores to test sentences with a familiar word than to comparable sentences with a novel word. We used these sentences to create 30 six-sentence test passages for each of the four words (*frog, doll, duck, ball*) that could be used during familiarization. Sentences were selected randomly, with replacement. Each passage contained one of the four words, which could, depending on the familiarization words, be familiar or novel. This was done for all 16 possible speaker pairs, and for the two definitions of recognition. All sequences were converted to attention functions using the procedure explained in Section 3.6, whereby we explore a range of values of the attention span parameter $\alpha$. **Figure 5** shows an example of one sequence, with two values of $\alpha$. The value of $\alpha$ varied between 0.01 and 0.3, in steps of 0.01. Previous experiments with the model have shown that this range covers all cognitively relevant phenomena (Bergmann et al., 2012, 2013).

In our model, we treat the continuous attention function as identical to the headturn angle. The higher the attention function, the more the head is turned toward the side lamp (c.f., **Figure 5**). To compute listening times given an attention function, we need an additional parameter to model the experimenter's decision wheter the head is turned sufficiently toward the side. For that purpose we use the parameter $\theta$ explained in Section 3.7. The total listening time corresponding to a passage is the cumulated time during which the value of the attention function is above $\theta$ (counting up to the moment when the attention function is below $\theta$ for more than 2 consecutive seconds). In the simulations we varied the value of $\theta$ between 0.1 and 1.5 in steps of 0.01. Although we cannot quantify the relation between $\theta$ and the headturn angle in an infant experiment, we can say that higher values of $\theta$ correspond to stricter criteria imposed by the experimenter. Values of $\theta > 1.5$ make the criterion so strict that most listening times become effectively zero. Very small values of $\theta$ yield listening times that are almost invariably equal to the duration of the passages.

To obtain an overview of the listening time differences as a function of $\alpha$ and $\theta$ we depict the results in the form of Hinton plots (**Figures 6**, **7**). The figures show the $\{\alpha, \theta\}$ combinations for

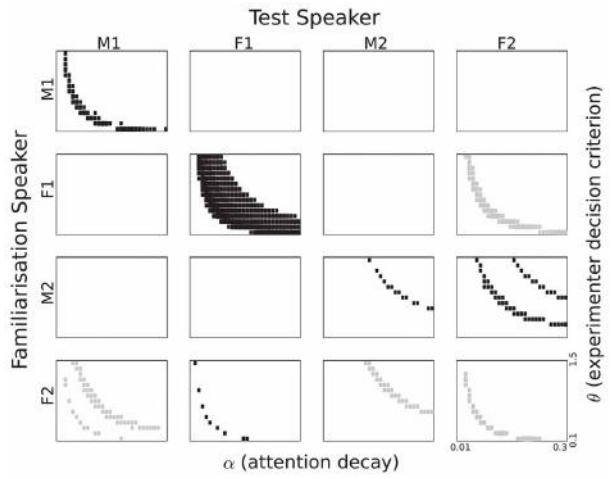



**FIGURE 6 | Listening time differences for all speaker pairings based on single episode activation.** The section of the parameter space displayed corresponds to 0.1 to 1.5 for θ and 0.01 to 0.3 for α. Rectangle size corresponds to the *p*-value in a two-sample *t*-test. Black rectangles correspond to a familiarity preference, grey rectangles to a novelty preference.

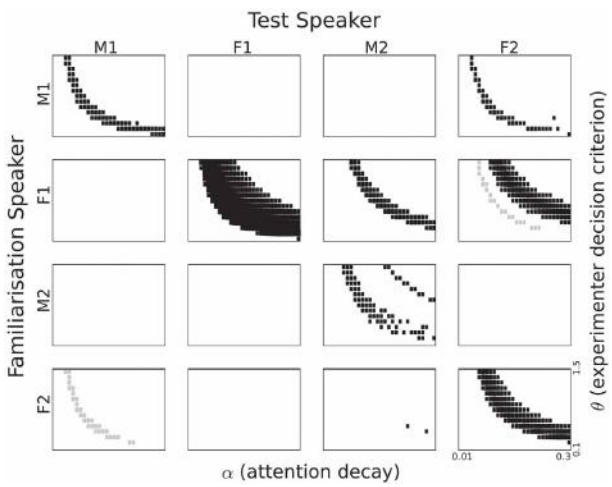



**FIGURE 7 | Listening time differences for all speaker pairings based on cluster activation.** The section of the parameter space displayed corresponds to 0.1 to 1.5 for θ and 0.01 to 0.3 for α. Rectangle size corresponds to the *p*-value in a two-sample *t*-test. Black rectangles correspond to a familiarity preference; grey rectangles to a novelty preference.

which the listening time difference between familiar and novel passages was significant with $p < 0.05$. The size of the rectangles in the figures corresponds to the significance level. If the listening time is longer for the familiar passages, the rectangles are black. Grey rectangles correspond to {α, θ} combinations in which there was a significantly longer listening time for the novel passages. *p*-values were computed using a two-sample *t*-test in which two sets of 120 passages were compared: 30 for each of the two words, which were used twice (as familiar and as novel) to remove biases caused by the fact that sentences corresponding to the words were of unequal length. We did not apply a correction for multiple comparisons for two reasons. First, it is not completely clear how many {α, θ} combinations must be included in a full comparison. For a substantial proportion of the combinations, the listening time difference is exactly zero, due to reasons that are independent of the goals of the present paper. When both α and θ are large, the attention function drops below the threshold θ more than 2 s before the end of the first sentence in a passage[9]. If both parameters have very small values, the attention function will stay above θ for the full duration of the passage. The {α, θ} pairs for which this happens might have to be excluded. One can take the position that listening time differences caused by the last sentence in a passage should also be discarded. The second reason for not adjusting the *p*-values is inspired by the shapes of the trajectories in the {α, θ} plane that can be seen in the figures. It is highly unlikely that continuous trajectories would emerge if there was no underlying process that causes the listening time differences. This procedure is similar to the procedures used in brain imaging, where the large number of comparisons between voxels would lose much of the relevant information if a straightforward adjustment would be applied, ignoring the underlying physical processes (Forman et al., 1995).

#### *5.3.1. Single episode activation*

Significant listening time differences based on internal single episode activation are displayed in **Figure 6** for all speaker pairings. The first thing that strikes the eye is the large difference between the four speakers. While three out of the four same-speaker pairs show a trajectory in the {α, θ} plane with a significant familiarity preference, it is also evident that the trajectory for Speaker F1 is much more robust than for the other speakers. For Speaker M2 we see a very thin trajectory. Interestingly, Speaker F2 appears to give rise to a novelty preference, despite the fact that we designed the model to yield a familiarity preference. It can also be seen that the trajectories are not always at the same area in the {α, θ} plane.

In addition to the same-speaker pairs, there are also between-speaker pairs that yield trajectories with significant differences. There is no unambiguous gender effect. The pair {M1, M2} shows no significance at all, but there are some pairings that show significant listening preferences. The patterns are not symmetric, as can be seen best for the pair M1 and F2. Familiarization with M1 gives no significant listening preferences when testing with F2, vice versa, there are substantial significant trajectories for M1 as test speaker. The lack of symmetry is perhaps most striking in the case of the two female speakers. When Speaker F1 utters the familiarization stimuli and Speaker F2 the test material, we see a novelty preference. However, when the roles are reversed between speakers a novelty preference emerges. We also see a novelty preference in the {F2, M2} pair.

[9]Up to the end of the first sentence in a passage the attention function depends only on the decay parameter α. The familiarity scores only take effect after the end of an utterance.

*5.3.1.1. Attention span and experimenter decision criterion.* In **Figure 6** it can be seen that significant listening time differences are obtained for a wide range of values for α (on the horizontal axis), except for speaker M2. The absence of significant differences between listening times to familiar and novel passages for speaker M2 for small values of α (long attention span) is caused by the fact that the attention function never drops below the θ threshold.

**Figure 6** shows an effect of the strictness with which the experimenter interprets the headturn angle, modeled by the parameter θ. For high values of θ significant listening time differences are only obtained in combination with long attention spans (lower values for α). As the value of θ decreases, significant listening time differences (both familiarity and novelty preferences) can be obtained with shorter attention spans (higher values for α). At this point we refrain from interpreting the parabolic shapes of the trajectories in the figure because a different quantization of α and θ would yield other shapes.

*5.3.1.2. Familiarity or novelty preference.* From comparing the data in **Table 1** and the patterns in **Figure 6** it can be seen that there is no straightforward relation between familiarity scores for individual sentences and listening preference. Apparently, the way in which sentences are concatenated to form a passage has a substantial effect. If a sentence that yields a relatively small familiarity score is followed by a relatively long sentence, the next reset of the attention function, at the end of that sentence, may come too late to avoid the cut-off of the 2-s rule.

For some speaker pairs we see a novelty preference. Perhaps the most striking example is when the speaker F2 utters all speech material, the more so because the familiarity scores for this speaker in **Table 1** suggests a familiarity preference with slightly higher values for familiar than for novel test items. However, when we base the attention function on the familiarity score of a single memory entry, it cannot be ruled out that the maximum value of a novel utterance is higher than the maximum of a familiar sentence. This can give rise to a novelty preference.

### 5.3.2. Cluster activation

The significantly different listening times as a function of the two parameters α and θ for the cluster activation definition of recognition can be seen in **Figure 7**. This definition corresponds to the assumption that infants treat all familiarization stimuli as referring to a single concept and that they aim to detect references to that concept in the test passages. Numerically, summing over the activations of all 10 familiarization entries in the memory to compute a familiarity score should make that score less sensitive to seemingly random effects.

In **Figure 7** we see a strong familiarity preference in all same-speaker pairs, even for speaker F2, for whom we found a novelty preference in the single episode activation case. Again, there is no unambiguous gender effect. The male speakers M1 and M2 share no pattern, while the relation between the two female speakers is quite complex. Perhaps the most striking effect is the clear familiarity preference for M2 as test speaker, if the familiarization speaker is F1. Again, we see that there is no straightforward relation between the sentence-based familiarity score data in **Table 2** and the significant listening time differences in **Figure 7**.

*5.3.2.1. Attention span and experimenter decision criterion.* Again, we see parabola-shaped patterns of significant differences in the {α, θ} plane. As α becomes larger, the decay of the attention function becomes more rapid, and a lower value of θ is needed to keep the attention function above threshold. As mentioned in the previous section, we refrain from interpreting those shapes since they depend on the quantization of the explored parameters.

*5.3.2.2. Familiarity or novelty preference.* All same-speaker pairs now show a clear familiarity preference. Apparently, reducing the impact of individual memory entries leads to overall more homogeneous familiarity scores. These scores in turn lead to a familiarity preference in listening times across all four speakers.

When Speaker F1 is used to familiarize the model and Speaker F2 as the test speaker, we see a familiarity preference for some {α, θ} combinations, and a novelty preference for other combinations. This suggests that minor variations in attention span in combination with small changes in the strictness of the experimenter can cause the result of an experiment to switch from a familiarity preference to a novelty preference. While this might indeed happen in infant studies, it cannot be ruled out that the switch seen in **Figure 7** is, at least in part, due to a property of the behavior generating module that is exaggerated by small changes in the decision threshold. The effect can be illustrated with the attention function for $\alpha = 0.25$ in **Figure 5**. If the first familiarity score would have been slightly larger, the duration of the time interval where the function is below the threshold θ might have become less than 2 s. If the familiarity score for the second sentence would have been higher, listening time would increase (even if the two-second rule would have cut off the experiment during the course of the third sentence in the passage). The same effect can be caused by small changes in the threshold θ. This can be observed in the simulations with familiarization stimuli from Speaker F1 and test passages from Speaker M2.

**Figures 8**, **9** provide additional support for the observation that small differences in familiarity scores, combined with specific values of α and θ, can result in switches between familiarity and novelty preference in our model. **Figure 8** shows the cumulative distributions of the familiarity scores of the sentences spoken by Speaker M2 if the familiarization Speaker was M2 himself (left panel) or F2 (right panel). It can be seen that when all stimuli stem from Speaker M2, the familiarity scores are slightly but systematically higher for familiar test sentences. This is different when F1 is the familiarization speaker. As long as the familiarity scores are low, the scores for novel sentences are slightly higher than the scores for familiarized sentences. When the familiarity scores get higher, we see a cross-over point, where the familiarity scores for the familiarized utterances become larger than the corresponding scores for the sentences in the novel condition. **Figure 9** depicts listening times to familiar and novel test sentences for two example speaker pairs (the same as in **Figure 8**) as a function of α with the assessment threshold θ set to 0.3. It can be seen in the left panel that the systematically lower scores for the novel sentences yield accordingly longer listening times in the familiar test condition for the whole range of values for α where listening time is not identical to the full duration of a passage. The right panel of the

plot shows a novelty preference for longer attention spans, which switches to a familiarity preference as the value for α increases.

**Figure 9** furthermore illustrates the general effect of α on the total listening time to novel and familiar passages. For small values of α, where the attention span is long and the attention function decays slowly, the total listening time is equal to the average total duration of the passages (six sentences with an average duration of slightly less than 3 s). As the value of α increases, which means that the attention span shortens, listening times decline. This is caused by a shift of the time point when the attention function drops below θ.

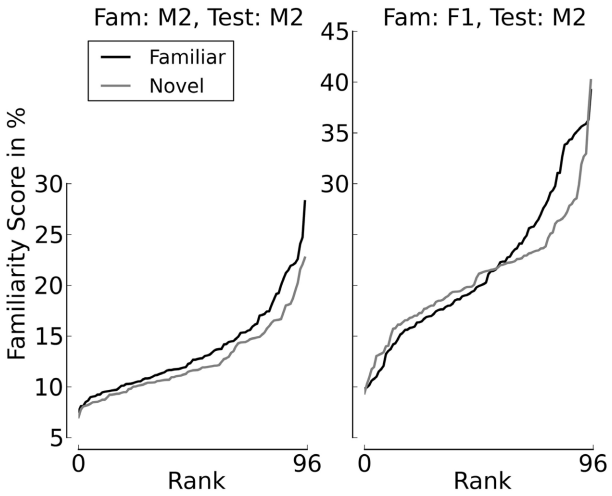



**FIGURE 8 | Familiarity scores for familiar and novel test sentences, sorted by rank.** The left panel depicts a clear familiarity preference. In the right panel, the preferences cross, with lower ranks showing a novelty preference.

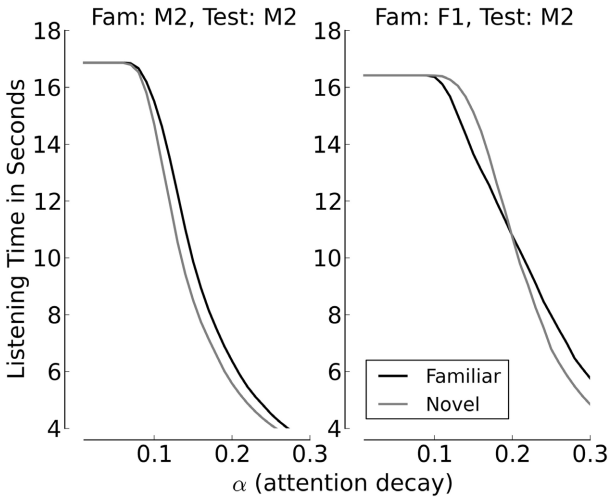



**FIGURE 9 | Listening times (in seconds) across the whole range of values for α, with θ = 0.3.** The left panel shows the listening times when the same speaker, M2, utters familiarization and test stimuli, the right panel shows listening times and when Speaker F1 utters the familiarization stimuli and Speaker M2 the test items.

## 6. DISCUSSION

In the present paper, we investigated four assumptions in the interpretation of experiments that use the HPP, a behavioral method to tap into infants' speech processing abilities. In addition, we investigated two implementation issues that may affect the outcomes of such experiments. Because the four assumptions are difficult to address in infant studies, we took recourse to computational modeling. To this end, we built a computational model that can simulate infant behavior (headturns) observed in HPP studies. The simulations address infant studies which investigated whether infants process test passages that contain words with which the infants were familiarized differently than similar passages that contain novel words (c.f., Jusczyk and Aslin, 1995).

Our model comprises several modules that operate in sequence, in a strict feed-forward architecture. We opted for this modular architecture because it enables us to investigate several processes that have been implicated in the interpretation of HPP studies in isolation. Most importantly, our model makes a distinction between the perceptual processing of the speech stimuli and the process that converts the result of perceptual processing into overt behavior. In addition, the model contains a component that simulates the decisions of the experimenter in HPP studies. Perhaps with the exception of the strict modularity and feed-forward architecture, we put a strong emphasis on making the model as cognitively plausible as possible. It processes real speech that is represented in a way we believe is neurally and cognitively defendable. The implemented matching procedure also can claim cognitive plausibility, if only because it can be combined with learning procedures that can operate in a strictly incremental and causal procedure, in which each input stimulus is used once (instead of iterating multiple times over a corpus of training stimuli).

The basic assumption in HPP studies is that different behaviors are caused by different results of processing the test stimuli. A second assumption in interpreting HPP experiments is that a listening preference for familiar (or novel) passages reflects some form of *recognition*. We defined recognition in two ways, corresponding to different hypotheses of how infants store and access familiarization stimuli during the test phase. The first definition of recognition proposes that an infant treats the familiarization stimuli as independent phenomena. In that interpretation, termed *single episode activation*, recognition was based on the single familiarization entry in the model's memory that matched a test sentence best. The alternative interpretation, *cluster activation*, corresponds to the hypothesis that the infant treats all familiarization stimuli as referring to a single phenomenon. Both definitions of recognition yielded systematic differences in the familiarity scores corresponding to familiar and novel test sentences. With cluster activation, more familiarity score differences were significant than when single episode activations were used. We believe that the larger number of statistically significant differences in the cluster activation case is, at least to a large extent, due to the fact that the sum of 10 activations is less susceptible to random variation than the maximum of a set of 10 values. Therefore,

our simulations do not allow to compare the cognitive plausibility of the two interpretations of the concept of recognition.

A third assumption is that recognition of words embedded in test passages, which were heard in isolation during familiarization, implies infants' ability to segment words from continuous speech. Our model does not rely on segmentation—the division of the speech stream into smaller units, such as words. We found differences between the results of processing sentences with familiarized and novel words and we could replicate infant listening preferences using a representation of the familiarization words and test sentences that have the exact same interpretation: as a bag of acoustic events. Therefore, our model has no need for segmentation procedures. Of course, the simulations do not prove that infants do not segment the speech input, but the experiments show that segmentation skills are not necessary to solve the task posed in the type of HPP studies modeled in the present paper following the work by Jusczyk and Aslin (1995).

In the present paper we do not address studies in which passages were used for familiarization, such as the work by van Heugten and Johnson (2012). However, Jusczyk and Aslin (1995) propose that the two types of experiments are equivalent, whereas the work by Nazzi et al. (2013) indicates that there might be different processes at stake. Addressing this issue is beyond the scope of the present paper and requires further modeling work in conjunction with a careful analysis of the outcome of infant studies that use either words or passages during familiarization.

A fourth assumption in HPP studies is that differences between individual infants do not affect the outcome of an experiment, as the main comparison (listening to novel or familiar test stimuli) takes place within participants. In our model, we simulated differences between infants in the form of varying attention spans. It appeared that if internal familiarity scores distinguish the two types of test stimuli, listening time differences can emerge for a fairly wide range of attention spans. Still, the simulations show that a very short attention span can obscure different familiarity scores in the overt behavior. We deliberately kept the module that converts the results of internal processing into overt behavior very simple, and probably even overly simplistic. We did so because there are no observation data that would allow us to construct a more plausible model. Yet, our simulations show convincingly that the relation between internal processing and externally observable behavior can be complex. Behavior generation can both obscure and enhance differences in the results of internal processing and recognition. In summary, our simulations suggest that the assumption that differences between infants do not affect the results of HPP experiments should be called into question.

We explicitly modeled the experimenter's categorization of infant behavior. Our simulations show that the criterion the experimenter applies can mask listening preferences or enhance them. In addition, there is a strong interaction between the strictness of the experimenter and the attention span of the infant participants. It appeared that slightly different combinations of the factors $\alpha$ (attention span) and $\theta$ (experimenter strictness) can enhance or obscure listening preference and may even lead to switches between familiarity and novelty preference for some combinations of familiarization and test speakers.

We biased our model toward a familiarity preference by focusing on the parts of memory that contain the previously familiarized speech stimuli. However, in various experiments using the HPP, novelty preferences have been observed. Several suggestions regarding the cause of such a preference have been made that implicate developmental or methodological factors (Hunter and Ames, 1988). It has been suggested that individual infants differ in their general input processing strategy (Houston-Price and Nakai, 2004). Novelty preferences might arise from a focus on aspects of the input that are not captured by what has been heard most recently. In our model, different processing strategies can be implemented by changing how familiarity scores are computed from the activations of the memory contents, or from how the familiarity scores are converted to observable behavior. For example, we could discard familiarity scores that exceed an upper bound, treating the corresponding sentences as "more of the same" and therefore uninteresting. In a similar vein, we could assume that attention is aroused by new experiences, rather than by recognizing known things. In such a setting an infant would pay attention to novel stimuli, perhaps not to recognize, but rather to extend the memory by attending to and storing the representations of novel sentences. Alternatively, if we assume that an infant switches from *learning* mode during familiarization to *recognition* mode during test, we might de-emphasize the activations of the familiarization entries in the Hippocampus in favor of the background utterances in the cortex.

The exact source of the novelty preferences generated by our model warrants further investigation into the details of the implementation of the individual modules. The simulations reported in this paper uncovered interactions between the attention function derived from the familiarity scores and the experimenter's decision criterion. This interaction is strengthened by the way in which we compute the familiarity scores. In our model these scores are the result of a sentence-based recognition process. The result is only available after the sentence is complete. Technically, it is possible to change the HAC-based sentence recognition into a continuous-time process (Versteegh and ten Bosch, 2013), but doing so would require the assumption that the memory contains word-like representations.

The voices of four different speakers were used in the present experiments to explore whether non-linguistic properties of the signal can influence the presence of listening preferences. When the speakers did not change between familiarization and test, most familiarity scores were statistically different. Depending on the definition of recognition, the difference for Speaker F2 was or was not statistically significant. In our model it is possible to investigate the voice characteristics that can affect the familiarity scores in great detail. Characteristics that can have an effect depend on the representation of the speech signals in the model. For example, the MFCC representations used in our simulations do not explicitly represent voice pitch, which is reflected in a lack of clear gender-specific effects in our simulations. The co-occurrence statistics in the HAC-representation (c.f., Section 3.2) are sensitive to differences in speaking rate, since they operate with fixed time lags between acoustic events. In this context it is interesting to note that speaker F2 had a slightly lower speaking rate than the other speakers. In addition, HAC-representations

can be sensitive to individual differences in pronunciation. The impact of pronunciation variation depends on the choice of words and passages, an issue that warrants further investigation. Pronunciation variation is a possible factor in infant studies as well. When different speakers are compared according to their accent, an extreme case of pronunciation variation, infants cannot detect words that recur between familiarization and test (Schmale and Seidl, 2009; Schmale et al., 2010). Both differences in speech rate and the possibility of pronunciation variants can also account for the model's mixed abilities to generalize across speakers.

Based on the investigation of the HPP in the present paper, we can make a number of predictions and recommendations for infant research. First, to faithfully measure infants' underlying speech processing abilities, it is helpful to consider their individual attention span. Attention span in the visual domain has been found to positively correlate with language development (Colombo, 2002; Colombo et al., 2008). Measuring individual attentional capabilities can thus at the same time shed light on infants' linguistic development and on an individual factor influencing their performance in HPP studies. Second, carefully defined testing procedures are necessary to allow for consistent and comparable assessments. While it is common practice within labs to have standardized procedures, there is only little exchange of precise assessment criteria across infant laboratories. For greater comparability of published results, a common assessment standard seems to be crucial. Third, an exchange of stimulus material to disentangle the properties of the speakers' voices from language-specific developmental pathways can help shed light on the factors in the stimulus material that can determine the outcome of HPP studies (Nazzi et al., 2013). Existing results using only one or a few speakers do not allow for general statements about the influence of speaker characteristics in HPP studies (c.f., Houston and Jusczyk, 2000; van Heugten and Johnson, 2012).

In summary, modeling the HPP illuminated the role of numerous factors that can determine the outcome of studies utilizing this method. The present paper exemplifies how modeling the task can help linking simulation results of presumed underlying cognitive abilities to overt infant behavior that can be measured experimentally.

## ACKNOWLEDGMENTS

The research of Christina Bergmann is supported by grant no. 360-70-350 from the Dutch Science Organization NWO.

## REFERENCES

Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., and van den Heuvel, H. (2010). "A speech corpus for modeling language acquisition: CAREGIVER," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, (Malta), 1062–1068.

Aslin, R. N. (2007). What's in a look? *Dev. Sci.* 10, 48–53. doi: 10.1111/j.1467-7687.2007.00563.x

Bergmann, C., Boves, L., and ten Bosch, L. (2012). "A model of the headturn preference procedure: linking cognitive processes to overt behaviour," in *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)* (San Diego, CA), 1–6. doi: 10.1109/DevLrn.2012.6400836

Bergmann, C., Boves, L., and ten Bosch, L. (2013). "A computational model of the head-turn preference procedure: design, challenges, and insights," in *Proceedings of the 13th Neural Computation in Psychology Workshop*. Singapore: World Scientific.

Bosch, L., Figueras, M., Teixidó, M., and Ramon-Casas, M. (2013). Rapid gains in segmenting fluent speech when words match the rhythmic unit: evidence from infants acquiring syllable-timed languages. *Front. Psychol.* 4:106. doi: 10.3389/fpsyg.2013.00106

Coleman, J. (2005). *Introducing Speech and Language Processing*. Cambridge: Cambridge University Press.

Colombo, J. (2002). Infant attention grows up: the emergence of a developmental cognitive neuroscience perspective. *Curr. Dir. Psychol. Sci.* 11, 196–200. doi: 10.1111/1467-8721.00199

Colombo, J., Shaddy, D. J., Blaga, O. M., Anderson, C. J., Kannass, K. N., and Richman, W. A. (2008). "Early attentional predictors of vocabulary in childhood," in *Infant Pathways to Language*, eds J. Colombo, P. McCardle, and L. Freund (New York, NY: Psychology Press), 143–168.

Driesen, J., ten Bosch, L., and Van hamme, H. (2009). "Adaptive non-negative matrix factorization in a computational model of language acquisition," in *Proceedings Interspeech* (Brighton).

Ferguson, C. J., and Heene, M. (2012). A vast graveyard of undead theories: publication bias and psychological science's aversion to the null. *Perspect. Psychol. Sci.* 7, 555–561. doi: 10.1177/1745691612459059

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., and Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn. Reson. Med.* 33, 636–647. doi: 10.1002/mrm.1910330508

Frank, M. C., and Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition* 120, 360–371. doi: 10.1016/j.cognition.2010.10.005

Gold, B., and Morgan, N. (2000). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music, Chapter 14*, New York, NY: J Wiley, 189–203.

Goldinger, S. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychol. Rev.* 105, 251–279. doi: 10.1037/0033-295X.105.2.251

Hirsh-Pasek, K., Kemler Nelson, D., Jusczyk, P., Cassidy, K., Druss, B., and Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition* 26, 269–286. doi: 10.1016/S0010-0277(87)80002-1

Hollich, G. (2006). Combining techniques to reveal emergent effects in infants' segmentation, word learning, and grammar. *Lang. Speech* 49, 3–19. doi: 10.1177/00238309060490010201

Houston, D. M., and Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 1570–1582. doi: 10.1037/0096-1523.26.5.1570

Houston, D. M., and Jusczyk, P. W. (2003). Infants' long-term memory for the sound patterns of words and voices. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 1143–1154. doi: 10.1037/0096-1523.29.6.1143

Houston-Price, C., and Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant Child Dev.* 13, 341–348. doi: 10.1002/icd.364

Hunter, M. A., and Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Adv. Infancy Res.* 5, 69–95.

Jusczyk, P. W. (1998). "Dividing and conquering linguistic input," in *Chicago Linguistic Society 34: The panels*, Vol. 2, eds C. Gruber, D. Higgins, K. S. Olson, and T. H. Wysocki. (Chicago, IL: University of Chicago), 293–310.

Jusczyk, P. W., and Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cogn. Psychol.* 29, 1–23. doi: 10.1006/cogp.1995.1010

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nat. Rev. Neurosci.* 5, 831–843. doi: 10.1038/nrn1533

Kumaran, D., and McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychol. Rev.* 119, 573–516. doi: 10.1037/a0028681

Lee, D., and Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Nazzi, T., Mersad, K., Sundara, M., Iakimova, G., and Polka, L. (2013). Early word segmentation in infants acquiring Parisian French: task-dependent and dialect-specific aspects. *J. Child Lang.* doi: 10.1017/

S0305000913000111. [Epub ahead of print].

Newman, R. S. (2008). The level of detail in infants' word learning. *Curr. Dir. Psychol. Sci.* 17, 229–232. doi: 10.1111/j.1467-8721.2008.00580.x

Saffran, J., Werker, J., and Werner, L. (2007). "The infant's auditory world: hearing, speech, and the beginnings of language," in *Handbook of Child Psychology, Cognition, Perception, and Language*, eds W. Damon, R. Lerner, D. Kuhn, and R. Siegler (Hoboken, NJ: Wiley), 59–108.

Schmale, R., Cristia, A., Seidl, A., and Johnson, E. K. (2010). Developmental changes in infants ability to cope with dialect variation in word recognition. *Infancy* 15, 650–662. doi: 10.1111/j.1532-7078.2010.00032.x

Schmale, R., and Seidl, A. (2009). Accommodating variability in voice and foreign accent: flexibility of early word representations. *Dev. Sci.* 12, 583–601. doi: 10.1111/j.1467-7687.2009.00809.x

Singh, L., Morgan, J. L., and White, K. S. (2004). Preference and processing: the role of speech affect in early spoken word recognition. *J. Mem. Lang.* 51, 173–189. doi: 10.1016/j.jml.2004.04.004

Van hamme, H. (2008). "HAC-models: a novel approach to continuous speech recognition," in *Proceedings Interspeech* (Brisbane), 2554–2557.

Van hamme, H. (2011). "On the relation between perceptrons and non-negative matrix factorization," in *Signal Processing with Adaptive Sparse Structured Representations Workshop* (Saint-Malo), 119.

van Heugten, M., and Johnson, E. (2012). Infants exposed to fluent natural speech succeed at cross-gender word recognition. *J. Speech Lang. Hear. Res.* 55, 554–560. doi: 10.1044/1092-4388(2011/10-0347)

Versteegh, M., and ten Bosch, L. (2013). "Detecting words in speech using linear separability in a bag-of-events vector space," in *Proceedings Interspeech* (Lyon).

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 31 May 2013; accepted: 08 September 2013; published online: 04 October 2013.*

*Citation: Bergmann C, ten Bosch L, Fikkert P and Boves L (2013) A computational model to investigate assumptions in the headturn preference procedure. Front. Psychol. **4**:676. doi: 10.3389/fpsyg.2013.00676*

*This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 22 October 2013
doi: 10.3389/fpsyg.2013.00767

# Experience and generalization in a connectionist model of Mandarin Chinese relative clause processing

***Yaling Hsiao and Maryellen C. MacDonald****

*Department of Psychology, University of Wisconsin-Madison, WI, USA*

***Edited by:***
*Franklin Chang, University of Liverpool, UK*

***Reviewed by:***
*Hartmut Fitz, Max Planck Institute for Psycholinguistics, Netherlands*
*Yasuhiro Shirai, University of Pittsburgh, USA*

***Correspondence:**
*Maryellen C. MacDonald, Language and Cognitive Neuroscience Lab, Department of Psychology, University of Wisconsin-Madison, 1202 W. Johnson St., Madison, WI 53705, USA*
*e-mail: mcmacdonald@wisc.edu*

Sentences containing relative clauses are well known to be difficult to comprehend, and they have long been an arena in which to investigate the role of working memory in language comprehension. However, recent work has suggested that relative clause processing is better described by ambiguity resolution processes than by limits on extrinsic working memory. We investigated these alternative views with a Simple Recurrent Network (SRN) model of relative clause processing in Mandarin Chinese, which has a unique pattern of word order across main and relative clauses and which has yielded mixed results in human comprehension studies. To assess the model's ability to generalize from similar sentence structures, and to observe effects of ambiguity through the sentence, we trained the model on several different sentence types, based on a detailed corpus analysis of Mandarin relative clauses and simple sentences, coded to include patterns of noun animacy in the various structures. The model was evaluated on 16 different relative clause subtypes. Its performance corresponded well to human reading times, including effects previously attributed to working memory overflow. The model's performance across a wide variety of sentence types suggested that the seemingly inconsistent results in some prior empirical studies stemmed from failures to consider the full range of sentence types in empirical studies. Crucially, sentence difficulty for the model was not simply a reflection of sentence frequency in the training set; the model generalized from similar sentences and showed high error rates at points of ambiguity. The results suggest that SRNs are a powerful tool to examine the complicated constraint-satisfaction process of sentence comprehension, and that understanding comprehension of specific structures must include consideration of experiences with other similar structures in the language.

**Keywords: Simple Recurrent Networks, relative clauses, sentence processing, Mandarin Chinese, working memory, connectionism**

## INTRODUCTION

Sentence comprehension is generally considered to be a complex constraint satisfaction process integrating probabilistic information from syntactic, semantic, prosodic, and discourse sources (e.g., MacDonald et al., 1994; Tanenhaus and Trueswell, 1995). This emphasis on multiple probabilistic constraints in sentence comprehension demands precise accounts of how constraints of different types and different strengths are weighed, so as to yield clearly testable models of comprehension. Unfortunately, compared to a large number of empirical studies in sentence comprehension, there are relatively few implemented computational models of sentence processing phenomena, which could illuminate the interaction of complex probabilistic constraints in sentence comprehension.

One issue that has been addressed in computational models of sentence comprehension is the role of computational capacity in accounts of human comprehension behavior. In particular, a key question in comprehension research is the separability between linguistic knowledge and the capacity to use that knowledge in comprehension and other language behavior. These issues are familiar in the competence-performance distinction that has traditionally distinguished much of linguistic and psycholinguistic research (e.g., Miller and Chomsky, 1963), but it also arises within psycholinguistic accounts of complex sentence comprehension—how much is human comprehension difficulty attributable to limitations on human working memory capacity, independent of people's experience with language? For example, several different accounts have been offered for the difference in comprehension difficulty for subject relative clauses and object relative clauses in English and many other languages. An example can be seen in sentences (1a–b), where the object relatives (1b) are generally found to be more difficult than subject relatives (1a).

(1a) Subject Relative Clause: The candidate [**who$_1$ attacked$_1$ the opponent**] won this election.
(1b) Object Relative Clause: The candidate [**who$_1$ the opponent attacked$_1$**] won this election.

A common argument for the difference in comprehension difficulty between these two sentence types points to the role of working memory, that the object relatives (1b) place higher working memory demands on the comprehender than do the subject

relatives (e.g., King and Just, 1991). In one variant of this view, the Dependency Locality Theory (Gibson, 1998), the working memory demands are tied to greater distance between related elements (shown with subscripts in 1a–b) in object relatives (1b) compared to subject relatives (1a). This additional distance in (1b) requires longer memory maintenance of the partially processed information ("the candidate") until it can be integrated with the action ("attacked"). Failure to maintain or retrieve the information disrupts comprehension. Thus, on this view, comprehension difficulty is directly tied to the capacity to maintain discontinuous elements during sentence comprehension.

These questions have also been addressed in computational models of sentence processing. Simple Recurrent Network (SRN) models of sentence comprehension have a computational capacity that is inherently tied to the model's experience with linguistic input (MacDonald and Christiansen, 2002; this is true of connectionist models more generally, e.g., McClelland and Elman, 1986). As originally implemented by Elman (1990), an SRN is a partially recurrent network equipped with an additional context layer that stores the output of the hidden layer, which is paired with the next input to the network. The model's task is to predict the next input word, and the presence of the context layer permits the prior linguistic context to influence the prediction of later words. SRNs have been applied to several different types of complex sentences (e.g., Elman, 1991, 1993; Christiansen, 1994; Christiansen and Chater, 1999), including the contrast in difficulty between subject and object relatives, as in (1a–b). Direct comparisons of human and model performance reveal important similarities. **Figure 1** shows MacDonald and Christiansen's (2002) SRN and Wells et al's. (2009) human subjects' reading times for subject and object relatives in studies in which experience with the two structures was explicitly manipulated in the model and in the human readers. MacDonald ad Christiansen's model had extensive experience with several kinds of simple sentences, while only 5% of the sentences in the training set contained relative clauses. Their model was tested at three different points in training to investigate a claim about how these models generalize from the common simple sentences, such as *The candidate attacked the opponent*, to relative clauses such as (1a–b). They hypothesized that interpretation of subject relatives would be aided by these sentences' similarities to simple sentences, and as a result, the model would rapidly learn to make accurate predictions for subject relatives, and it would show little effect of additional training, as it had already benefited from the overlap with the common simple sentence "neighbors." By contrast, object relatives have an idiosyncratic word order that is not aided by extensive experience with simple sentences, and MacDonald and Christiansen predicted that as a result, the model would be extremely sensitive to the degree of direct experience with object relatives. **Figure 1** shows that these predictions were supported, and it also shows that human reading times in the Wells et al. study were similarly influenced by a training manipulation, in which additional exposure to subject and object relatives over the course of a month affected participants reading patterns for object relatives (right panel) but not subject relatives (left panel).

An important feature of SRNs is that they generalize over their training experiences with individual sentences to find regularities across the training items (Elman, 1990; St John and McClelland, 1990). The results in **Figure 1** reflect generalization in MacDonald and Christiansen's SRN—in that object and subject relatives were presented equally often but were not equally difficult, because the model generalized from simple transitive sentences to the similar subject relatives. Several other researchers have pursued this point, including Fitz et al. (2011), who used a dual-path SRN that models both sentence production and sentence semantics. Although these models are not meant to be accounts of the acquisition process, similar generalization effects have been found in child language acquisition. For example, Yip and Matthews (2007) found that the relative clauses that emerged earlier in Cantonese (object relatives) do so because of their word order resemblance to the dominant word order in simple sentences (for similar effects in other languages, see Diessel, 2004, 2007; Diessel and Tomasello, 2005; Ozeki and Shirai, 2007).

These results hold promise for SRN accounts of sentence comprehension, but to date, their application has been quite limited, and in particular the models have not typically incorporated realistic frequencies of various sentence types, which would better allow researchers to understand how complex probabilistic constraints are weighed in comprehension. In this paper, we take steps toward meeting this challenge with an SRN model of relative clause comprehension that accurately represents critical elements of the distributional knowledge that humans bring to bear in

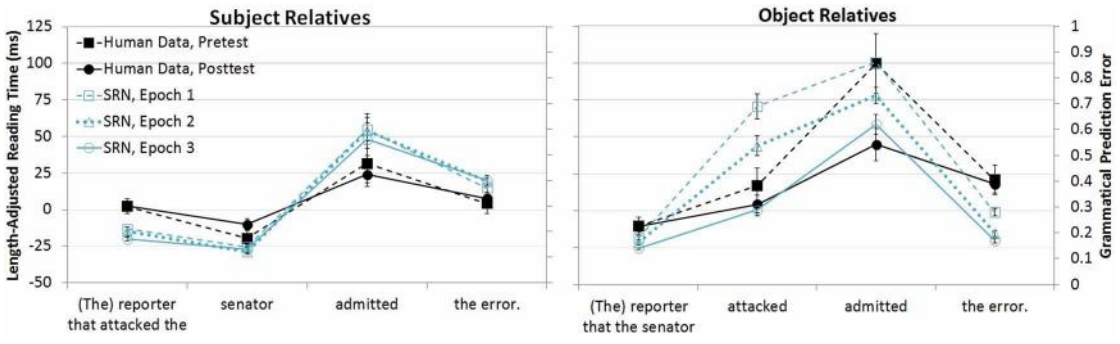



**FIGURE 1 | Comparison of SRN performance in MacDonald and Christiansen (2002) and human reading times in Wells et al. (2009).**

interpreting these structures. Rather than elaborating accounts of English relative clause comprehension, about which an enormous body of research exists, we address relative clause comprehension in Mandarin Chinese. As we detail below, Mandarin relative clause comprehension is particularly interesting because (a) the relative clause structure is quite different from English, (b) some key findings about comprehension difficulty show potentially opposite patterns than in English and in many other languages, (c) there is a fair amount of controversy concerning comprehenders' performance in various structures and discourse contexts, and (d) the number of important factors affecting comprehension appears to be too large to be manipulated together in empirical studies. Thus, a computational account of Mandarin relative clause processing has the opportunity to have a substantial impact informing the nature of constraint satisfaction processes in Mandarin and more generally.

## MANDARIN RELATIVE CLAUSES

Relative clauses in English and many other languages are "head-first," meaning that the relative clause appears after the head noun, as in *the candidate [that attacked the opponent]*. In many of such languages, comprehenders show a clear pattern of finding subject relatives easier than object relatives (English: e.g., King and Just, 1991; Gibson et al., 2005; Dutch: e.g., Frazier, 1987; German: e.g., Schriefers et al., 1995; French: e.g., Frauenfelder et al., 1980). Other languages, including Korean, Japanese, and Mandarin, have a "head-final" relative clause structure, such that the relative clause precedes the head noun, as in the Mandarin examples in (2). In these examples, in which the relative clause modifies the subject of the main clause, the relative clause begins the sentence, ending with the relativizer DE (equivalent to the English *that* in this context), followed by the noun phrase (head) being modified, in this case *candidate*. This head-final relative clause word order is a critical factor in accounts of cross-linguistic differences in relative clause processing, because the different relative clause structures create different degrees of distance between dependent elements, as shown by subscripts in (2).

(2a) Subject-modifying subject relative:
[攻擊 競爭 者 的] 候選人$_1$ 贏了 這場 選舉
[$e_1$ attack opponent DE] candidate$_1$ won this election
The candidate who attacked the opponent won this election.

(2b) Subject-modifying object relative:
[競爭者 攻擊 的] 候選人$_1$ 贏了 這場 選舉
[opponent attack $e_1$ DE] candidate$_1$ won this election
The candidate who the opponent attacked won this election.

Most studies of relative clause processing in Korean and Japanese have suggested that subject relatives are easier than object relatives (Japanese: e.g., Miyamoto and Nakamura, 2003; Korean: Kwon et al., 2010), similar to English and other head-first languages. Some researchers suggest that this pattern points to a universal subject preference for relative clause processing, as first proposed by Keenan and Comrie under the term "Accessibility Hierarchy" (1977), which argues that noun phrases at the subject position are the easiest to be relativized due to their higher syntactic position, compared to noun phrases lower in the hierarchy, such as genitive (though cf. Ozeki and Shirai, 2007; Yip and Matthews, 2007). However, the pattern of subject preference does not appear to apply to Mandarin, which differs in several ways from the other languages considered here. Mandarin has the head-final relative clause structure, like Japanese and Korean, but whereas Japanese and Korean have a rich system of case marking on nouns that presumably aids relative clause processing, case-marking is non-existent in Mandarin. Mandarin is also different in that it has the dominant word order of subject-verb-object (SVO) in main clauses, like English and many European languages, but with the head-final relative clause structure that is absent in these languages. This combination of SVO basic word order, head final relative clause structure and absence of case marking is attested in the world's languages only in four Sino-Tibetan languages such as Bai, and other Chinese languages like Mandarin, Cantonese, and Hakka (Keenan, 1985; Haspelmath et al., 2005).

These features make Mandarin an interesting test case to disentangle the influences of various potential factors in relative clause processing. To date, empirical studies have yielded conflicting results, with some studies finding the typical cross-linguistic pattern of subject relatives being easier than object relatives (e.g., Lin and Bever, 2006), and others finding the opposite result (e.g., Hsiao and Gibson, 2003). This reversal of the dominant cross-linguistic pattern finds a clear interpretation in the Dependency Locality account. Hsiao and Gibson argued that subject relatives (e.g., 2a) were more difficult than object relatives (e.g., 2b) because they have higher storage and integration costs, owing to the longer distance between dependencies in subject relatives: there are more intervening words between the filler and the gap and thus more new discourse referents and incomplete dependencies in a subject relative than in an object relative. Support for this memory-based view comes from a study in which the added difficulty of subject relatives was higher in participants with lower working memory span (Chen et al., 2008) and patients with aphasia (Su et al., 2007). Several studies manipulating other factors, such as relative clause topicalization (Lin and Garnsey, 2011) and context (Gibson and Wu, 2013), also found a similar object relative advantage.

However, a reading time advantage for subject relatives has also been reported in several studies. Vasishth et al. (2013) modified Hsiao and Gibson (2003)'s materials and failed to replicate their effect, instead finding that object relatives were harder than subject relatives. Lin and Bever (2011) found no difference between the two types of relative clauses modifying the main clause subject (as in 2a–b) but found shorter reading times for subject relatives than object relatives when the relative clause was modifying the main clause object, such that subject relatives like (3a) were easier than object relatives like (3b). They also manipulated whether the participants were informed that they were reading relative clauses and about which noun positions they were modifying. They found that participants who were not informed had the most difficulty reading object-modifying ORCs, whose word order in combination with the matrix clause (i.e., the first three

words in 3b) could lead the readers into a garden path of a simple sentence.

(3a) Object-modifying subject relative:
選民 支持 [ 攻擊 競爭者 的 ] 候選人
voter support [$e_1$ attack opponent DE] candidate$_1$
Voters support the candidate who attacked the opponent.

(3b) Object-modifying object relative:
選民 支持 [競爭者 攻擊 的 ] 候選人
voter support [opponent attack $e_1$ DE] candidate$_1$
Voters support the candidate who the opponent attacked.

Still other studies investigated the effect of lexical semantics and found it to have a modulating influence on the difficulty of the two relative clause types. Wu et al. (2012) manipulated the animacy of both the head noun and the relative clause noun and found that comprehension difficulty in both sentence types was strongly dependent on whether the sentence contained a canonical animacy configuration, in which animate entities acted on inanimate ones. These results resonate well with the Mak et al. (2002, 2006) and Traxler et al. (2002, 2005) studies in other languages, in that animacy serves an important cue for thematic role assignment, which in turn affects ambiguity resolution (Gennari and MacDonald, 2008).

More generally, these results, together with absence of case marking and head-final relative clause structure in Mandarin, suggest that difficulty in relative clause comprehension may be strongly modulated by temporary ambiguities in sentences, where comprehenders initially interpret nouns and verbs in the input as being part of a main clause only to realize later that they had encountered a relative clause.

Summarizing over these various studies of Mandarin relative clause processing, it is impossible to draw conclusions about overall comprehension difficulty of the two sentence types. The inconsistency across studies and the sensitivity of the results to multiple factors such as animacy and modifying position may suggest that comprehenders are able to use very detailed information about patterns of relative clause use in comprehending these structures. Thus, while many prior studies argued for one structure being absolutely easier than the other, a closer look at the materials in these studies suggests researchers' conclusions are limited by their methodological choices and stimulus items. For example, while claims about relative clause processing difficulty are typically phrased to cover all relative clauses of a certain type (such as subject relatives), in fact, most of the previous studies examined only a narrow subset of relative clauses within a given type. For example, most studies have examined only relative clauses that modified main clause subjects, and containing only animate head nouns and relative clause nouns. Such relative clauses are in reality rare in the linguistic environment (Pu, 2007; Wu et al., 2012; Vasishth et al., 2013). As a result, the Mandarin relative clause research, which could in principle be extremely informative about cross-linguistic regularities and differences in complex sentence interpretation, is instead marginalized by a lack of consensus and by overly-broad claims based on a narrow range of stimulus materials.

While is effectively impossible to combine all the potentially important factors in a single empirical study that would allow us to observe complex interactions among them, it is possible to examine many of these effects in an SRN. We conducted our study in two phases, reflecting the fact that an attractive property of SRNs is their strong sensitivity to the distributional patterns in the language, a property that also appears in human comprehension. As described in Study 1, we examined the distributional statistics of relative clauses in a large corpus, extensively hand-coding the corpus data for features we believe to be critical in relative clause comprehension. In Study 2 we used this distributional information to develop a finite state grammar from which we developed training materials that faithfully represented key properties of relative clauses identified in the corpus. We used these training materials to train an SRN that was exposed to several different types of simple sentences and relative clauses. Model performance was then compared to human reading time data from previous studies. Because the corpus analysis and, therefore, the training set, were so detailed, we can compare experiment findings with model performance specifically for the types of sentences used in several empirical studies. In this way, we aim to use the model to help resolve the conflicting findings in the literature and be able to identify broader themes in Mandarin relative clause processing and cross-linguistic differences and commonalities in sentence comprehension more generally.

## STUDY 1—CORPUS ANALYSIS

Humans' comprehension of relative clauses is influenced not only by their prior experience with relative clauses but also by their experience with other sentences in the language. For example, MacDonald and Christiansen (2002) suggested that English speakers' experiences with simple transitive sentences aided comprehenders' interpretation of subject relative clauses, which are similar in that nouns in analogous sentence positions receive the same thematic role assignments. In this sense simple transitive sentences are helpful "neighbors" of subject relatives in English, as experience with these highly frequent simple sentences allows generalization to the rare subject relative structure (see also Fitz et al., 2011). MacDonald and Christiansen found a similar neighborhood effect in their SRN, attributable to the overlap in word order between the two sentence types, as the SRN does not assign thematic roles or interpret sentences. Conversely, humans' prior experiences with other kinds of sentences can increase the difficulty of relative clause comprehension. Gennari and MacDonald (2008) showed that object relative clauses in English contain temporary ambiguities for which the object relative clause is often an infrequent and disfavored interpretation. In this case, experience with other, more frequent sentence types affects the process of ambiguity resolution, leading to difficulty in interpreting object relatives. We will call these alternative interpretations "competitors," recognizing that both neighbor and competitor effects reflect comprehenders' generalizations over their prior linguistic experiences.

Mandarin relative clauses may similarly be affected by competitors and neighbors. First, Mandarin head-final relative clauses

exhibit temporary ambiguities such that the unfolding linguistic input has several alternative interpretations. Second, certain Mandarin relative clauses have highly similar word orders to some more frequent simple sentences; generalization over these common neighbor sentences should help relative clause processing. Note that the same sentence type may serve both competitor and neighbor functions at different points, in that it might provide an alternative interpretation that affects ambiguity resolution early in processing but that following a point of disambiguation, certain overlap with a relative clause may help in eventual relative clause interpretation (see Fitz et al., 2011, for further discussion). Appendix A describes some of the main neighbors and competitors for relative clauses in Mandarin.

To investigate the range of competitors and neighbors of Mandarin relative clauses, which should be informative for both human and computational work, and to develop a realistic training corpus for our model, we conducted a corpus analysis that enabled us to calculate the statistics of various types of relative clauses, crucial competitor structures that could increase comprehension difficulty, and neighboring structures that could reduce comprehension difficulty.

Here it is necessary to introduce some clarifying terminology, because it can be confusing to discuss both subject and object relative clauses (in which the modified head noun is the subject or object of the relative clause verb, respectively) crossed with the main clause subject vs. object modifying positions (in which the relative clause-modified noun is either the subject or object of the main clause; see examples 2a–b and 3a–b above). For the remainder of this paper, we will refer to relative clauses as RCs, and the subject and object relatives as SRCs and ORCs, respectively. We will continue to spell out subject- vs. object-modifying positions in the main clause, so that full word descriptions are associated with main clause factors and acronyms are used for the embedded clauses.

**Table 1** shows the competing and neighbor structures that we investigate in the present study and the ambiguities or facilitation they create at different points of an SRC and an ORC. The cells with gray shading indicate sentence types in the training set for which we do not expect large effects on RC processing in the model. As the table shows, some sentences may serve both competitor and neighbor functions at different points in the sentence.

## METHODS

We used Tgrep2 1.15 (Rohde, 2005) to extract sentences from a parsed corpus of spoken and written language, the Penn Chinese Treebank 7.0 (Xue et al., 2010). The corpus consists of more than one million words in more than 50,000 sentences, with sources from Chinese newswire, broadcast news, magazine news, broadcast conversation programs, web newsgroups, and others. The Tgrep2 search patterns used to extract the sentences are contained in Appendix B. Our aim was to retrieve every subject- and object-modifying ORC, every subject- and object-modifying SRC (with both transitive and intransitive verbs), every single-clause pro-drop construction, and every overt subject single-clause simple sentence in the corpus. A total of 6255 sentences were extracted.

**Table 1 | The competing and neighbor structures (listed on the top) we examined in the current study and the ambiguities and facilitation they created at different points for SRCs and ORCs at the two matrix positions (listed on the left).**

| | | **Overt subject simple sentences: N V {N}** | **Pro-drop simple sentences: V {N}** | **Subject-modifying intransitive SRCs: [V DE] N V . . .** | **Object-modifying intransitive SRCs: N V [V DE] N.** |
|---|---|---|---|---|---|
| Subject-modifying | SRCs: [V N DE] N V . . . | | Competitor: before DE neighbor: after DE (for V N word order) | Neighbor: early (promotes RC interpretation of first V) | |
| | ORCs: [N V DE] N V . . . | Competitor: before DE (interpret the initial N V order as start of simple sentence) neighbor: after DE (similar N V N order) | | | |
| Object-modifying | SRCs: N V [V N DE] N. | | | | Neighbor: early (promotes RC interpretation of first V) |
| | ORCs: N V [N V DE] N. | Competitor: from beginning (interpret RC N as object N of simple sentence) | | | |

*N, noun, V, verb, DE, relativizer, SRC, subject relative clause, ORC, object relative clause, [Square brackets] indicate the relative clause, {Curly brackets} indicate optional direct object N, which is present following transitive verbs and absent following intransitive verbs.*

*Note: Because the model did not include those verbs that can be both transitive and intransitive (e.g., "I ate" vs. "I ate an apple"), intransitive SRCs were not listed as a competitor to transitive SRCs.*

We limited extraction of simple sentences to those with only a single clause, because our model was not designed to process other types of multi-clause sentences. However, we extracted all RCs that modified a sentential subject or object regardless of whether the rest of the sentence was a simple main clause or a more complex sentence. We cast this wider net for RCs because they are fairly rare in Mandarin, and extracting literally every RC in the corpus would give us the best estimate of the relative frequency of different RC types and animacy configurations. These methodological choices yielded a higher ratio of RC sentences to non-RC sentences than in the whole corpus (because many other non-RC multi-clause sentences were not extracted). Nonetheless, simple sentences outnumbered RCs at about 5:1 in the extracted set of 6255 sentences.

For each sentence, noun animacy was coded by hand, as NP animacy is known to affect Mandarin RC processing (Wu et al., 2012), and we expected that the animacy configuration of simple sentences could also serve as important experience for RC processing. Coding followed the criteria that animate nouns refer to living entities that possess agency and volition to perform an action, while inanimate nouns refer entities without these properties (Hundt, 2004). For example, plants are living things but do not have volition, and thus they were coded as inanimate. Nouns that represented a group of people, such as organizations, countries, etc. (e.g., *the school that taught me math*), were considered animate. However, when these nouns were used purely as locations (e.g., *the school that I went to*), they were coded as inanimate. Coding was performed by two native speakers of Mandarin who were instructed in these coding criteria.

## RESULTS

**Table 2** reports the frequencies of all simple sentences that were coded, and **Table 3** reports all relative clauses. Inspection of these tables reveals the following patterns:

(1) Simple sentences with an overt subject (a potentially helpful neighbor of ORCs) ($N = 3309$) were more frequent than pro-drop simple sentences (a competitor for some RC interpretations, $N = 1885$), and sentences with RCs ($N = 1061$) were less frequent than either of these simple sentence types. Thus, both some competitor interpretations and helpful neighbor interpretations are more frequent than RCs themselves.

(2) The majority of (overt) main clause subjects were animate and the majority of main clause objects were inanimate. As main clauses are more common than RCs, these patterns of main clause animacy may influence expectations for RC animacy configurations.

(3) There were relatively more subject-modifying RCs ($N = 636$, 60% of all RCs) than object-modifying RCs ($N = 415$). SRCs (transitive and intransitive combined) occurred more often than ORCs in both modifying positions. Transitive SRCs were the most frequent among the three types of RCs.

(4) ORCs, regardless of modifying position, mostly had inanimate head nouns and animate RC embedded nouns, consistent with patterns observed in English (Roland et al., 2007). Transitive SRCs had a higher proportion of animate heads and inanimate RC embedded nouns in the subject-modifying position, whereas there was not a big difference in head animacy but a preference of inanimate RC embedded nouns in the object-modifying position. Intransitive SRCs at both modifying positions had more inanimate heads than animate heads. These relatively high rates of inanimate head nouns, which generally followed the patterns of animacy usage in main clauses, could be attributed to the high percentage of sentences like "The growth rate rose," in the newswire genre, which comprises a large subset of the Chinese Treebank 7.0 corpus.

**Table 2 | Token frequencies of overt subject simple sentences and pro-drop sentences found in Chinese Treebank 7.0.**

| | Simple sentences with overt grammatical subject | | | | | |
|---|---|---|---|---|---|---|
| | Animate subject nouns | | | Inanimate subject nouns | | |
| | Object noun type | | | Object noun type | | |
| **Verb Type** | **Animate** | **Inanimate** | **None** | **Animate** | **Inanimate** | **None** |
| Intransitive | | | 162 | | | 702 |
| Transitive | 355 | 1477 | | 121 | 492 | |

| | Pro-drop sentences (with grammatical subject omitted) | | |
|---|---|---|---|
| | Object noun type | | |
| **Type** | **Animate** | **Inanimate** | **None** |
| Intransitive | | | 295 |
| Transitive | 539 | 1051 | |

*Gray cells mark non-existent combinations of verb transitivity and object type; for example, intransitive verbs by definition have no direct objects, and so cells representing the animacy coding of objects are not relevant for intransitive verbs.*

**Table 3 | Token frequencies of subject- and object-modifying SRCs (transitive and intransitive) and ORCs at found in Chinese Treebank 7.0**

| | Animate head nouns | | | Inanimate head nouns | | |
|---|---|---|---|---|---|---|
| | Relative clause (RC) noun type | | | Relative clause (RC) noun type | | |
| **RC-type** | **Animate** | **Inanimate** | **None** | **Animate** | **Inanimate** | **None** |
| **SUBJECT-MODIFYING RCs** | | | | | | |
| SRC, Intransitive | | | 14 | | | 56 |
| SRC, Transitive | 61 | 209 | | 68 | 31 | |
| ORC | 11 | 0 | | 163 | 23 | |
| **OBJECT-MODIFYING RCs** | | | | | | |
| SRC, Intransitive | | | 34 | | | 74 |
| SRC, Transitive | 30 | 62 | | 27 | 76 | |
| ORC | 10 | 0 | | 79 | 23 | |

*Relative Clause Nouns are the Relative clause object in SRC Transitive sentences and the Relative clause subject in ORC sentences. Gray cells reflect non-existent sentence types (because transitive verbs must have a relative clause noun and intransitive verbs do not have a direct object noun).*

## DISCUSSION

This corpus analysis yielded several important patterns. First, the configuration of animacy in the corpus, in both main clauses and relative clauses, is strikingly similar to findings in other languages and also replicates and extends previous Mandarin corpus studies (Pu, 2007; Wu et al., 2012). Key results here include the tendency for main clause subjects to be animate and objects to be inanimate, the tendency of SRC heads to be animate and ORC heads to be inanimate, and the rarity of RCs with two nouns of the same animacy (both animate or both inanimate, Wu et al., 2012). The general similarity of these patterns to main and relative clause usage in other languages (e.g., Bock and Warren, 1985; Roland et al., 2007; Gennari and MacDonald, 2009) suggests that the difference in patterns of relative clause interpretation in Mandarin vs. other languages does not lie in different animacy configurations and must instead reflect other critical cross-linguistic differences. Candidates for other important cross-linguistic differences include the very high degree of temporary ambiguity encountered during Mandarin RC processing, owing to the combination of head-final RCs and the absence of case marking, and also particular patterns of potentially helpful neighboring structures. The complexity of the potential interactions here is quite large, and in the next study, we use a SRN to explore the effect of these linguistic patterns on relative clause processing.

## STUDY 2—SIMPLE RECURRENT NETWORK

We chose an SRN to model Mandarin RC processing because of its potential to simulate word-by-word human reading times, including in prior studies of relative clause processing (MacDonald and Christiansen, 2002; Wells et al., 2009; Fitz et al., 2011). We trained the model on a mix of relative clauses, helpful neighbor sentences, and competitor sentences (contributing to temporary ambiguities) in proportions based on the Study 1 corpus analysis to investigate how these varied experiences could jointly contribute to relative clause processing. It is important to note that SRNs are not simulating human language comprehension *per se* but are instead a simulation of a component thought to be a part of comprehension and a factor related to reading times, namely prediction of upcoming input. Thus, while terms such as "ambiguity resolution," "alternative interpretations," and "garden-path effect" are common in descriptions of human reading times, the model is not adopting any interpretation or calculating meaning. Ambiguity created by the existence of multiple interpretations in human sentence processing is more adequately termed as "conditional indeterminacy" in the case of an SRN. As the probabilities of grammatical continuations are more varied, the higher the prediction error is for the model. Thus, for both model and humans, indeterminacy is costly, and the model is taken to represent one important aspect of the ways in which input can be ambiguous for humans.

## METHODS

The SRN used the backpropagation learning algorithm. It contained a context layer in addition to input, output and hidden layers. These layers were connected by trainable weights, except that the context layer directly copied the activations of the hidden layer from the previous time tick. The input pattern (the words of a sentence) is activated in the input layer one word at a time and then propagated onto the hidden layer and the output layer at time step $t$. The weights are adjusted by comparing the output activations to the desired output. At the next time step $t + 1$, the output activation on the hidden layer at time step $t$ is copied to the context layer and then projected backed to the hidden layer to pair with the current input.

### *Model specification*

As shown in **Figure 2**, the model contained 28 localist units in the input and output layer, representing 27 words, plus one "end-of-sentence" marker. Although the network did not model semantics, animacy was captured distributionally, meaning that units designating "animate" nouns appeared frequently as sentence subjects and rarely as objects, while units designating inanimate nouns had the opposite pattern. The grammatical categories included in this model were the following: Animate Nouns (7 units), Inanimate Nouns (7 units), Transitive Verbs (6 units), Intransitive Verbs (6 units), RC marker DE (1 unit), and End-of-Sentence marker (1 unit). There were 40 units each in the hidden layer and the context layer. The learning rate was set to 0.05, with momentum of 0.9 and the batch size of 1. The simulation was conducted with the software Lens (Version 2.63) (Rohde, 1999).

### *Training*

Based on the frequency information obtained from Study 1, we calculated the bigram transitional probabilities from one grammatical category to another, as shown in **Table 3**. Transitional probabilities were calculated from the probability of occurrence of Y given the previous input X. For example, given the occurrence of a transitive verb (VT), the next word was an animate noun (aniN) 18% of the time, an inanimate noun (inaN) 73% of the time and the start of an object modifying relative clause (objRC) 9% of the time. Similarly, the probability of the occurrence of a subject-modifying SRC with an animate head and intransitive verb is $0.7 \times 0.16 \times 0.11 \times 0.2 = 0.0025$, meaning that there are ~25 sentences that contain this type of RC in the 10,000-sentence training corpus.

The model's training faithfully reflected the relative frequencies of RCs found in the corpus, but as with all computational models, it is a simplification of the knowledge that humans bring to the task. These simplifications include the absence of semantics, the limited vocabulary, the omission of other uses of DE

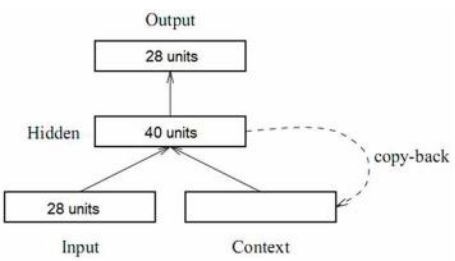



**FIGURE 2 | Architecture of the Simple Recurrent Network used in Study 2.**

(see Appendix A), and having a higher proportion of RCs in the training set than in the corpus as a whole. Some of these features (RC frequency, definitive disambiguation at DE) may make RCs proportionally less difficult for the model than for humans, and other features (absence of fine grained semantics or real world or discourse context, smaller neighborhood effects) could make RCs yield proportionally more error for the model than would be expected based on human reading times. Given that all models necessarily include simplifications, the choices here represent a good starting point with which to examine Mandarin RC processing.

A python script was written based on the finite state grammar with the transitional probabilities in **Table 4** to produce the 10,000 training sentences. The grammar did not permit sentences with multiple RC embeddings because multiply embedded RCs are very rare in the human corpus. The sentences encompassed the following structures: transitive/intransitive simple sentences with/without subject, intransitive/transitive subject relative clauses, and transitive object relative clauses. The three types of relative clauses (SRCs with transitive and intransitive verbs and ORCs) could modify either main clause subjects or objects. The script selected words within a given category at random, with equal probability for each word. As with other SRNs, there were very few word units in the model, and all words in a category (such as animate nouns, transitive verbs) were equally frequent as the other words in the category.

### *Model assessment*

The model performance was evaluated via Grammatical Prediction Error (GPE), which has been shown to relate well to behavioral measures such as reading times and grammaticality judgments (Christiansen and Chater, 1999). The measure is based on the idea that, because the current output activation is generated by the model reflecting the context in the previous time ticks, the model should activate what is expected/grammatical and should not activate what is ungrammatical, as defined by the training corpus. The GPE therefore incorporates the concepts of "hits" (sum of activation in grammatical units) and "false alarms" (sum of activation in ungrammatical units), with "misses" (sum of insufficient activation of grammatical units), as described in Christiansen and Chater (1999) and shown below.

$$\text{GPE} = 1 - \frac{\text{hits}}{\text{hits} + \text{false alarms} + \text{misses}}$$

The misses are derived from the difference between the actual output activation from the target activation, based on the transitional probabilities in the training corpus. For example, the target activation of the units representing VT (a transitive verb) at the sentence initial position should sum up to about 0.3 because that value is the sum of all the sentence types with a sentence-initial VT: first, there is 30% chance that the sentence starts as a pro-drop sentence, which is in turn 79% likely to be composed of a transitive verb with an object noun phrase, yielding $0.3 \times 0.79 = 0.24$. Second, a sentence-initial VT may start an SRC, with a probability of $0.7 \times 0.16 \times 0.58 = 0.065$. Combining the probabilities of the two situations ($0.24 + 0.065$), we should obtain a target activation for the sentence-initial VT of around 0.3. If in this example the sum of the total output activation in the VT units is actually 0.26, then the misses should be 0.04. Inside the sentence, the target activation distribution was affected by prior context. For example, following the sentence-initial VT, grammatical continuations were aniN (animate noun), inaN (inanimate noun), VT, and VI (intransitive verb). If the next word was an aniN, the VT+aniN fragment could be the start of a) a pro-drop simple sentence with an animate patient ($0.3 \times 0.79 \times 0.18 = 0.043$) or b) a subject-modifying SRC ($0.7 \times 0.16 \times 0.58 \times 0.35 = 0.023$). The total probability of VT + aniN occurring in the whole training corpus was $0.043 + 0.023 = 0.066$ but the probability of aniN following VT should be weighted among all four possible continuations. That is, summing up the probabilities of sentence-initial VT + aniN (0.066), VT + inaN ($0.3 \times 0.79 \times 0.73 + 0.7 \times 0.16 \times 0.58 \times 0.65 = 0.215$), VT + VT ($0.3 \times 0.79 \times 0.09 \times 0.58 = 0.012$), VT

**Table 4 | Finite state grammar with corpus-based bigram transitional probabilities.**

| **S → subNP + VP (0.7) / VP (0.3) (meaning that 70% of sentences had a subject NP and 30% were pro-drop sentences, without a subject NP)** | |
|---|---|
| **VP → VI (0.21) / VT + objNP (0.79) (i.e., 21% intransitive verbs and 79% transitive verbs with direct object NPs)** | |
| **subNP → aniN(0.5) / inaN (0.34) / subRC (0.16)** | **objNP → aniN(0.18) / inaN (0.73) / objRC (0.09)** |
| subRC (modifying matrix subject) | objRC (modifying matrix object) |
| → SRC_VI (0.11): | → SRC_VI (0.26): |
| VI + DE + aniN(0.2)/ inaN(0.8) | VI + DE + aniN(0.31)/ inaN(0.69) |
| → SRC (0.58): | → SRC (0.47): |
| (0.35) VT + aniN + DE + aniN(0.47)/ inaN(0.53) | (0.29) VT + aniN + DE + aniN(0.53)/ inaN(0.47) |
| (0.65) VT + inaN + DE + aniN(0.87)/ inaN(0.13) | (0.71) VT + inaN + DE + aniN(0.45)/ inaN(0.55) |
| → ORC (0.31): | → ORC (0.27): |
| (0.88) aniN + VT + DE + aniN(0.06)/ inaN(0.94) | (0.80) aniN + VT + DE + aniN(0.11)/ inaN(0.89) |
| (0.12) inaN + VT + DE + aniN(0)/ inaN(1) | (0.20) inaN + VT + DE + aniN(0)/ inaN(1) |

*S, sentence; NP, noun phrase; VP, verb phrase; VI, intransitive verb; VT, transitive verb; subNP, subject noun phrase; objNP, object noun phrase; aniN, animate noun; inaN, inanimate noun; subRC, subject-modifying relative clause; objRC, object-modifying relative clause; SRC_VI, subject relative clause with intransitive verb; SRC, subject relative clause with transitive verb; ORC, object relative clause; DE, relative clause marker.*

+ VI ($0.3 \times 0.79 \times 0.09 \times 0.11 = 0.002$) in the corpus, the total probability of sentences starting with an VT should be around 0.30, as calculated above. Among all four legal continuations, the relative proportion of aniN appearing after VT was $0.066/0.30 = 0.22$, which should be the summed target activation for all the units representing aniN. The same procedure applied to the other continutations.

It should be noted that overestimation was also implicitly penalized due to the fact that dislocated activation found in some units means insufficient activation in some other units because the total output activation sums to around 1.

The GPE ranges from 0 to 1, with 0 being perfectly accurate in predicting the grammatical categories of the current word based on prior context (that is, 0 error) and 1 being completely incorrect in doing so.

GPEs reflect the prediction error of the next word based on the cumulative context (e.g., the GPE of Word 3 is directly affected by Word 2, which in turn is affected by Word 1), and therefore may implicitly simulate the spillover effect observed in human reading patterns, where reading of one word may be affected by properties of prior input.

### *Testing*

Ten networks with different random initial weights, ranging from 1 to -1 with the mean of 0, were created and trained on the 10,000 sentences in the training corpus. These 10 networks simulated the role of "participants" in empirical studies, each of whom may have come from different backgrounds with varying prior experiences and skills. The models were trained on one pass through the training set and were tested using novel test sentences (sentences not contained in the training set) that allowed us to assess the major types of relative clauses that have been investigated in the comprehension literature. There were 16 different types of test sentences, each with 10 tokens, for a total of 160 test sentences. The 16 sentence types were obtained from crossing factors to yield the relative clause types shown in **Table 3**: two modification positions (main clause subject, object) x relative clause type (SRC vs. ORC) × head noun animacy (animate, inanimate) and RC noun animacy (animate, inanimate). Intransitive SRCs (also shown in **Table 3**) were not included in the test set because there are no human reading time data in the literature. The GPE scores were calculated at each word in the critical RC region, the head noun, and the word after the head noun.

## RESULTS

In the sections below, we first present results of the model performance at each modifying position (main clause subject or object), noting the general relationships to human data. Then in later sections we present a detailed comparison between the model GPEs and the results of specific experiments. Statistical analyses were conducted with mixed effect models with maximal random effects of participants and items using the lme4 packages in R. Significance values were estimated by likelihood ratio test as suggested in Barr et al. (2013).

### *Subject-modifying RCs*

**Figure 3** presents the mean GPEs of SRCs and ORCs at each of five word positions comprising a sentence-initial relative clause, a head noun, and the next word. Statistical analyses were conducted to examine three effects—RC type, RC noun animacy, and head noun animacy—along with their interactions. Note that there are no effects of head noun animacy through the first three words, because this factor does not appear in the sentence until the head noun is reached at the fourth word position. The head animacy effects also illustrate a crucial effect of GPE calculations. Before the head noun was encountered, only the effects of RC type, embedded noun animacy, and their interaction were factored into the analyses from Word 1 to DE. The effect of head animacy and its interactions with other two factors were considered only at the head and head + 1 positions.

The most important result for these sentences is the contrast in difficulty at the head position for SRCs (left graph) and ORCs (on the right). As we noted in **Tables 2**, **3** SRCs modifying a main clause subject are more ambiguous than ORCs: a sentence-initial verb might be the start of an SRC, but competitor pro-drop structures (simple sentences without an overt grammatical subject) are

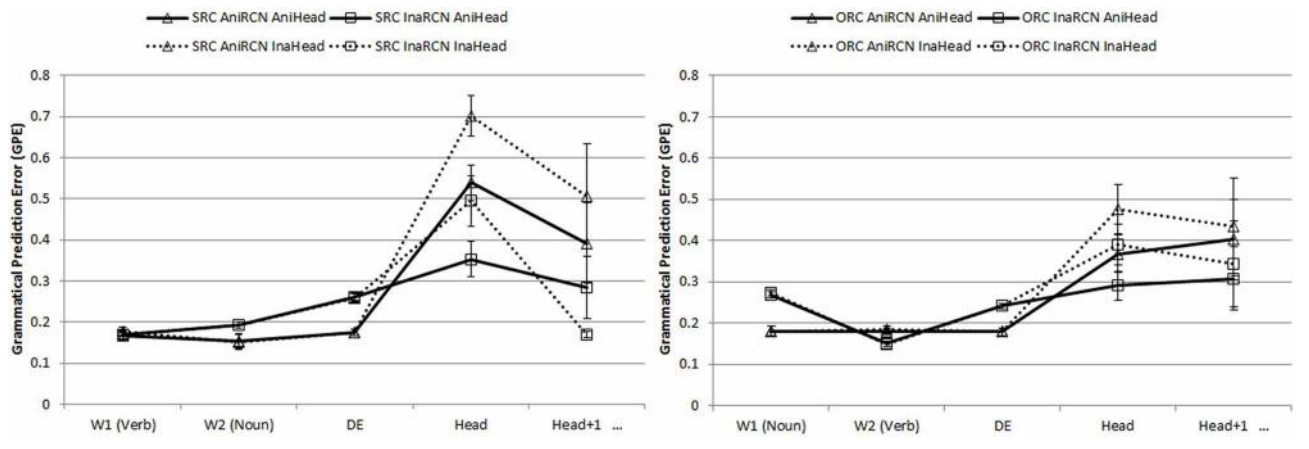



**FIGURE 3 | Word-by-word GPEs of subject-modifying SRCs (left panel) and ORCs (right panel) in the critical RC region, the head, and next word after the head.**

much more common. Sentence-initial object relatives are comparatively less ambiguous, and the effects these different amounts of ambiguity are clear in the model, with reliably higher GPEs for subject relatives than object relatives at the head position (detailed analyses at each word position are given below). Thus, even though sentence-initial SRCs with transitive verbs are more frequent than ORCs at a ratio of almost 2:1, SRCs are substantially harder for the model. This result replicates major findings in the comprehension literature (Hsiao and Gibson, 2003; Gibson and Wu, 2013) and shows how the model's behavior on a particular structure is not simply a reflection of the frequency of that structure. We will elaborate these points after describing the effects at each word position.

***Word-by-word analyses.*** At W1 (RC object for SRCs, RC verb for ORCs), both main effects of RC type ($\beta = -0.009$, $SE = 0.003$, $t = -3.08$, $p = 0.002$) and animacy of the RC noun ($\beta = 0.090$, $SE = 0.008$, $t = 11.69$, $p < 0.001$) were significant. SRCs were easier than ORCs, and RCs with animate embedded nouns were easier than those with inanimate embedded nouns. An RC type x animacy of RC noun interaction was present ($\beta = -0.093$, $SE = 0.008$, $t = -11.24$, $p < 0.001$), primarily because ORCs with inanimate RC nouns were the hardest among all conditions.

At W2, the RC type main effect showed an SRC advantage ($\beta = -0.031$, $SE = 0.013$, $t = -2.34$, $p < 0.001$). The main effect of embedded noun animacy was such that RCs with inanimate embedded nouns were easier than those with animate embedded nouns ($\beta = -0.033$, $SE = 0.003$, $t = -10.70$, $p < 0.001$). An RC type x RC noun animacy effect existed ($\beta = 0.074$, $SE = 0.012$, $t = 6.39$, $p < 0.001$), where SRCs were particularly hard with inanimate RC nouns.

At the RC marker DE, no RC type main effect existed. The RC noun animacy effect remained ($\beta = 0.063$, $SE = 0.007$, $t = 8.71$, $p < 0.001$), where RCs with animate embedded nouns were easier. A significant interaction effect of RC type x embedded noun animacy was present ($\beta = 0.022$, $SE = 0.007$, $t = 3.05$, $p = 0.008$). SRCs with inanimate RC nouns were particularly hard.

At the head, a main clause transitive or intransitive verb, head animacy was added as a factor in the analysis. All three main effects were significant. The RC type effect showed ORC advantage ($\beta = 0.017$, $SE = 0.052$, $t = 3.32$, $p = 0.001$), different from earlier in the RC region. Animacy of the embedded noun stayed significant ($\beta = -0.077$, $SE = 0.028$, $t = -2.78$, $p = 0.043$), with inanimate embedded nouns being easier. Head animacy was also reliable ($\beta = 0.011$, $SE = 0.031$, $t = 3.57$, $p < 0.001$), where inanimate headed-RCs yielded higher error. Interaction effects were not significant.

At the head + 1 position, no main effects or interactions were present.

## DISCUSSION

Beyond the greater difficulty of SRCs compared to ORCs at the head noun, we observed several other reliable effects in RCs modifying main verb subjects. First, we observed a tendency of early difficulty for RCs with inanimate embedded nouns. Encountering an inanimate noun early in the sentence, a position where animate nouns usually appear as an agent and the grammatical subject, yielded high error. Second, there was an effect of head noun animacy at the head noun, such that animate-headed RCs yielded lower error, reflecting the high frequency of animate nouns being at the main clause subject position in the training corpus. This result has been attested in other corpus data (Wu, 2009) and behavioral studies (Wu et al., 2012).

Third, differences between the two RC types changed over the course of the sentence. ORCs initially yielded higher error rates than SRCs, mostly driven by high error for the (unusual) sentence-initial inanimate noun in ORCs. That is, the performance of the model here shows that simple sentences were initially a competitor (error is high for inanimate nouns sentence-initially, because in the more common simple sentences, sentence-initial nouns are animate). These results are compatible with the results of reading time studies that manipulated noun animacy (Wu et al., 2012). However, other empirical data that found an ORC advantage also exist [e.g., Hsiao and Gibson, 2003; Gibson and Wu, 2013; cf. (Lin and Bever, 2006), for non-significant differences]. As one reviewer noted, GPEs at the sentence-initial verb in SRCs were no larger than GPEs at the sentence-initial noun in ORCs, despite the fact that sentence-initial nouns are much more common than sentence-initial verbs (a 70:30 ratio). The reason may be attributed to the prevalence of verb-starting sentences (i.e., pro-drop sentences and SRCs) in the training corpus and the rather low number of simple sentences compared to the realistic statistics in the language. Sentences starting with a transitive verb ($0.3 \times 0.79 + 0.7 \times 0.16 \times 0.58 = 30\%$) in the training set occurred almost as often as sentences starting with an animate noun ($0.7 \times 0.5 = 35\%$). Sentences starting with an inanimate noun, however, were much fewer ($0.7 \times 0.34 = 23.8\%$) and thus generated higher error. The error triggered by noun-starting sentences as reflected in transitional probabilities averaged across animate and inanimate nouns is 29.5%, even lower than the 30% for transitive verb-starting sentences. The lower errors in SRCs than ORCs may have been the reflection of such subtle difference in the transitional probability. However, at the same time, this may reflect lower sensitivity of the model to sentence-initial variation, in the absence of prior context, or to the comparatively low number of simple sentences to RCs in the training set, which may have affected sentence initial predictions more than in later regions where prior context is available. Future work with a more varied corpus and more training may clarify this result.

At the RC marker DE, the SRC advantage disappeared and became the opposite pattern at the head: ORCs yielded lower error than SRCs. This effect is consistent with many prior studies in Mandarin that found lower reading times for ORCs (e.g., Hsiao and Gibson, 2003; Lin and Garnsey, 2011; Gibson and Wu, 2013). Importantly, this higher error for SRCs over ORCs appears even though SRCs are more frequent: the model's performance does not simply reflect RC frequencies but also the frequency of competitors and neighbor structures. Next, we look at the model performance on object-modifying RCs, which are more constrained by previous context.

### *Object-modifying RCs*

To facilitate comparisons across conditions, all test sentences for object-modifying RCs began with an animate main clause subject noun and a transitive verb. As the sentence unfolded to the main clause object position, possible grammatical continuations were more limited than in the sentence-initial RCs reviewed above, owing to the greater preceding context in object-modifying items. **Figure 4** shows the trajectory of GPE values for SRCs and ORCs in the RC region, head, and the end-of-sentence marker.

Examination of **Figure 4** reveals a general pattern of ORCs being harder than SRCs, in contrast to the subject-modifying relative clauses in **Figure 3**, for which ORCs were initially slightly harder, and then SRCs were substantially harder after the head noun. This result is also found in the human comprehension literature: Lin and Bever (2006) found longer reading times for object-modifying ORCs than SRCs. Due to the initial unusual word order of two consecutive verbs (main clause verb and RC embedded verb), SRCs are disambiguated early, yielding a small amount of early difficulty but later being easier, because this conjunction of two verbs removes competitors other than an SRC. By contrast, ORCs remain ambiguous: the first word in ORCs in combination with the matrix clause context (the N V N order) created a strong bias toward a simple sentence, which was hard to revise even at the end.

***Word-by-word findings.*** Here we call the starting word of an object-modifying RC "W3" (RC verb for SRCs, RC subject for ORCs) because it was preceded by the main clause subject and verb and therefore was the third word in the sentence. At this word, the main effect of RC type indicated that SRCs were significantly harder than ORCs ($\beta = 0.137$, $SE = 0.017$, $t = 7.96$, $p < 0.001$). Animacy of the RC embedded nouns had a significant effect too ($\beta = -0.094$, $SE = 0.006$, $t = -16.30$, $p < 0.001$), where the inanimate RC nouns induced lower GPEs. The interaction between RC type and embedded noun animacy was significant ($\beta = 0.090$, $SE = 0.008$, $t = 11.16$, $p < 0.0001$). ORCs appeared to have a larger GPE difference between the ones with animate RC nouns (higher) and the ones with inanimate RC nouns (lower), whereas the two kinds of SRCs almost had overlapping GPEs.

At W4 (RC object for SRCs, RC verb for ORCs), the previous significant effects reversed: SRCs were now easier than ORCs ($\beta = -0.072$, $SE = 0.024$, $t = -2.98$, $p = 0.008$) and animate RC nouns instead produced more accurate predictions than inanimate ones ($\beta = 0.118$, $SE = 0.024$, $t = 4.97$, $p < 0.001$). Interaction between the two factors was significant ($\beta = -0.109$, $SE = 0.026$, $t = -4.16$, $p < 0.001$). ORCs with inanimate embedded nouns had particularly high GPEs.

At the RC marker DE, RC type was non-significant. The RC noun animacy effect was still present ($\beta = 0.051$, $SE = 0.006$, $t = 8.32$, $p < 0.001$). Animate RC nouns again were preferred to inanimate ones. Interaction between the two was significant ($\beta = -0.058$, $SE = 0.011$, $t = -5.53$, $p < 0.001$).

At the head, the main effect of RC type was significant, with GPEs of ORCs higher than those of SRCs ($\beta = -0.20$, $SE = 0.098$, $t = -2.08$, $p < 0.001$). There was also a significant main effect of head noun animacy ($\beta = -0.13$, $SE = 0.032$, $t = -4.12$, $p < 0.001$). Inanimate heads were preferred to animate heads. Other effects and interaction were not significant.

At the end-of-sentence maker, the errors went down to nearly zero for all conditions. No effects were present.

## DISCUSSION

Compared to the subject-modifying position, the model performance exhibited an opposite pattern at the object-modifying site: SRCs were easier than ORCs at the head noun. Error rates reflected early effects of low frequency sequences, followed by lower error rates because these low frequency sequences are themselves highly predictive of subsequent input (e.g., the very rare V V sequence in SRCs yields high error, but a V V sequence strongly predicts subsequent input, leading to low error at the next word). Overall, the SRN's performance on object-modifying RCs once again confirmed the model's sensitivity to both the structural and lexical statistics of the target RC structures and those of

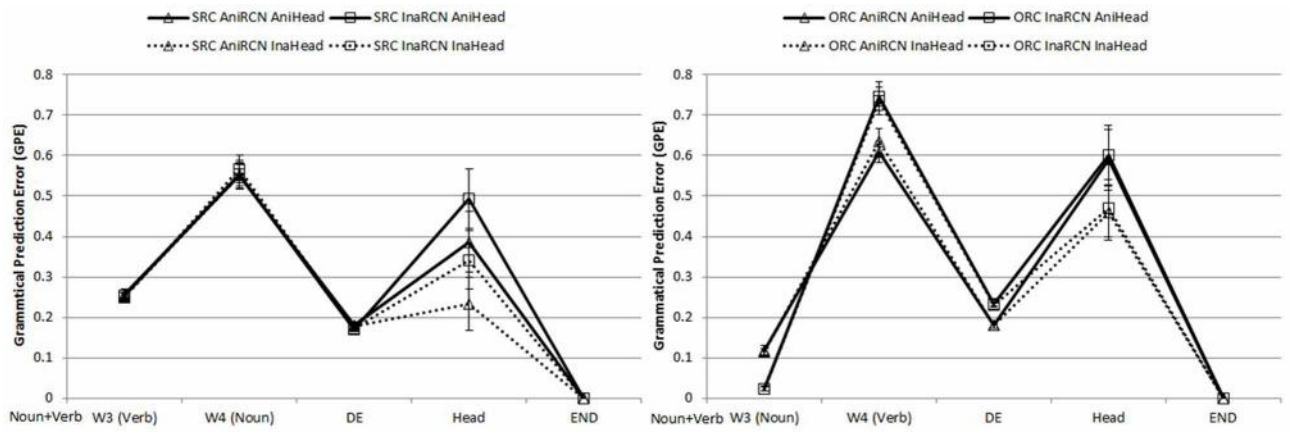



**FIGURE 4 | Word-by-word GPEs of object-modifying SRCs (left panel) and ORCs (right panel) in the critical RC region, the head, and next word after the head.** All sentences began with a Noun + Verb sequence, not shown in the graphs.

competing and neighbor structures under the constraining prior context of a main clause subject and verb. The processing advantage of SRCs over ORCs at the head noun resembles the findings in Lin and Bever (2006), which could be a result of the overall higher structural frequency and the lack of competing structures of SRCs at this position. In the next section, we investigate the effect of modifying position by analyzing SRCs and ORCs at both modifying positions together.

#### *Contrasting subject vs. object modification*

From the separate analyses above of the two modification sites, we observed almost opposite patterns of GPEs at the head. Prior RC studies have assessed RC difficulty at various regions: words inside the RC, at the head, at the head + 1 position, even further down until the end of sentence, or average across the whole sentence. It is very true that the current model showed a dynamic nature in its processing patterns across the RC, the head, and post-head. Due to the head-final order of RCs in Mandarin, the head is considered the first position where the processing pattern between SRCs and ORCs is differentiated. The working memory-based account argues that integration of related elements in an RC (filler and gap) occurs at this point. Prior empirical studies of Mandarin RCs typically found robust effect of RC type starting from the head Therefore, to compare the difficulty between two RC types at two modification positions, we examined the GPEs at the head position. It was not only the position that reflected the accumulated processing difficulty across the RC but also the first position that reflected the head noun animacy effect.

**Figure 5** displays the average GPEs of SRCs and ORCs at the word following the head noun, across both levels of head noun animacy and modification site. ORCs were easier than SRCs when main clause subjects were modified, whereas SRCs were easier than ORCs modifying main clause objects. Animate heads were preferred to inanimate heads as matrix subjects while inanimate heads were preferred to animate heads as matrix objects.

At this word position, even though there were no significant main effects of RC type, head noun animacy, and modifying position, there were significant interactions of modifying position with RC type ($\beta = 0.37$, $SE = 0.043$, $t = 8.76$, $p < 0.001$) and head noun animacy ($\beta = 0.24$, $SE = 0.043$, $t = 5.61$, $p < 0.001$). These results suggest that sentential environment (subject vs. object modification) played an important role in the processing the two types of RCs, in combination of the animacy properties of the head. Prior human reading time studies typically have considered only a fraction of the conditions shown, and the conflicting results found in some studies may stem from failures to consider the full pattern of data. Therefore, we examine representative Mandarin RC studies and compare them with our model performance in the next section.

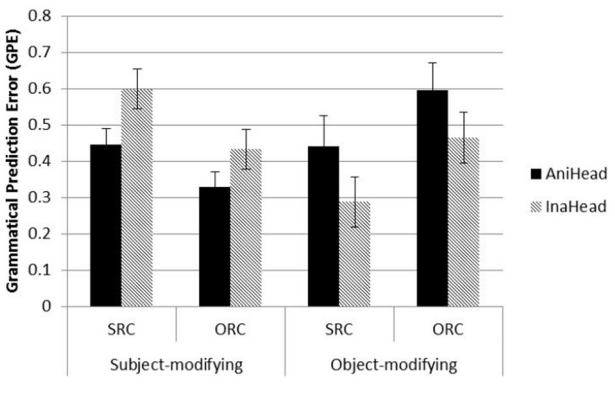



**FIGURE 5 | GPEs of SRCs and ORCs at the head noun.** AniHead = animate head noun, InaHead = inanimate head noun.

## COMPARISON WITH PREVIOUS MANDARIN RC READING TIME STUDIES

In this section we compare the SRN's performance to self-paced reading studies of Mandarin RC processing. As **Table 5** shows, there are three major patterns of results in the Mandarin relative clause literature. The first major result is that in studies of subject-modifying relative clauses, ORCs are easier than SRCs, a reverse of the typical pattern in most other languages. Studies yielding this pattern include Gibson and Wu (2013) and Hsiao and Gibson (2003, though only for multiply-embedded sentences in the latter study. An ORC advantage occurred only at the first two words of singly-embedded sentences and Vasishth et al. (2013) reported instead an SRC advantage at the head using Hsiao and Gibson's singly-embedded materials), Su et al. (2007), Chen et al. (2008), and Lin and Garnsey (2011). The SRN captured the major result of subject-modifying ORCs being easier than SRCs. Compared to subject-modifying ORCs, subject-modifying SRCs yielded reliably higher error rates in the model at the head noun ($\beta = 0.17$, $SE = 0.052$, $t = 3.33$, $p = 0.003$) and in an average of DE and the head together (as in Gibson and Wu's analysis), the ORC preference was also robust ($\beta = 0.08$, $SE = 0.023$, $t = 3.60$, $p = 0.002$).

Many of the articles listed in the top row of **Table 5** described the ORC advantage as a general tendency of Mandarin sentence processing and working memory limitations, even though their studies investigated only a subset of RCs, namely RCs modifying main clause subjects and containing entirely animate nouns. This brings us to the second and third patterns in **Table 5**, which reflect the non-universality of the results in the top row. The second major pattern, reported by Lin and Bever (2006), is the effect of modification position. They found that object-modifying RCs were harder than subject-modifying RCs, reflected in reliable reading time differences at many word positions. On first glance, that result does not seem to correspond to the model's performance, but if we examine the model's behavior in the exact sentence types they tested (all animate nouns), the model and human data look much more similar. For this subset of conditions, Lin and Bever's RC-type effect was replicated in our model, with lower GPEs for SRCs compared to ORCs at several word positions (pre-DE: $\beta = -0.04$, $SE = 0.016$, $t = -2.54$, $p = 0.009$; head: $\beta = -0.20$, $SE = 0.025$, $t = -8.08$, $p < 0.001$). We also roughly replicated their main effect of modifying position (subject-modifying easier) at the pre-DE region (W1 + W2) ($\beta = -0.18$, $SE = 0.016$, $t = -11.26$, $p < 0.001$), and also at the head ($\beta = -0.22$, $SE = 0.025$, $t = -8.81$, $p < 0.001$). Interactions were significant at these two positions too (pre-DE:

**Table 5 | Comparison of experimental materials and major findings of representative human reading time studies.**

| Major data patterns | Experiment conditions | Studies | Specific manipulations | Model replication? |
|---|---|---|---|---|
| ORCs easier than SRCs at the head noun and/or nearby words | Only subject-modifying RCs, all nouns animate | Gibson and Wu (2013) | Supportive context | Yes |
| | | Hsiao and Gibson (2003) | Doubly-embedded RCs (ORC advantage only at pre-DE region in singly-embedded RCs) | |
| | | Su et al. (2007) | Aphasic patients | |
| | | Chen et al. (2008) | Memory spans | |
| | | Lin and Garnsey (2011) | Topicalization | |
| Non-replication of Hsiao and Gibson (SRCs easier than ORCs) replication of Gibson and Wu | | Vasishth et al. (2013) | Hsiao and Gibson (2003) materials | No |
| | | | Gibson and Wu (2013) materials | Yes |
| Object-modifying SRCs easier than ORCs | Both subject- and object-modifying (all nouns animate) | Lin and Bever (2006) | | Yes |
| No difference between SRCs and ORCs with preferred animacy configuration | Animacy of RC noun and head noun (all subject-modifying) | Wu et al. (2012) | Contrastive animacy of RC noun and head noun | Yes; model shows small differences where Wu et al. find little or no difference |

$\beta = -0.06$, $SE = 0.021$, $t = -2.84$, $p = 0.005$; head: $\beta = 0.38$, $SE = 0.036$, $t = 10.51$, $p < 0.001$). When considering each modifying position separately, Lin and Bever did not find any RC type difference for subject-modifying RCs but a significant difference at DE and head for object-modifying RCs. Our model replicated the strong effect of SRC advantage at the object-modifying position ($\beta = -0.20$, $SE = 0.025$, $t = -8.16$, $p < 0.001$) that Lin and Bever (2006) found.

The third major pattern is the effect of noun animacy on RC processing, exemplified by Wu et al's. (2012) manipulation of both head and RC noun animacy within subject-modifying relative clauses. Their study did not fully cross head and RC noun animacy and included only contrastive animacy conditions (one noun animate and one inanimate). They found that in the preferred animacy configuration (in which the animate noun was the agent of the RC verb and the inanimate noun the theme), SRCs and ORCs didn't differ in processing difficulty. However, with the dispreferred animacy configuration (inanimate agents, animate patients), ORCs (such as the Mandarin equivalent of *The hiker that the rocks crushed*) were read particularly slowly and an SRC advantage emerged. Our model performance did show such SRC preference for these unusual animacy RCs early at W1 ($\beta = -0.09$, $SE = 0.008$, $t = -11.12$, $p < 0.001$) and later at DE ($\beta = -0.07$, SE = 0.007, $t = -10.04$, $p < 0.001$), but the ORC became easier at the head ($\beta = 0.41$, $SE = 0.053$, $t = 7.68$, $p < 0.001$). For RCs with the preferred animacy configuration, SRCs showed early advantage at W1 ($\beta = -0.01$, $SE = 0.006$, $t = -2.18$, $p = 0.03$) but later switched to ORC advantage at DE ($\beta = 0.08$, $SE = 0.011$, $t = 07.25$, $p < 0.001$) and switched back at the head ($\beta = -0.12$, $SE = 0.050$, $t = -2.48$, $p = 0.02$) but the effect was rather reduced compared to the disfavored animacy configurations. Furthermore, when considering all possible animacy configurations (i.e., AniRCN + AniHead, InaRCN + AniHead, AniRCN + InaHead, InaRCN + InaHead) rather than only the two that Wu et al. (2012) investigated, we found no difference between the favored configurations, namely animate-headed SRCs and inanimate-headed ORCs, at every word except for W1 ($\beta = -0.06$, $SE = 0.011$, $t = -5.28$, $p < 0.001$). Thus, the model shows comparatively small differences in the same conditions that Wu et al. find little or no difference in reading times.

In sum, the model captured major patterns of comprehension difficulty across several empirical studies, despite the fact that these studies are often thought to conflict with one another. The model's performance suggests that the inconsistencies in the literature are more apparent than real and stem from different experimental materials used in the experiments, which focus on a small subset of relative clause and animacy types. Whereas it is impossible to manipulate all relevant factors within a single self-paced reading experiment, the current SRN model could incorporate 16 types of test sentences in a $2 \times 2 \times 2 \times 2$ design. The model data present a more comprehensive picture, in which RC type, modifying position, head noun animacy, and RC noun animacy all have an effect. These data suggest that relative clause difficulty in Mandarin depends on a complex interplay of probabilistic constraints from animacy and other information gleaned from prior experience. Thus, contra many claims in the literature (e.g., Hsiao and Gibson, 2003; Lin and Bever, 2006; Gibson and

Wu, 2013; Vasishth et al., 2013) none of the empirical results warrant broad conclusions about ORCs or SRCs being universally easier or harder.

## GENERAL DISCUSSION

In this paper, we presented an SRN simulation of the processing of Mandarin relative clauses. We had two related goals. First, we wanted to use the model to investigate issues that are difficult to test in human experiments. We suspected that controversies in the empirical literature stemmed at least in part from complex interactions among a number of factors such that when researchers designed materials tapping different subsets of the factors, different results obtained. Constraints on human studies, such as biases or priming effects that arise when comprehenders encounter many sentences of the same type, typically prevent anything more complex than a $2 \times 2$ design in sentence processing studies. By contrast, our $2 \times 2 \times 2 \times 2$ design presented in 16 test sentence types showed that the four different factors we examined (RC type, modifying position, head animacy, and RC noun animacy) interacted in complex ways in the model. The results from the model closely track reading time patterns from a number of human comprehension studies and suggest that there is no overall SRC or ORC preference in Mandarin. Instead, the results strongly depend on which types of relative clauses are contrasted. In relative clauses with the animacy configurations most commonly investigated in human studies to date, ORCs are easier than SRCs when they modify main clause subjects but the reverse is true when they modify main clause objects, but other patterns are obtained with different animacy configurations, which can be seen most clearly in **Figure 5**. These results suggest that claims for broad categories of relative clause difficulty in Mandarin are premature at best.

Second, we wanted to use an SRN's ability to generalize over similar items (MacDonald and Christiansen, 2002; Fitz et al., 2011) to investigate how competitor and neighbor interpretations affect RC interpretations. Here our corpus analysis showed that RCs in Mandarin can be highly ambiguous for human comprehenders. When the model was trained on the resulting training set, prediction error varied with all four factors investigated here.

An important component of this second goal was our incorporation of animacy information in the corpus analyses and in the model. Human comprehension patterns clearly show the importance of animacy information in relative clause processing (Wu et al., 2012), and the model also captured these animacy effects, despite having no conceptual information that would typically be used to code distinctions between animate and inanimate entities. These results show the power of the sequential information associated with animacy, that sentences containing animate entities have different distributional patterns in the language than those with inanimate nouns. Chang (2009) made a related argument in his examination of cross-linguistic variation in language production, that learning over the distributional regularities of various animacy configurations in a language is critical for explaining why animacy has different effects in a relatively strict word order language such as English compared with a relatively free word order language such as Japanese—people are learning the sequential information associated with sentences of different types, and the animacy-structure patterns vary as a function of the rigidity of the word order in the language. His results, as well of those of our model, suggest that sequential learning is an important adjunct to conceptual information in accounts of animacy effects in sentence-level language use.

Another important component of our investigation of neighborhood effects is the link between generalization over similar sentences and a topic that initially seems unrelated: the role of working memory in accounts of language comprehension. Because of its generalization over many sentence types, the model shows patterns of difficulty analogous to those in human studies. Critically, it does so in a system in which experience influences the model's computational capacity (and thus its ability to make accurate predictions for upcoming input, MacDonald and Christiansen, 2002). This result contrasts with claims questioning the adequacy of experience-based accounts in RC processing. Levy et al. (in press) and Levy and Gibson (2013) have assessed experience via calculations of a word's surprisal—the conditional probability of that word given prior context. They suggest that surprisal does not correctly predict the full pattern of human relative clause reading times, and they argue that human comprehension difficulty requires supplementing surprisal with an account of human memory burdens, as in Gibson's (1998) Locality theory, in which RC difficulty varies with load in an experience-independent working memory. Given our own results and MacDonald and Christiansen's success in using an SRN to model humans' reading times of English RCs, we think it is premature to reject all experience-based accounts on the basis of failures of particular implementations (particular surprisal instantiations). First, as Frank (2009) notes, the success of Levy and colleagues' surprisal calculation varies with the richness of the prior input. Thus, it may be that a larger or more realistic corpus over which to calculate conditional probabilities would yield a better account of humans' experience and consequently better prediction of reading times. However, we suspect there is a second reason why SRNs can yield different predictions than Levy and colleagues' surprisal results, concerning how context is represented and transformed into predictions. As it has been implemented to date, surprisal is based on an aggregation of past instances, which is used to calculate the conditional probability of upcoming input. By contrast, the SRN is a learning model that compresses and transforms its experience into an internal representation as it learns (Elman, 1990; Tabor et al., 1997; Frank, 2009). As a result, the SRN generalizes over neighboring structures and shows behavior that is not always a sum of instances in the training set, such as when the more frequent subject-modifying SRC sentences yield higher error rates than the rarer subject-modifying ORC sentences. A more direct comparison of SRNs and various instantiations of surprisal should be an important step in better understanding the role of experience in RC processing (see also Frank, 2009). In the meantime, we see no need to complicate our experience-based account with an additional component as in Levy and Gibson's proposal, and indeed we see the success of our SRN as further evidence for the non-independence of experience and computational capacity/working memory

(McClelland and Elman, 1986; MacDonald and Christiansen, 2002).

### FUTURE DIRECTIONS

The SRN we used was trained on a set of sentences based on realistic human linguistic experiences, gleaned from a detailed corpus analysis. Although other corpus analyses exist for Mandarin RCs (Hsiao and Gibson, 2003; Pu, 2007; Wu, 2009; Vasishth et al., 2013), our study was unique in its large scale (incorporating many non-RC structures), its hand-coding of animacy across both RC and non-RC sentences, and its use in training the SRN. The combination with the SRN is crucial here because RC processing difficulty in both human and model is not simply a function of the frequency of RCs. We consider the following points important steps to further improve the current model.

The claim that Mandarin RC processing is tied to uncertainty of predictions echoes similar claims for other languages (e.g., Gennari and MacDonald, 2008). The SRN offers important opportunities to more stringently test these claims in future research. For example, to test whether simple pro-drop sentences are truly a competitor for certain RCs, the current model (containing pro-drop sentences) can be compared to one trained on a variant of Mandarin without pro-drop (i.e., replacing all pro-drop sentences in the training set with overt subject variants). We are pursuing these and other model comparisons aimed at elucidating the role of neighbors/competitors in the input. In addition, we intend to manipulate the amount of training, as in MacDonald and Christiansen (2002), to further observe the developmental trends in learning and generalizing. Thus, SRNs can help trace the sources of processing difficulty using methods that are impossible to use with human comprehenders.

Another future endeavor should involve a more accurate portrayal of the rate of RCs and other related structures in Mandarin. For example, the multiple functions of *DE* (see Appendix A), which has been considered a disambiguating cue in RCs can in actuality create ambiguities depending on the semantic context. Structures involving these other *DE* uses should be considered potential competitors. Another typological feature in Mandarin that might be relevant is the possibility of null object (e.g., He saw (the movie).) with supportive context, contrastive to null subject considered in the current model. The effects of null object and its combination with null subject are yet to be explored. Of course models necessarily remain simplifications of the entire linguistic experience (and simplifications of many other dimensions of human cognition). Model expansions therefore should not just cover more data but provide new insight into how people weigh multiple probabilistic constraints during sentence interpretation.

An important feature of our training set was its coding of distributional aspects of lexical meaning, as originally demonstrated by Elman (1990). That is, the model had no explicit semantic representations but it came to distinguish animate and inanimate nouns and transitive and intransitive verbs by virtue of their different distributions in the training set. These fine-grained discriminations were crucial for the model's account of animacy effects in RC processing. These and other studies of distributional semantic effects in SRNs (Elman, 1990; St John and McClelland, 1990) show that there is potential for future work to incorporate other distributional "semantic" effects. There is also potential to make the distributions more natural, thereby capturing some additional ambiguity effects not in the present model. For example, for the current model, all verbs in the training set were either 100% transitive or 100% intransitive, with no optionally transitive verbs such as *eat*, which can occur either with or without a direct object. Optionally transitive verbs add additional indeterminacy, in that a model encountering an optionally transitive verb will be uncertain about an upcoming direct object, an effect which could modulate differences between SRCs (which can have transitive or intransitive verbs) and ORCs (which must have transitive verbs).

## CONCLUSION

The current study confirmed SRN models as a promising tool in modeling human sentence processing, and, in this particular case, appropriate to examine intricate and complicated dynamics of Mandarin RC processing. The architecture of SRNs allows flexibility in modeling multiple effects in a single model, whereas manipulating a large number of factors human studies is nearly impossible. Mandarin is typologically unique in its conjunction of head-final RCs and head-initial SVO basic word, and yet in some sense the model's behavior looks very similar to that of SRN models of English RCs (MacDonald and Christiansen, 2002; Fitz et al., 2011). That is, the patterns of SRC vs. ORC difficulty are wildly different for the two languages, but in both cases, the models are strongly affected by the balance of RCs, competitors and neighbors. The modeling results suggest that rather than arguments for universal Locality (Gibson, 1998) or universal SRC preference (Lin and Bever, 2006), the real universals in human RC processing are exquisite sensitivity to the statistical regularities of across many different types of input.

## ACKNOWLEDGMENTS

This work was supported by a grant from the National Science Foundation (BCS 1123788). We thank Jinman Li for helpful discussions of Mandarin and Tim Rogers for many helpful discussions of the model.

## REFERENCES

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bock, K., and Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition* 21, 47–67. doi: 10.1016/0010-0277(85)90023-X

Chang, F. (2009). Learning to order words: a connectionist model of heavy NP shift and accessibility effects in Japanese and English. *J. Mem. Lang.* 61, 374–397. doi: 10.1016/j.jml.2009.07.006

Chen, B., Ning, A., Bi, H., and Dunlap, S. (2008). Chinese subject-relative clauses are more difficult to process than the object-relative clauses. *Acta Psychol.* 129, 61–65. doi: 10.1016/j.actpsy.2008.04.005

Christiansen, M. H. (1994). *Infinite Languages, Finite Minds: Connectionism, Learning and Linguistic Structure*. Unpublished doctoral dissertation, University of Edinburgh.

Christiansen, M. H., and Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cogn. Sci.* 23, 157–205. doi: 10.1207/s15516709cog2302_2

Diessel, H. (2004). *The Acquisition of Complex Sentences*. Cambridge:

Cambridge University Press. doi: 10.1017/CBO9780511486531

Diessel, H. (2007). A construction-based analysis of the acquisition of East Asian relative clauses. *Stud. Second Lang. Acquis.* 29, 311–320. doi: 10.1017/S0272263107070167

Diessel, H., and Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language* 81, 1–25. doi: 10.1353/lan.2005.0169

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* 7, 195–225. doi: 10.1007/BF00114844

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition* 48, 71–99. doi: 10.1016/0010-0277(93)90058-4

Fitz, H., Chang, F., and Christiansen, M. H. (2011). "A connectionist account of the acquisition and processing of relative clauses," in *The Acquisition of Relative Clauses: Processing, Typology and Function*, Vol. 8, ed E. Kidd (Amsterdam: John Benjamins), 39–60.

Frank, S. L. (2009). "Surprisal-based comparison between a symbolic and a connectionist model of sentence processing," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, eds N.A. Taatgen and H. van Rijn (Austin, TX: Cognitive Science Society), 1139–1144.

Frauenfelder, U., Segui, J., and Mehler, J. (1980). Monitoring around the relative clause. *J. Verb. Learn. Verb. Behav.* 19, 328–337. doi: 10.1016/S0022-5371(80)90257-1

Frazier, L. (1987). Syntactic processing: evidence from Dutch. *Nat. Lang. Linguist. Theory* 5, 519–559. doi: 10.1007/BF00138988

Gennari, S. P., and MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *J. Mem. Lang.* 58, 161–187. doi: 10.1016/j.jml.2007.07.004

Gennari, S. P., and MacDonald, M. C. (2009). Linking production and comprehension processes: the case of relative clauses. *Cognition* 111, 1–23. doi: 10.1016/j.cognition.2008.12.006

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E., Desmet, T., Grodner, D., Watson, D., and Ko, K. (2005). Reading relative clauses in English. *Cogn. Linguist.* 16, 313. doi: 10.1515/cogl.2005.16.2.313

Gibson, E., and Wu, H.-H. I. (2013). Processing Chinese relative clauses in context. *Lang. Cogn. Processes* 28, 125–155. doi: 10.1080/01690965.2010.536656

Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B. (eds.). (2005). *The World Atlas of Language Structure*. Oxford: Oxford University Press.

Hsiao, F., and Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition* 90, 3–27. doi: 10.1016/S0010-0277(03)00124-0

Hundt, M. (2004). Animacy, agentivity, and the spread of the progressive in Modern English. *Engl. Lang. Linguist.* 8, 47–69. doi: 10.1017/S1360674304001248

Keenan, E. L. (1985). "Relative clauses," in *Language Typology and Syntactic Description,* Vol II*: Complex Constructions*, ed T. Shopen (Cambridge: Cambridge University Press), 141–170.

Keenan, E. L., and Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguis. Inq.* 8, 63–99.

King, J., and Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. *J. Mem. Lang.* 30, 580–602. doi: 10.1016/0749-596X(91)90027-H

Kwon, N., Lee, Y., Goron, P. C., Kluender, R., and Polinsky, M. (2010). Cognitive and linguistic factors affecting subject/object asymmetry: an eye-tracking study of pre-nominal relative clauses in Korean. *Language* 82, 546–582. doi: 10.1353/lan.2010.0006

Levy, R., Fedorenko, E., and Gibson, E. (in press). The syntactic complexity of Russian relative clauses. *J. Mem. Lang*. doi: 10.1016/j.jml.2012.10.005

Levy, R., and Gibson, E. (2013). Surprisal, the PDC, and the primary locus of processing difficulty in relative clauses. *Front. Lang. Sci.* 4:229. doi: 10.3389/fpsyg.2013.00229

Li, C. N., and Thompson, S. A. (1981). *Mandarin Chinese: A Functional Reference Grammar*. Berkeley, CA: University of California Press.

Lin, C. J. C. (2006). *Relative Clause Processing in Typologically Distinct Languages: A Universal Parsing Account*. Unpublished doctoral dissertation, University of Arizona, Tucson, AZ.

Lin, C.-J. C., and Bever, T. G. (2006). "Subject preference in the processing of relative clauses in Chinese," in *Proceedings of the 25th West Coast Conference on Formal Linguistics*, eds D. Baumer, D. Montero, and M. Scanlon (Somerville, MA: Cascadilla Proceedings Project), 254–260.

Lin, C.-J. C., and Bever, T. G. (2011). "Garden path and the comprehension of head-final relative clauses," in *Processing and Producing Head-final Structures*, eds H. Yamashita, Y. Hirose, and J. L. Packard (New York, NY: Springer), 277–297.

Lin, Y. B., and Garnsey, M. (2011). "Animacy and the resolution of temporary ambiguity in relative clause comprehension in Mandarin," in *Processing and Producing Head-final Structures*, eds H. Yamashita, Y. Hirose, and J. L. Packard (New York, NY: Springer), 241–276.

MacDonald, M. C., and Christiansen, M. H. (2002). Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychol. Rev.* 109, 35–54. doi: 10.1037/0033-295X.109.1.35

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101, 676–703. doi: 10.1037/0033-295X.101.4.676

Mak, W. M., Vonk, W., and Schriefers, H. (2002). The influence of animacy on relative clause processing. *J. Mem. Lang.* 47, 50–68. doi: 10.1006/jmla.2001.2837

Mak, W. M., Vonk, W., and Schriefers, H. (2006). Animacy in processing relative clauses: the hikers that rocks crush. *J. Mem. Lang.* 54, 466–490. doi: 10.1016/j.jml.2006.01.001

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

Miller, G. A., and Chomsky, N. (1963). "Finitary models of language users," in *Handbook of Mathematical Psychology*, eds R. D. Luce, R. Bush, and E. Galanter (New York, NY: John Wiley), 419–492.

Miyamoto, E. T., and Nakamura, M. (2003). "Subject/object asymmetries in the processing of relative clauses in Japanese," in *Proceedings of the 22nd WCCFL*, eds G. Garding and M. Tsujimura (Somerville, MA: Cascadilla Press), 342–355.

Ozeki, H., and Shirai, Y. (2007). Does the Noun Phrase Accessibility Hierarchy predict the difficulty order in the acquisition of Japanese relative clauses? *Stud. Second Lang. Acquis.* 29, 169–196. doi: 10.1017/S0272263107070106

Pu, M.-M. (2007). The distribution of relative clauses in Chinese discourse. *Discourse Process.* 43, 25–53. doi: 10.1207/s15326950dp4301_2

Rohde, D. (1999). *LENS: the Light, Efficient Network Simulator*. Technical Report CMU-CS-99-164. Pittsburgh, PA: Carnegie Mellon University.

Rohde, D. (2005). *TGrep2 Manual*. Available online at: http://tedlab.mit.edu/~dr/Tgrep2/tgrep2.pdf

Roland, D., Dick, F., and Elman, J. L. (2007). Frequency of basic English grammatical structures: a corpus analysis. *J. Mem. Lang.* 57, 348–379. doi: 10.1016/j.jml.2007.03.002

Schriefers, H., Friederici, A. D., and Kuhn, K. (1995). The processing of locally ambiguous relative clauses in German. *J. Mem. Lang.* 34, 499–520. doi: 10.1006/jmla.1995.1023

St John, M. F., and McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artif. Intell.* 46, 217–257. doi: 10.1016/0004-3702(90)90008-N

Su, Y., Lee, S., and Chung, Y. (2007). Asyntactic thematic role assignment by Mandarin aphasics: a test of the Trace-Deletion Hypothesis and the Double Dependency Hypothesis. *Brain Lang.* 101, 1–18. doi: 10.1016/j.bandl.2006.12.001

Tabor, W., Juliano, C., and Tanenhaus, M. K. (1997). Parsing in a dynamical system: an attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Lang. Cogn. Processes* 12, 211–271. doi: 10.1080/016909697386853

Tanenhaus, M. K., and Trueswell, J. C. (1995). "Sentence comprehension," in *Handbook of Perception and Cognition*. Vol. 11*: Speech, Language and Communication*, eds J. L. Miller and P. D. Eimas (San Diego, CA: Academic Press), 217–262.

Traxler, M. J., Morris, R. K., and Seely, R. E. (2002). Processing subject and object relative clauses: evidence from eye movements. *J. Mem. Lang.* 47, 69–90. doi: 10.1006/jmla.2001.2836

Traxler, M. J., Williams, R. S., Blozis, S. A., and Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *J. Mem. Lang.* 53, 204–224. doi: 10.1016/j.jml.2005.02.010

Vasishth, S., Chen, Z., Li, Q. and Kuo, G. (2013). Processing Chinese relative clauses: evidence for the universal subject preference. *PLoS ONE* 8:

e77006. doi: 10.1371/journal.pone.0077006

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., and MacDonald, M. C. (2009). Experience and sentence processing: statistical learning and relative clause comprehension. *Cogn. Psychol.* 58, 250–271. doi: 10.1016/j.cogpsych.2008.08.002

Wu, F. (2009). *Factors Affecting Relative Clause Processing in Mandarin*. Unpublished doctoral dissertation, University of Southern California, Los Angeles, CA.

Wu, F., Kaiser, E., and Andersen, E. (2012). Animacy effects in Chinese relative clause processing. *Lang. Cogn. Processes* 27, 1489–1524. doi: 10.1080/01690965.2011.614423

Xue, N., Jiang, Z., Zhong, X., Palmer, M., Xia, F., Chiou, F.-D., et al. (2010). *Chinese Treebank 7.0*. Philadelphia, PA: Linguistic Data Consortium.

Yip, V., and Matthews, S. (2007). Relative clauses in Cantonese-English bilingual children: typological challenges and processing motivations. *Stud. Second Lang. Acquis.* 29, 277–300. doi: 10.1017/S0272263107070143

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 04 June 2013; accepted: 30 September 2013; published online: 22 October 2013.*

*Citation: Hsiao Y and MacDonald MC (2013) Experience and generalization in a connectionist model of Mandarin Chinese relative clause processing. Front. Psychol. **4**:767. doi: 10.3389/fpsyg.2013.00767*

*This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*

## APPENDIX A

### AMBIGUITY AND COMPETITOR INTERPRETATIONS FOR MANDARIN RELATIVE CLAUSES

Due to the head-final feature of Mandarin relative clauses and the fact that the relative pronoun DE appears at the end of a relative clause, Mandarin relative clauses contain many temporary ambiguities. Even at the relative clause marker DE, the structure could still be ambiguous because DE has other functions in Mandarin beyond the relative pronoun. Below, we detail some of these ambiguities—strings that could turn out to be a relative clause of some type but could also turn out to be some other kind of structure. In some cases, these structures may conflict with the relative clause interpretation at one sentence position but may at other times facilitate the RC interpretation (be a neighbor).

#### *Simple sentences*

The dominant Subject-Verb-Object word order in Mandarin creates different amounts of ambiguity for object relative clauses occurring at different main clause positions. For example, an object relative clause modifying a main clause subject begins with the sequence "N-V" (as in 2b) which can be interpreted as a main clause subject and verb (the simple sentence is a temporary competitor), but later regions of the sentence reveal that this initial N-V sequence was instead part of a relative clause. The prevalence of Mandarin Subject-Verb-Object simple sentences could potentially aid the processing of sentences that share word order similarities (MacDonald and Christiansen, 2002; Fitz et al., 2011). Thus, the fact that simple transitive sentences and subject-modifying object relatives initially have the same word order creates first an ambiguity, and then when the simple sentence interpretation is removed by additional input, the simple sentences are a helpful neighbor, in that past experience with simple sentences can help interpretation of these relative clauses.

By contrast, simple transitive sentences are not a competitor or neighbor for subject relative clauses modifying a main clause object (as in 3a). These structures contain a sequence of two verbs early in the sentence, which rules out a simple sentence interpretation of the input. Subject relatives might therefore be somewhat difficult early when the unusual sequence of two verbs is encountered, but overall, these sentences are less ambiguous than the object relatives, for which the simple sentence interpretation can persist over more input.

These examples show that generalization and competition from simple sentences manifest differently in the two main clause positions. These patterns are important because the dominant results in the comprehension literature find that subject relatives in subject-modifying position are harder than object relatives (Hsiao and Gibson, 2003; Gibson and Wu, 2013), while the opposite pattern obtains for object-modifying RCs (Lin, 2006; Lin and Bever, 2006).

#### *Pro-drop sentences*

One important competitor structure for relative clauses comes from simple main clause sentences in which the subject NP has been omitted. Mandarin is a pro-drop language (meaning that in context, grammatical subjects are omissible; Li and Thompson, 1981), so that simple sentences such as "VT-Object" exist as complete sentences in the language. These pro-drop sentences are temporary competitors for subject-modifying subject relatives, which have the sequence "VT-Object-DE-Subject." As more input comes in to disambiguate the structure, the comprehender may recover from the initial difficulty and be aided by this resemblance in word order with pro-drop sentences. Such facilitation effect for subject relatives may be even more pronounced and activated sooner at the main clause object site. The unusual word order of two verbs in a row after the main clause subject creates short-term difficulty but meanwhile also provides a reliable cue for a subject relative clause and facilitates processing of the rest part.

#### *Subject relative clauses with intransitive verbs*

Subject relatives with intransitive verbs are rarely examined in prior empirical studies. Structurally, they resemble subject relatives with transitive verbs, with the lack of a direct object. Thus, experiences with one structure could be likely transferable to the other.

#### *Other structures with DE*

The surface word order of Mandarin relative clauses is identical to many other structures that also have the particle DE. In addition to serving as a relativizer in a relative clause, DE is also a marker for adjectives and adverbs and also appears in possessives and appositives or simply as a phrase-final particle. With these many functions of DE, which is sometimes considered the disambiguating point for a relative clause, the relative clause structure could still be ambiguous after DE and permits various interpretations; the only definitive disambiguation is discourse context. For example, in the phrase "respect teacher DE parents," the DE could be interpreted as a relativizer of a subject relative: "the parents that respect the teacher," or as a possessive marker in a nominalized VP gerund: "respecting the teacher's parents." The appositive structure with DE also shares identical word order with a subject relative, such as in "respect teacher DE policy," which could be either interpreted as "the policy that respects teacher" (RC interpretation) or "the policy about respecting teachers" (appositive interpretation).

Because the SRN in Study 2 will not code for semantics, except for animacy, we focus on only the relativizer use of DE. This represents a simplification compared to the full natural language.

Summarizing from above, we can see that these relevant structures, some of which may well surpass the frequency of relative clauses in comprehenders' linguistic experiences, impose influences on the processing of both the typical subject relative clauses with transitive verbs and object relative clauses, at different main clause site.

## APPENDIX B

**Table B1 | Tgrep2 search patterns for structures in Study 1.**

| | |
|---|---|
| Simple sentences, transitive | /^IP/<(/^NP-SBJ/!</^-NONE-/!<<DEC)<(/^VP/<<VV<<(/^NP-OBJ/!</^-NONE-/!<<DEC))!>>/^IP/ |
| Simple sentences, intransitive | /^IP/<(/^NP-SBJ/!</^-NONE-/!<<DEC)<(/^VP/<<VV!<</^NP-OBJ/)!>>/^IP/ |
| Pro-drop sentences, transitive | /^IP/<(/^NP-SBJ/</^-NONE-/)<(/^VP/<<VV<<(/^NP-OBJ/!</^-NONE-/!<<DEC))!>>/^IP/ |
| Pro-drop sentences, intransitive | /^IP/<(/^NP-SBJ/</^-NONE-/)<(/^VP/<<VV!<</^NP-OBJ/)!>>/^IP/ |
| SRC, transitive | /^NP-SBJ/ (or /^NP-OBJ/ for object-modifying)<<(/^IP/<(/^NP-SBJ/<(/^-NONE-/!<*PRO*!<*pro*)<(/^VP/<<VV<< (/^NP-OBJ/!</^-NONE-/))$ (DEC../^NP/)) |
| SRC, intransitive | /^NP-SBJ/(or /^NP-OBJ/ for object-modifying)<<(/^IP/<(/^NP-SBJ/<(/^-NONE-/!<*PRO*!<*pro*)<(/^VP/<<VV!<</^NP-OBJ /))$ (DEC../^NP/)) |
| ORC | /^NP-SBJ/(or /^NP-OBJ/ for object-modifying)<<(/^IP/<(/^NP-SBJ/!</^-NONE-/)<(/^VP/<<VV<<(/^NP-OBJ/</^-NONE-/))$(DEC../^NP/)) |

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 19 November 2013
doi: 10.3389/fpsyg.2013.00828

# Self-organizing map models of language acquisition

***Ping Li*[1]* and *Xiaowei Zhao*[2]**

[1] *Department of Psychology and Center for Brain, Behavior, and Cognition, Pennsylvania State University, University Park, PA, USA*
[2] *Department of Psychology, Emmanuel College, Boston, MA, USA*

***Edited by:***
*Julien Mayor, University of Geneva, Switzerland*

***Reviewed by:***
*Colin Davis, Royal Holloway University of London, UK*
*Michael Thomas, Birkbeck College, UK*

***Correspondence:**
*Ping Li, Department of Psychology and Center for Brain, Behavior, and Cognition, Pennsylvania State University, 452 Moore Building, University Park, PA 16802, USA*
*e-mail: pul8@psu.edu*

Connectionist models have had a profound impact on theories of language. While most early models were inspired by the classic parallel distributed processing architecture, recent models of language have explored various other types of models, including self-organizing models for language acquisition. In this paper, we aim at providing a review of the latter type of models, and highlight a number of simulation experiments that we have conducted based on these models. We show that self-organizing connectionist models can provide significant insights into long-standing debates in both monolingual and bilingual language development. We suggest future directions in which these models can be extended, to better connect with behavioral and neural data, and to make clear predictions in testing relevant psycholinguistic theories.

**Keywords: SOM, connectionism, language acquisition, vocabulary spurt, lexical aspect, age of acquisition, cross-language priming**

## INTRODUCTION

The parallel distributed processing (PDP) models have stimulated tremendous interests in computational models of language and led to intense debates regarding the nature and representation of language. Today, more than a quarter century after the original PDP volumes (McClelland et al., 1986; Rumelhart and McClelland, 1986), connectionism has become a powerful tool as well as a conceptual framework for us to understand many important issues in language learning, processing, and impairment. According to the connectionist framework, many critical aspects of human cognition are emergent properties, and language is an example *par excellence*. Language as a hallmark of human behavior thus received in-depth treatment in the original PDP volumes, and connectionist language models have flourished in the last 25 years. It is important to note that these models may involve significantly different computational architectures, for example, with regard to both representation structures (e.g., localist vs. distributed representation) or learning mechanisms (supervised vs. unsupervised learning). In this article, we focus on a type of unsupervised connectionist learning models, the self-organizing maps (SOMs)[1], and illustrate ways in which SOM-based connectionist models can be used effectively to study the acquisition and processing of both first and second languages.

### SELF-ORGANIZING MAPS

In contrast to the classic PDP learning models (e.g., of the type learned via back-propagation), unsupervised learning models use no explicit error signals to adjust the weights between input and output. These models span a wide range of learning algorithms, including the Adaptive Resonance Theory (ART; Grossberg, 1976a,b; see Hinton and Sejnowski, 1999 for a collection of unsupervised models). Here we focus on a particular unsupervised learning algorithm called SOM (Kohonen, 2001), which has been widely used in modeling language learning and representation (see Li and Zhao, 2012 for a bibliography).

A standard SOM consists of a two-dimensional topographic map for the organization of input representations, where each node is a unit on the map that receives input via the input-to-map connections. At each training step of SOM, an input pattern (e.g., the phonological or semantic information of a word) is randomly picked out and presented to the network, which activates many units on the map, initially randomly. The SOM algorithm starts out by identifying all the incoming connection weights to each and every unit on the map, and for each unit, compares the weight vector with the input vector. If the unit's weight vector and the input vector are similar or identical by chance, the unit will receive the highest activation and is declared the "winner" (the Best Matching Unit or BMU, see **Figure 1** for an example). Once a unit becomes highly active for a given input, its weight vector and that of its neighboring units are adjusted, such that they become more similar to the input and hence will respond to the same or similar inputs more strongly the next time. In this way, every time an input is presented, an area of nodes will become activated on the map (the "activity bubbles") and the maximally active nodes are taken to represent the input.

Equation 1 shows how the activations of the nodes on the map are calculated. Considering a node $k$ that has a vector $m_k$ associated with it to represent the weights of the input connections to it. Given an input vector $\mathbf{x}$ (e.g., the phonological or semantic information of a word), the localized output response $\alpha$ of node $k$ is computed as:

$$\alpha_k = \begin{cases} 1 - \dfrac{\|x - m_k\| - d_{\min}}{d_{\max} - d_{\min}}, & \text{if } k \in N_c \\ 0, & \text{otherwise} \end{cases}, \qquad (1)$$

where $N_c$ is the set of neighbors of the winner $c$ [for which $\alpha_c = \max_k(\alpha_k) = 1$], $d_{\min}$ and $d_{\max}$ are the smallest and the largest Euclidean distances of $\mathbf{x}$ to node's weight vectors within

[1] There are other connectionist models that have neither PDP nor SOM architectures that have been applied to language studies (see Davis, 1999; Bowers, 2002; see Li and Zhao, 2012 for a bibliography).

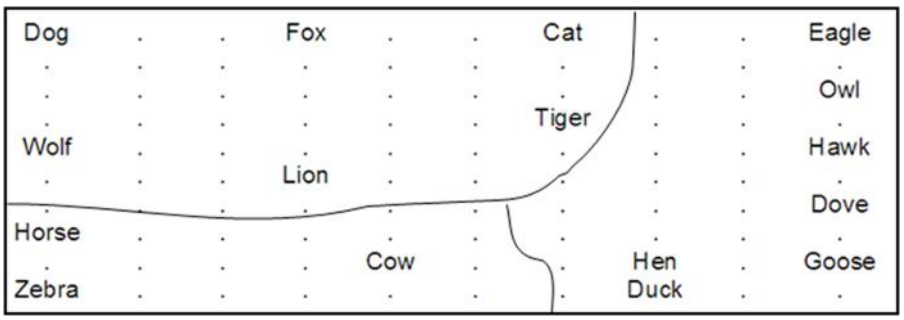



**FIGURE 1 | An illustration of the learned semantic categories in a SOM model.** Concepts with similar features/attributes are grouped together such as *horse* and *zebra*.

$N_c$. Initially activation occurs in large areas of the map, that is, large neighborhoods, but gradually learning becomes focused and the size of the neighborhoods reduces to only one node (the winner), which has an activation level of one. This process continues until all the input patterns elicit specific response units in the map (i.e., the BMUs).

As a result of this self-organizing process, the statistical structure implicit in the input is captured by the topographic structure of the SOM. In this newly formed topographic structure (the new representation), similar inputs will end up activating nodes in nearby regions, yielding meaningful activity bubbles that can be visualized on the 2-D space of the map. Equation 2 shows how the weights of the nodes around a winner or BMU are updated:

$$m_k(t+1) = m_k(t) + \beta(t) \cdot [x - m_k(t)] \text{ for all } k \in N_c. \quad (2)$$

Here, $\beta(t)$ is the learning rate for the map, which changes with time $t$. If the node $k$ belongs to the nodes in the neighborhood of the winner $c$, its weight should be adjusted according to this equation; otherwise, it remains unchanged.

Self-organizing maps have several important properties that make them particularly well suited to the study of language acquisition. First, as unsupervised learning systems, SOMs require no explicit teacher; learning is achieved by the system's organization in response to the input. Such networks provide computationally relevant models for language acquisition, given that in real language learning children do not receive constant feedback about what is incorrect in their speech (such as the kind of error corrections provided by supervised learning algorithms). Second, self-organization in these networks allow for the gradual formation of structures as changes of activity bubbles on 2-D maps, as a result of extracting an efficient representation of the complex statistical regularities inherent in the high-dimensional input space (Kohonen, 2001). In particular, the network organizes information first in large areas of the map and gradually zeros in on to smaller areas (decreasing neighborhoods); this zeroing-in is a process from diffuse to focused patterns, as a function of the network's continuous adaptation to the input characteristics. Third, the SOM can fall into a topography-preserving state once learning is achieved, which means nearby areas in the map respond to inputs with similar features. This property is consistent with known topographic features of certain areas of the brain where topographic maps are formed, especially in the primary motor, visual, and somatosensory cortical areas (Haykin, 1999; Spitzer, 1999; Miikkulainen et al., 2005). Although the association cortex in the human brain may be much more dynamic and less topographically organized (see Sporns, 2010), the SOM architecture does allow researchers to model the emergence of higher-level cognitive processes (see Miikkulainen, 1993, 1997), including the emergence of lexical categories as a gradual process and natural outcome of language learning (see Li, 2003).

As in other computational models, training in a SOM involves the use and manipulation of free parameters such as map size (number of nodes in the network), neighborhood size (initial radius of the training nodes), learning rate, etc. Appropriate values of these parameters often lead to fast convergence of training or better overall performance of the model, and decisions need to be made by the modeler in advance, given the nature and complexity of the modeling task. As a general yardstick, for example, the size of the map should generally be three to four times the size of the input units to be learned, whereas the size of the initial radius should be sufficiently large (e.g., 1/4 of the map size) to allow for reorganization of the map's topography, depending on how much plasticity the modeler wants to give to the network[2]. Both neighborhood size and learning rate in most SOM models take a linear decrease function (e.g., Miikkulainen, 1997), although some studies have tied the change of neighborhood size to quantization errors of the network (see Li et al., 2007 and discussion below). The modeler must constantly evaluate the impact of different values of the free parameters in affecting the performance of the model and speed of convergence. As an example, Richardson and Thomas (2008) systematically examined the influence of the free parameters, along with some other factors, on SOM's ability to simulate critical periods in cognitive development.

### HEBBIAN LEARNING RULE

A highly influential learning mechanism that has also been computationally implemented in connectionist models is the Hebbian learning rule, due to the Canadian neuroscientist Donald Hebb

[2]These are based on our modeling experience and the modeler needs to consider their utility in light of the task and complexity of input–output relationships in each simulation.

(Hebb, 1949). In considering how biological neural networks could work, Hebb hypothesized that when neuron A is persistently and repeatedly engaged in exciting neuron B, the efficiency of A in firing B will be increased due to some growth process or metabolic changes taking place in one or both neurons. In other words, the strength between A and B is increased as a result of their frequent associations in neural activities. Hebb's hypothesis is the basis of the slogan "neurons that fire together wire together." It provides an important background for connectionism, as well as a biologically plausible mechanism for how associative learning and memory could occur at the neural level, as it is clearly related to long-term potentiation (LTP) in biological systems (Munakata and Pfaffly, 2004). Mathematically, the Hebbian learning rule can be expressed as Eq. 3:

$$\Delta w_{kl} = \beta.\alpha_k.\alpha_l, \tag{3}$$

where $\beta$ is a constant learning rate, and $\Delta w_{kl}$ refers to change of weights from input $k$ to $l$ and $\alpha_k$ and $\alpha_l$ the associated activations of neurons $k$ and $l$. The equation indicates that the connection strengths between neurons $k$ and $l$ will be increased as a function of their concurrent activities.

Although the Hebbian learning rule itself is not explicitly included in the SOM algorithm discussed above, it has been a very useful mechanism for connectionist language models based on SOM. In particular, several SOMs can be linked together via associative connections trained by Hebbian learning (see Miikkulainen, 1993, 1997 for this approach). As shown in several models discussed next, when Hebbian learning is incorporated, the SOM model has strong implications for language acquisition: it can account for the process of how the learner establishes relationships between word forms, lexical semantics, and grammatical morphology, on the basis of how often they co-occur and how strongly they are co-activated in the representation.

## SOM-BASED CONNECTIONIST MODELS OF LANGUAGE

### MODELS WITH SINGLE SOMs

Many early SOM-based connectionist models include just one layer of SOM, which usually only accounts for a particular aspect of language in which the researchers are interested. A good example is the classic work of Ritter and Kohonen (1989) that demonstrates that SOM networks can capture the semantic structure of words. These authors tested a single SOM with inputs representing the meaning of words that were generated from two methods. The first method is a feature-based method, according to which a word's meaning is represented by a vector and each dimension of this vector represents a possible descriptive feature or attribute of the concept. The value of the dimension could be 0 or 1, indicating the absence (0) or presence (1) of a particular feature for the target word. For example, the representations of *dove* and *hen* are very similar except one dimension representing the flying feature (dove = 1, hen = 0). Specifically, Ritter and Kohonen (1989) generated a detailed representation of 16 animals based on 13 attributes, trained a SOM with the 16 animal words, and found that the network was able to form topographically organized representations of semantic categories associated with the 16 animal words; see an example in **Figure 1**.

Ritter and Kohonen's (1989) second method of representing meanings of words is a statistics-based method, according to which the researchers generated a corpus consisting of three-word sentences randomly formed from a list of nouns, verbs, and adverbs (e.g., *Dog drinks fast*). A trigram window is applied to the corpus, and the co-occurrence frequencies of the word in the middle of the trigram with its two closest neighbors are calculated. This generates a co-occurrence matrix, which forms the basis of each word's "average context vector," a combination of the average of all the words preceding the target word and that of all the words following it. The researchers then used these vectors as input to the SOM, and training on the SOM again indicated topographically structured semantic/grammatical categories on the map, like the example shown in **Figure 1**. Ritter and Kohonen's (1989) pioneering work clearly shows that categories implicitly in the linguistic environment (input streams) can be extracted by the SOM[3]. The properties of a topographic-preserving map as demonstrated by Ritter and Kohonen (1989) provide the basis for SOM as a model to simulate empirical findings regarding semantic representation and semantic priming (see Spitzer, 1999 and discussions of SEMANT and DevLex-II below).

Silberman et al. (2007) introduced the SEMANT model to simulate the associations of words/concepts in a semantic network. SEMANT includes one SOM that handles semantic information that was extracted from large-scale corpora based on the method of Li et al. (2000) using the CHILDES database (MacWhinney, 2000). SEMANT also integrates a component of episodic memory simulated by the lateral connections among the units on the SOM. The basic idea here was that the semantically related words tend to occur together in a linguistic content, and therefore their episodic associations tend to be strong. Simulation results for the model showed that SEMANT was able to replicate the empirical findings from psycholinguistics, such as effects of semantic priming that indicate faster response to the related word than unrelated words (e.g., faster lexical decision times for *nurse* than to *bread* upon hearing *doctor*; Neely and Durgunoglu, 1985).

In addition to semantic learning, SOM networks have also been used for simulating phonological development. For example, Guenther and Gjaja (1996) introduced a single-layer SOM to simulate the "perceptual magnet effect" (Kuhl, 1991) in infants' phonetic learning. In particular, the authors first trained the SOM (or "auditory map" as so named by the authors) with input sound patterns which contained formant information of different phonemic categories (such as /r/ and /l/ in American English). They then presented the network with test sounds similar to the phonemes that the network was trained on. When a test stimulus was presented to the map, the activities of all nodes on the map were calculated. Each node's activity level was further used to multiply its "preferred stimulus" (the input vector that activated the node most strongly), and the resulting products for all the nodes were added together and normalized to serve as the map's "population vector." Guenther and Gjaja (1996) measured

[3]Based on this method, Zhao et al. (2011) developed a software (Contextual Self-Organizing Map Package) to derive corpus-based semantic representations in multiple languages using word co-occurrence statistics.

the population vectors that corresponded to each test stimulus, and used these measures to represent the map's overall perception of the particular test stimulus. Consistent with results from empirical studies of human listeners, the SOM-based modeling results showed a warpping of perceptual space, that is, the acoustic patterns near the center/prototype of the learned sound categories are perceived as more similar to each other than to those patterns further away from the center, a trademark of "perceptual magnet effect."

#### MODELS WITH MULTIPLE SOMs

Although connectionist models with only one layer of SOM have be successful in simulating different aspects of language one at a time (e.g., semantic learning or phonological learning), researchers have realized that in natural language contexts the user or learner is engaged in a process in which phonological, lexical, semantic, and orthographic information often occurs simultaneously. An integrated SOM-based model must be able to simulate this process. Multiple SOMs that are interconnected have thus been developed in response to this requirement.

One of the earliest attempts to construct a full-scale multiple SOM language model was Miikkulainen (1997; see also Miikkulainen, 1993). He introduced the DISLEX model, which includes different SOMs connected through associative links via Hebbian learning. In DISLEX, each map is dedicated to a specific type of linguistic information (e.g., orthography, phonology, and semantics), and is trained as a standard SOM. In the training of the network, an input pattern activates a node or a group of nodes on one of the input maps, and the resulting activity bubble propagates through the associative links and causes an activity bubble to form in the other map. The activation of co-occurring lexical and semantic representations leads to continuous organization in these maps, and most importantly, to adaptive formations of associative connections between the maps. The DISLEX model was successfully used to simulate certain impaired language processes such as dyslexia and aphasia (Miikkulainen, 1997), and has also been applied to simulate bilingual representation (Miikkulainen and Kiran, 2009), bilingual aphasia (Kiran et al., 2013), and the acquisition of Chinese reading by elementary school children (Xing et al., 2004).

Using the basic idea of multiple SOMs connected via associative links, Li et al. (2004) developed the Developmental Lexicon (DevLex) model to simulate children's early lexical development. Similar to DISLEX, DevLex is a multi-layer self-organizing model with cross-layer links trained by Hebbian learning. Unlike DISLEX that uses the standard SOM learning algorithm, it includes two growing SOMs which recruit additional nodes in response to task demands in learning, and these new nodes are inserted in the topographic structure of the existing map as the network's learning progresses. The growth of new nodes is dependent upon accuracy of learning (e.g., as more errors occur more nodes are inserted). One growing maps handles the semantic representation and another the phonological representation of words. DevLex takes advantages of the SOM properties discussed earlier (see Introduction). The Li et al. (2004) simulations showed that it developed topographically organized representations for linguistic categories over time, modeled lexical confusion as a function of word density and semantic similarity, and displayed age-of-acquisition effects in the course of learning a growing lexicon. These results matched up with patterns from empirical studies of children's early lexical development. DevLex later evolved into the DevLex-II model, which we will discuss in the next section (Li et al., 2007).

Mayor and Plunkett (2010) introduced a self-organizing model to account for fast mapping in early word learning. Their simulations particularly focused on the generalization property of word–object associations based on the taxonomic/categorical relationship of objects. Their model included two SOMs with one receiving visual input (the *object*) and another acoustic input (the *word*). The connections of the two maps were adjusted by Hebbian learning rule, which emphasizes that the cross-layer weights are reinforced as the object and the word are simultaneously presented to their model. Mayor and Plunkett (2010) pointed out that this joint presentation of an object and its corresponding name reflected the results of the development of infants' joint-attentional activities with their caregivers. Although the visual inputs to this model were random dot matrices artificially generated, the model could simulate several interesting patterns in children's early lexical and category development, such as the taxonomic constraint that indicates children tend to give a new object a known name in the same category (e.g., seeing a tiger for the first time and immediately call it a cat, which they already learned). The authors also argued that an efficient, pre-established categorization capacity is a prerequisite to successful word learning. This argument is highly consistent with data from other simulations of early lexical development such as those based on DevLex-II (see discussion under the Section "Modeling Vocabulary Spurt").

Recently, Kiran et al. (2013) presented data from the DISLEX model that simulated patterns of bilingual language recovery in aphasic patients. A distinct feature of the Kiran et al.'s (2013) model was that they applied it to simulate empirical patterns of, on a case-by-case basis, each of the 17 patients who underwent treatment following injury. Their simulation results showed a close match with real behavioral data from individual patients, and this is a testimony that computational models can closely reflect realistic linguistic processes from realistic language users (rather than from simplified or idealized situations). In order to do so, Kiran et al.'s (2013) model incorporated important variables underlying patterns of behavior, including the patient's language history with regard to age of L1 and L2 acquisition, proficiency, and the dominance of the treatment language. More impressive was the model's ability to predict the efficacy of rehabilitation in each of the bilingual's languages. In reality, each bilingual patient underwent rehabilitation treatment for only one of their languages (English or Spanish) due to empirical constraints, but the computational model was trained for recovery in both languages following lesion, thus showing considerable advantage and flexibility of the model as compared with examination of the actual patient. It is important to note that in empirical studies the researcher, when faced with the injured patient, cannot go back to study the patient's pre-lesion condition, whereas in computational modeling the researcher can examine the intact model, lesion it, and track the performance of the same model

before and after lesion, as was done by Kiran et al. (2013) in their study.

## DevLex-II: A SCALABLE SOM-BASED CONNECTIONIST MODEL OF LANGUAGE

In this section, we present the details of the DevLex-II model (Li et al., 2007), a SOM-based connectionist model designed to simulate processes of language learning in both the monolingual and bilingual situations. In a number of studies (Zhao and Li, 2009, 2010, 2013), we have tested the model's ability in accounting for patterns of first (L1) and second (L2) language acquisition. We say that the model is "scalable" because it can be used to simulate a large realistic linguistic lexicon, in single or multiple languages, and for various bilingual language pairs (such as Chinese–English, Spanish–English, etc.). In what follows we will first discuss some key features of the DevLex-II architecture and then discuss the applications of the model to various L1 and L2 phenomena to illustrate how models based on multiple SOMs can be used effectively to address critical issues in L1 and L2 acquisition.

### ARCHITECTURE OF THE MODEL

Considering the features of previous models (DISLEX, DevLex), the DevLex-II model builds on the basic structure as described above: multiple SOMs which are connected via Hebbian learning. The architecture of the model is illustrated in **Figure 2**. The model includes three basic levels for the representation and organization of linguistic information: phonological content, semantic content, and the articulatory output sequence of the lexicon. The core of the model is a SOM that handles lexical-semantic representation. This SOM is connected to two other SOMs, one for input (auditory) phonology, and another for articulatory sequences of output phonology. Upon training of the network, the semantic representation, input phonology, and output phonemic sequence of a word are simultaneously presented to the network. This process can be analogous to that of a child hearing a word and performing analysis of its semantic, phonological, and phonemic information.

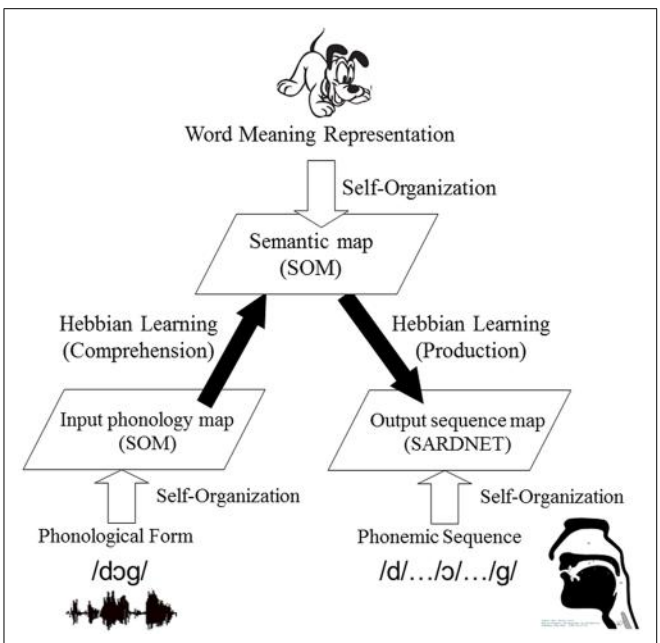



**FIGURE 2 | A sketch of the DevLex-II model.** See text for explanation of the architecture of the model.

On the semantic and phonological levels, the network constructs the representations based on the corresponding linguistic input according to the standard SOM algorithm. On the phonemic output level, DevLex-II uses an algorithm called SARDNET (James and Miikkulainen, 1995), a SOM-based temporal sequence learning network. The addition of the SARDNET algorithm to the model is based on considerations that word production is a temporal sequence ordering problem, and that language learners face the challenge to develop better articulatory control of the phonemic sequences of words.

In this architecture, at each training step, phonemes are input into the sequence map one by one, according their order of occurrence in a word. The winning unit of a phoneme is found and the weights of nodes in its neighborhood are adjusted; meanwhile, the activation levels of the winners responding to phonemes preceding the current phoneme will be adjusted by a number $\gamma^d$, where $\gamma$ is a constant and $d$ is the distance between the location of the current phoneme and the preceding phoneme that occurred in the word. This adjustment is intended to model the effect of phonological short-term memory during the learning of articulatory sequences; the activation of the current phoneme could be accompanied by some rehearsal of previous phonemes due to phonological short-term memory, which deepens the network's or the learner's knowledge of previous phonemes. The $\gamma$ here is chosen to be $<1$ (0.8 in our case), in order to model the fact that phonological short-term memory tends to decay with time. For a word with $n$ phonemes, the output of the winner responding to the $j^{\text{th}}$ phoneme will be $1 + \gamma + \gamma^2 + \ldots\ldots + \gamma^{n-j}$, which is a geometric progression, and can be written as:

$$\alpha_{\text{winner}}(j) = \frac{(1-\gamma^{n-j+1})}{1-\gamma}. \quad (4)$$

According to this equation, when the representation of all the phonemes in a word is received by the output sequence map, the activation of some nodes (e.g., the first winner) will be larger than 1, so they need to be normalized to the range between 0 and 1. Thus, the node in response to the first phoneme of the word will have the largest activation, followed by sequentially decaying activations of other phonemes in the sequence of the word.

In DevLex-II, the associative connections between maps are trained via the Hebbian learning rule, as in DevLex and DISLEX. As training progresses, the weights of the associative connections between the frequently and concurrently activated nodes on two maps will become increasingly stronger. After the cross-map connections are stabilized, the activation of a word form can evoke the activation of a word meaning via form-to-meaning links, which models word comprehension. If the activated unit on the semantic map matches the correct word meaning, we determine that the network correctly comprehends this word; otherwise the network makes a comprehension error. Similarly, the activation of a word meaning can trigger the activation of an output sequence

via meaning-to-sequence links, which models word production. If the activated units on the phonemic map match the phonemes making up the word in the correct order, we determine that the network correctly produces this word; otherwise the network makes a production error.

## PLASTICITY AND STABILITY IN THE MODEL

Since we aim at designing a scalable model that is suitable to simulate learning in different linguistic contexts (monolingual and bilingual), we must consider a fundamental problem called "catastrophic interference" (see French, 1999 for a review). Keeping the network's plasticity for learning new words often causes it to lose its stability for old knowledge; conversely, a network that is too stable often cannot adapt itself very well to the new learning task. This problem has been termed the "plasticity–stability" dilemma in neural networks since the 1970s (Grossberg, 1976a,b). To resolve this problem for our studies (in particular the bilingual learning studies discussed below), we introduced two new features into DevLex-II.

The first is a self-adjustable *neighborhood function*. In the standard algorithm of SOM, the radius of the neighborhood usually decreases according to a fixed training timetable (see earlier discussion on SOM modeling parameters). This type of development in the network, though practically useful, is subject to the criticisms that (1) learning is tied directly and only to time (amount) of training, but is independent of the input-driven self-organizing process; and (2) the network often loses its plasticity for learning new inputs when the neighborhood radius becomes very small. DevLex-II considered these potential problems by using a learning process in which the neighborhood size is not totally locked with time, but is adjusted according to the network's learning outcome (experience). In particular, neighborhood function depends on the network's error level on each layer averaged across all the input patterns. Here, a "quantization error" (as used in Kohonen, 2001) of an input pattern is defined as the Euclidean distances of the input pattern to the weight vector of its winner or BMU.

A second way in which we have attempted to solve the plasticity–stability problem is to manage the training process to be more realistic for learning: for the input phonology map and the semantic map, during each training epoch, once a unit is activated as a BMU, it will become ineligible to respond to other inputs in the current training epoch. In this way, the old words are kept untouched in the training, whereas the new words can be represented by new units on the maps. A similar procedure is also used for the output sequence map at the word level, where the same phoneme in different locations of a word will be mapped to different, but adjacent, nodes on the map. This mechanism resembles DevLex's growing map process in which new nodes are recruited for novel inputs as computational resources become scarce (see Li et al., 2004 for an algorithm in new node recruitment, as also discussed earlier). The use of a different but adjacent unit for new input is also empirically plausible: psycholinguistic research suggests that when young children encounter a novel word they tend to map it to a different category or meaning for which the child has not yet acquired a name (Markman, 1994; Mervis and Bertrand, 1994).

## LINGUISTIC REALISM OF THE MODEL

Many connectionist models of language are based on the use of artificially generated lexicon that is often limited in size. Such use of synthetic or highly simplified vocabularies provides certain modeling conveniences, but it lacks linguistic realism and is out of touch with the learner's true lexical experience. As a step forward, we considered two methods in which our modeling data was constructed. First, in all of our studies with DevLex-II (Li et al., 2007; Zhao and Li, 2009, 2010, 2013), we used input based on realistic linguistic stimuli. For example, in several studies our simulation material was based on the vocabulary from the MacArthur–Bates Communicative Development Inventories (the CDI; Dale and Fenson, 1996), which allowed us to model a lexicon size of up to 1000 words. Second, we coded the input to our model as vector representation of the phonemic, phonological, or semantic information of words, extracted in the following ways: (1) PatPho, a generic phonological representation system, was used to generate the sound patterns of words based on articulatory features of different languages (Li and MacWhinney, 2002; Zhao and Li, 2010); (2) statistics-based methods were used to generate semantic representations of training stimuli from large-scale corpus data (e.g., CHILDES database; MacWhinney, 2000) or from computational thesauruses (e.g., WordNet database; Miller, 1990), as done in Li et al. (2007) and Zhao and Li (2009, 2010, 2013; see also earlier discussion about generating semantic representations in SOM-based models). Given the input representations constructed in the above manner, the DevLex-II model receives each representation sequentially in the training (i.e., one word at a time, in randomized order of training), approximating the word learning process in the realistic learning environment.

## DevLex-II MODELS OF MONOLINGUAL LANGUAGE ACQUISITION

Many interesting empirical phenomena have been examined in the field of monolingual language development; for example, in the study of lexical development, researchers have investigated patterns such as the vocabulary spurt, age of acquisition (AoA) of vocabulary, relationship between comprehension and production, motherese and role of input, word frequency effect, lexical category development, fast mapping, lexical overextension, U-shaped development, and so on (see Clark, 2009; Saxton, 2010). Connectionist approaches have been fruitfully applied to study these phenomena in the past two decades (see Westermann et al., 2009 for a review; see Li and Zhao, 2012 for a bibliography). The DevLex-II model was originally designed to account for several of the phenomena listed above.

### *Modeling vocabulary spurt*

Vocabulary spurt refers to a period of extremely rapid growth of vocabulary starting around 18–24 months of age in children. A large number of studies have examined vocabulary spurt in young children (see Goldfield and Reznick, 1990; Bates and Carnevale, 1993 for example). Despite the empirical research in documenting the outcome of timing of vocabulary spurt, the underlying mechanisms for when and how vocabulary spurt occurs has been an issue of intense debate. To provide a computational account of this phenomenon, Li et al. (2007) trained a DevLex-II model

to learn 591 English words extracted from the toddler's vocabulary list of the English CDI. Their model incorporated several key features of learning and representation for lexical development, including the multiple SOMs that were used for simulating comprehension and production for the same items, along with realistic phonological and semantic input patterns of the lexical items.

**Figure 3** presents the average receptive and productive vocabulary sizes across the course of DevLex-II training, averaged across 10 simulation trials. The simulation results show a clear vocabulary spurt, preceded by a stage of slow learning and followed by a performance plateau. On average, the model's productive vocabulary did not accelerate until about 35–40 epochs, one-third into the total training time, reflecting the model's early protracted learning of the representations of word forms, meanings, and sequences, and their associative connections. Once the basic organization of the lexicon was acquired in terms of lexical and semantic categories and their associations, vocabulary learning accelerated, which occurred quickly after 40 epochs of training.

Although the figure shows only the results of the associative connections (form-to-meaning for comprehension, and meaning-to-sequence for production), the hit rates for these connections depended directly on the shape or precision of self-organization in the separate feature maps (see Figure 3 on Li et al., 2007). In other words, the period of rapid increase in vocabulary size may have been prepared by the network's slow learning of the structured representation of phonemic sequence, word phonology, and word semantics, as well as its learning of the mappings between these characteristics of the lexicon. Once the basic structures were established on the corresponding maps, the associative connections between maps could be reliably strengthened to reach a critical threshold through Hebbian learning.

**Figure 3** also shows that the vocabulary spurt occurred for both production and comprehension, rather than being restricted to only one modality, consistent with empirical studies (Reznick and Goldfield, 1992). Previous empirical studies have largely focused on children's word production, but a few researchers have also questioned whether a comprehension vocabulary spurt could exist. Our DevLex-II model was able to simulate a spurt pattern in both comprehension and production. Interesting, although both types of spurt were present in our simulations, the comprehension spurt occurred earlier than the production spurt, which is consistent with the argument that comprehension generally precedes production (Clark and Hecht, 1983) and in the case of lexical acquisition, a spurt in the receptive vocabulary could start much earlier (e.g., from 14 months of age; see Benedict, 1979).

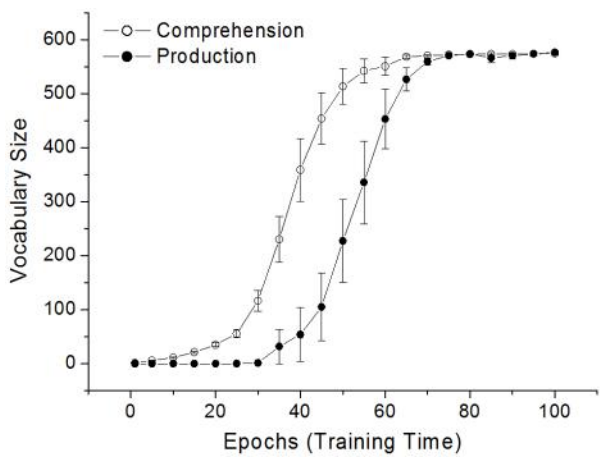



**FIGURE 3 | Vocabulary spurt in the learning of the 591 CDI words by DevLex-II.** Results are averaged across 10 simulation trials. Error bars indicate standard error of the mean (figure adapted from Li et al., 2007, reproduced with permission from Wiley and Sons, Inc.).

Our simulation results as shown in **Figure 3** also indicated that there were significant individual differences between different simulation trials, even when all simulations had the same modeling parameters. Most interestingly, the largest variations tended to coincide with the rapid growth or spurt period. Examining the individual trials in detail, we found that different simulated networks could differ dramatically in the onset time of their vocabulary spurt. In the 10 simulation trials, the rapid increase of vocabulary size in production could begin from as early as epoch 30, or from as late as epoch 60, but in each case there was a clear spurt process. While some of these variations may be random effects (due primarily to the network's random initial weight configurations and the random order of training words), others were systematic differences as a function of learning the complexity of the lexical input, especially the different stimulus properties such as word frequency and word length. Our simulation results clearly indicate the higher the word frequency, or the shorter the word length, the earlier the vocabulary spurt (see discussions in Li et al., 2007, Section 3.3).

### *Modeling early phonological production*

DevLex-II is also able to simulate patterns in early phonological production. **Table 1** presents a list of typical examples from the same network discussed above on word productions at different training times. These errors parallel children's early word pronunciations (Foster-Cohen, 1999), such as omission of consonants at the end of a word (e.g., output for "bib" at epoch 50), deletion of a consonant from consonant clusters (e.g., output for "smile" and "glue" at epoch 60), and substitution of consonants with similar phonemes (e.g., producing /b/ for /d/ in "bird"). These errors can be attributed to (a) incomplete meaning-to-phoneme links, and (b) incomplete sequence learning of phonemes. Our modeling results showed clearly how children's early phonemic errors can arise from incomplete consolidation of word sequences, amplified by limitations in the learner's phonemic memory.

An examination of **Table 1** also indicates other interesting patterns. For example, in two different simulation trials, responding to the word *sock*, the network gave two different patterns of production error, the deletion of consonant /k/, and the substitution of it with /t/ (see **Table 1** for the two cases of *sock*). Given that the simulation trials had the same training parameters with the only difference in initial weights and training order of words, this difference reflects individual variations that are similar to those found both within and across different developmental stages in children (Menn and Stoel-Gammon, 1993).

**Table 1 | Sample production errors from DevLex-II in learning English vocabulary.**

| Words | Epochs (training time) | | | | | |
|---|---|---|---|---|---|---|
| | **30** | **40** | **50** | **60** | **80** | **100** |
| Foot (/fʊt/) | /ʋuː/ | /fʊ/ | /fʊ/ | /fʊ/ | /fʊt/ | /fʊt/ |
| Dog (/dɔːg/) | – | /d/ | /d/ | /dɔː/ | /dgɔː/ | /dɔːg/ |
| Sock (/sɑːk/) | /ʃd/ | /suː/ | /suː/ | /sɑː/ | /sɑː/ | /sɑː/ |
| Bib (/bɪb/) | /ɑː/ | /br/ | /bɪ/ | /bɪ/ | /bɪb/ | /bɪb/ |
| Apple (/æpəl/) | – | /æp/ | /æp/ | /æp/ | /æpə/ | /æpəl/ |
| Cat (/cæt/) | /a/ | /cæ/ | /cæ/ | /cæ/ | /cæt/ | /cæt/ |
| Brush (/brəʃ/) | /n/ | /bən/ | /bən/ | /bəʃ/ | /bəʃr/ | /brəʃ/ |
| Smile (/smaɪl/) | /pəiiː/ | /ɪma/ | /ɪma/ | /maɪl/ | /maɪl/ | /maɪl/ |
| Glue (/gluː/) | /ɪ/ | /g/ | /g/ | /guː/ | /guː/ | /guː/ |
| Sock (/sɑːk/) | /a/ | /sɑːt/ | /sɑːt/ | /sɑːt/ | /sɑːt/ | /sɑːt/ |
| Hide (/haɪd/) | /ɪb/ | /ɪhb/ | /haɪb/ | /haɪb/ | /haɪb/ | /haɪb/ |
| Bird (/bɜːd/) | /ʋ(d;n)/* | /bb/ | /bɜːb/ | /bɜːb/ | /bɜːb/ | /bɜːb/ |
| Bottle (/bɑːtəl/) | /tæ/ | /btɑːəæ/ | /btɑːəæ/ | /btɑːəæ/ | /bɑːtəæ/ | /bɑːtəæ/ |
| Glasses (/glæsəz/) | /ts/ | /æziːn/ | /gæziːn/ | /gæsəz/ | /gæsəzl/ | /glæsəz/ |

*"–" indicates no unique output of the word since the word is confused with other words on the semantic map at the current time.*
**Both the phonemes /d/ and /n/ on the phonemic map were the best matching units (BMUs) in response to the semantic representation of "bird."*

Finally, most of the examples in **Table 1** also reflect a general developmental shift in phonological pattern formation. At the earliest stages of learning, the network's productions were highly simplified and often very different from the target pronunciations. During the middle and late stages of learning, with the emergence of self-organized phonemic structure and the developing associative links, correct productions increased gradually. At these stages, the production errors, though still present, were much closer to the target pronunciations but some also had typical error patterns as discussed above. The coexistence of correct and incorrect word pronunciations corresponds to empirical patterns in children's phonological development from babbling to word production (Menn and Stoel-Gammon, 1993; Foster-Cohen, 1999). The transition from incorrect sequences, omissions, and substitutions to correct pronunciations indicates that our model was able to capture developmental changes in early phonological acquisition.

### *Modeling the acquisition of grammatical and lexical aspect*

Linguists generally distinguish between two kinds of aspect, grammatical aspect and lexical aspect. Grammatical aspect is related to aspectual distinctions which are often marked explicitly by linguistic devices (e.g., English auxiliary *be* plus inflectional suffix *ing* to mark ongoing activities). Lexical aspect, on the other hand, refers to the characteristics inherent in the temporal meanings of a verb, for example, whether the verb encodes an inherent end point of a situation (e.g., telic verb like *arrive* vs. atelic verb like *run*), or whether the verb is inherently stative or punctual (stative verb like *believe* vs. punctual verb like *break*). Research has shown that there is a strong interaction between these two types of aspect in the process of children's early lexical and grammatical acquisition (see Li and Shirai, 2000 for a review); for example, initially children's use of grammatical inflections is restricted to specific verbs (e.g., using *-ed* only with punctual verbs), and only later on it approaches the adult linguistic pattern.

We wanted to see whether a multi-layer SOM-based model is able to capture the developmental patterns of child language in the acquisition of lexical and grammatical aspect, and whether a connectionist network void of pre-specified categories can acquire verb aspectual categories that have been claimed to be innate (cf. the "language bioprogram hypothesis" of Bickerton, 1984). To simulate the acquisition of aspect, DevLex-II was trained on 184 English verbs across four growth age stages (or input ages: 13–18, 19–24, 25–30, and 31–36 months old). Each of the 184 verb types was chosen if it occurred in the parental speech of CHILDES database (MacWhinney, 2000) for 50 or more times within a certain age group mentioned above (see Zhao and Li, 2009 for details). We examined the network's acquisition of imperfective/progressive aspect marker *ing*, habitual aspect markers *-s* and perfective aspect marker *-ed* in connection with the acquisition of three semantic categories of lexical aspects (*activity*, *telic*, and *stative* verbs). Here, we defined the correct production of the aspect form for any given verb in the same way as done in Li et al. (2007): for example, if the word *kicking* is shown to the semantic map, production is counted correct only when the consecutively activated nodes on the output phonemic map are the BMUs for /k/ /I/ /k/ /I/ /ŋ/ in the correct sequence.

**Table 2** presents a comparison of our simulation results with empirical patterns from parental speech. First, looking at the simulation data, we found that the network's production of inflectional markers across the four input ages are highly consistent with the empirical patterns: the use of imperfective aspect (*-ing*) is closely

**Table 2 | Percentage of use of tense-aspect suffixes with different verb types across input age groups in DevLex-II's production and in input data based on parental speech (adapted from Li and Zhao, 2009, reproduced with permission from Mouton de Gruyter).**

| Verbs | Tense-aspect suffixes | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Age 1;6 | | | Age 2;0 | | | Age 2;6 | | | Age 3;0 | | |
| | *-ing* | *-ed* | *-s* | *-ing* | *-ed* | *-s* | *-ing* | *-ed* | *-s* | *-ing* | *-ed* | *-s* |
| **Network production** | | | | | | | | | | | | |
| Activity | 73 | 0 | 29 | 69 | 27 | 33 | 61 | 24 | 35 | 62 | 30 | 37 |
| Telic | 27 | 75 | 14 | 21 | 53 | 28 | 32 | 62 | 27 | 31 | 62 | 26 |
| Stative | 0 | 25 | 57 | 10 | 20 | 39 | 7 | 14 | 38 | 7 | 8 | 37 |
| **Parental input data** | | | | | | | | | | | | |
| Activity | 63 | 23 | 29 | 62 | 26 | 26 | 63 | 22 | 33 | 60 | 29 | 35 |
| Telic | 31 | 62 | 29 | 31 | 58 | 26 | 29 | 66 | 25 | 32 | 59 | 24 |
| Stative | 6 | 15 | 43 | 7 | 16 | 48 | 8 | 12 | 42 | 8 | 12 | 41 |

associated with activity verbs that indicate ongoing process, while the use of perfective aspect (*-ed*) is closely associated with telic verbs that indicate actions with endpoints or end results. In particular, in early child English, *-ing* is highly restricted to activity verbs, *-ed* restricted to telic verbs, and *-s* restricted to stative verbs, as demonstrated by Bloom et al. (1980). Our network, having received input patterns based on parental speech from the CHILDES database, behaved in the same way as children do. For example, at input age 1;6, the network produced *-ing* predominantly with activity verbs (73%), *-ed* overwhelmingly with telic verbs (75%), and *-s* with stative verbs (57%). Such associations were observed at all four stages (especially for *-ing* and *-ed*), but they became attenuated over time.

Second, we analyzed the input dataset to our network (based on child-directed parental speech), and found that there was also a clear consistency between the input and the network's production. In the input data there are clear associations between *-ing* and activity verbs, *-ed* and telic verbs, and that these associations are strong throughout the four input ages, as shown also by Shirai (1991) in an empirical analysis. The degree to which the network's production matches up with the input patterns indicates that DevLex-II was able to capture the statistical co-occurrences relationship between lexical aspect (verb types) and grammatical aspect (verb morphology) in the input. While this is hardly surprising for a connectionist model, our results also indicate that DevLex-II's productions were not simply verbatim mimics of what's in the input by recording individual words and suffixes and their co-occurrence. This is important in that our network has derived (but not simply reproduced) the type–suffixes association patterns from the linguistic input. The simulation results showed that the associations between verb types and suffixes were stronger in the network's productions than they were in the input data received by the network, particularly for the early training stages (i.e., more restrictive associations between verb semantics and inflectional suffixes, for example, between telic verbs and *-ed*). DevLex-II at early stages behaved more restrictively than what is in the language input with respect to the correlations between lexical aspect and grammatical aspect, which matches up well with empirical observations from child language (see Li and Shirai, 2000 for review).

## DevLex-II MODELS OF BILINGUAL LANGUAGE ACQUISITION

While the above discussion highlighted two domains (vocabulary and grammatical morphology) in which DevLex-II was applied to first language (L1) acquisition (see Zhao and Li, 2013 for a full list of DevLex-II applications), the utility of the model as a general model of language acquisition has also been tested further in the study of second language (L2) acquisition. Below we discuss how DevLex-II has been applied to examine a range of key issues in bilingualism.

### *Modeling age-of-acquisition effects*

Much of the current debate about the nature of L2 learning and how it differs from L1 learning stems from the "critical period" hypothesis. Indeed, interests in the critical period hypothesis have led *Science* magazine in its 125th anniversary issue to designate the understanding of critical periods of language acquisition as one of the top 125 big science questions in all scientific domains of inquiry for the next quarter century (*Science*, vol. 309, July 1, 2005). Recent studies, however, have argued against the original account of Lenneberg (1967) that there is a biologically based critical period for language acquisition due to brain lateralization; instead, the evidence points to cognitive and linguistic mechanisms underlying the AoA effects seen with both L1 and L2 acquisition (see MacWhinney, 2012; Li, in press). For example, Johnson and Newport (1989) suggested that language learning in childhood confers certain cognitive advantages precisely because of the child's limited memory and cognitive resources (the "less is more" hypothesis; see also Elman, 1990). Hernandez and Li (2007) suggested that different sensorimotor processing and control characteristics could underlie child vs. adult learning and processing differences (the "sensorimotor integration hypothesis"; see also Bates, 1999). Finally, MacWhinney (2012) suggested that certain risk factors (e.g., entrenchment of L1, negative transfer, and social isolation) with late learners but

not early learners could be responsible for the age-related learning effects in language acquisition (the "unified competition model" hypothesis).

In an effort to provide computational insights into the AoA effects, Zhao and Li (2010) applied the DevLex-II model to 1000 words, 500 in Chinese as L1 and 500 in English as L2, selected from the CDI database (Dale and Fenson, 1996). These words were presented to the model systematically in three different learning scenarios: simultaneous learning of L1 and L2; early L2 learning; and late L2 learning. For simultaneous learning, the two lexicons were presented to the network and trained in parallel (see Li and Farkaš, 2002 for a previous example in this training regime). For early L2 learning, the onset time of L2 input to the model was slightly delayed relative to that of L1 input, that is, training on L2 vocabulary occurred at a point after 1/5 of the entire L1 vocabulary had been presented to the network. For late L2 learning, the onset time of L2 input was significantly delayed relative to that of L1, that is, training on L2 vocabulary occurred at a point after 4/5 of the entire L1 vocabulary had been presented to the network. Specifically, the simultaneous learning situation is analogous to a situation in which children are raised in a bilingual family and receive linguistic inputs from the two languages simultaneously (e.g., Li and Farkaš, 2002 used input based on the two parents' different language input). The early learning situation could be compared to the situation in which bilinguals acquired their L2 early in life (e.g., in early childhood) while the late learning situation to that of a bilingual's learning of L2 later in life (e.g., after puberty).

One key pattern from our simulations is illustrated in **Figure 4**, which shows how lexical representations from the two languages are distributed differently in the three different learning conditions. Here, black regions indicate those nodes that represent the L2 (English) words whereas white regions the L1 (Chinese) words learned by the model. Specifically, if a unit's weight vector is the closest to the input vector of an English word, the unit is marked in black. If a unit's weight vector is most similar to the input pattern of a Chinese word, the unit is marked in white.

It is clear from **Figure 4** that the relative onset time of L2 vs. L1 plays an important role in modulating the overall representational structure of the L2[4]. For the simultaneous acquisition situation (**Figures 4A,B**), DevLex-II showed clear distinct lexical representations of the two languages at both the phonological and semantic levels and within each language. The results suggest that simultaneous learning of two languages allows the system to easily separate the lexicons during learning, consistent with the simulation patterns from Li and Farkaš (2002). In the case of sequential acquisition, if L2 was introduced into learning early on, the lexical organization patterns were similar (though not identical) to those found in simultaneous acquisition, as shown in **Figures 4C,D**. The differences were in terms of the slightly smaller spaces occupied by the L2 words (English) as compared to the lexical space occupied by L1[5]. The L2 lexicon was still able to establish its separate territory of lexical representation. However, if L2 was introduced to learning at a late stage, the lexical organization patterns were significantly different from those found in simultaneous acquisition, as shown in **Figures 4E,F**. No large independent areas for the L2 representation appeared this time. Indeed, the L2 representations appeared to be parasitic or auxiliary to those of L1 words: compared with L1 words, the L2 words occupied only small and fragmented regions, and were dispersed throughout the map. There were small L2 chunks that were isolated from each other, and interspersed within L1 regions. Interestingly, the parasitic nature of the L2 representation is also reflected in the locations of the L2 words in the map, which was determined by the similarity of the L2 words to the established L1 words in meaning (for semantic map) or in sound (for phonological map).

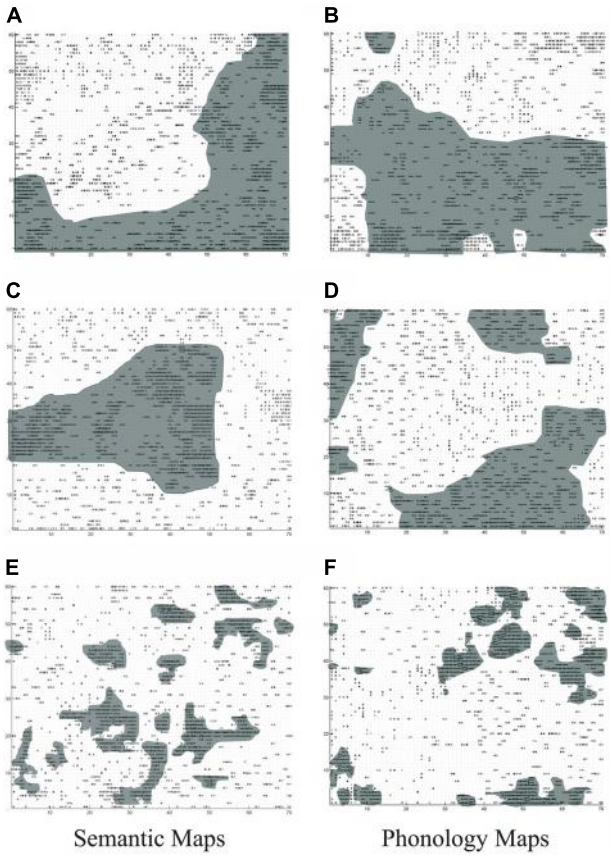



**FIGURE 4 | Bilingual lexical representation of semantics and phonology respectively as a function of age of acquisition (AoA).** Dark areas correspond to L2 (English) words. **(A,B)** Simultaneously learning; **(C,D)** early L2 learning; **(E,F)** late L2 learning.

The biologically based account of a critical period as originally put forth by Lenneberg (1967) is intuitively appealing, but the modeling results presented here indicate that we can simulate

[4]At this point we consider such segregations in the representation between L1 and L2 at the lemma level rather than a deeper semantic level, given the complexity associated with semantic and conceptual relations across languages (see Pavlenko, 2009 for a discussion).

[5]Similar results were obtained when English was trained as the L1 and Chinese the L2.

critical period-like effects without invoking any significant changes in the architecture or mechanisms in the network. A significant contribution of connectionist models to the understanding of development, according to Elman et al. (1996), is that these models do not involve pre-determined or pre-specified categories or underlying differences, and yet the simulated data show that categories or differences in these models emerge as a result of learning itself across a developing learning history. The "age" effects that were simulated in our model may reflect the changing dynamics inherent in learning and the interactions between the two languages across different types of learning situation. The idea that the learning process itself can lead to differences in the dynamics and outcomes of development is not new (see Elman et al., 1996; Thomas and Johnson, 2008).

What is new from our studies is that the age-related effects, traditionally attributed to inherent properties of the learner, emerged in our models as a result of learning the same L2 targets at different time points of learning. The effects of dynamic interactions in the two competing languages clearly speak for the perspective of competition, entrenchment, and plasticity in accounting for critical period effects (see General Discussion for more discussion).

### *Modeling cross-language priming*

One important goal of simulation is to provide a mechanistic account of the observed behavioral phenomena found in empirical studies (see, for example, Richardson and Thomas, 2008 for discussion). Capitalizing on the above findings of the impact of simulated age effects on bilingual lexical organization, Zhao and Li (2013) extended DevLex-II to simulate cross-language semantic priming in connection with the AoA effect. Cross-language priming has been a vital empirical method in the literature for testing semantic representations in bilinguals, and many studies have shown that in such a paradigm bilinguals respond faster to translation equivalents or semantically related words across languages than to unrelated pairs of words from the two languages (named as *translation priming* and *semantic priming*, respectively). Zhao and Li (2013) implemented a spreading activation mechanism in DevLex-II so that cross-language priming could be modeled. This mechanism involves two parts: (1) nodes on a map were fully connected with each other via lateral connections, and their weights were trained via Hebbian learning, triggered by the joint presentations of translation equivalents. This type of associative connections was added to DevLex-II specifically for modeling priming effects; (2) spreading activation from a prime word to a target word could occur via two paths, one through the lateral connections and one within the semantic map[6].

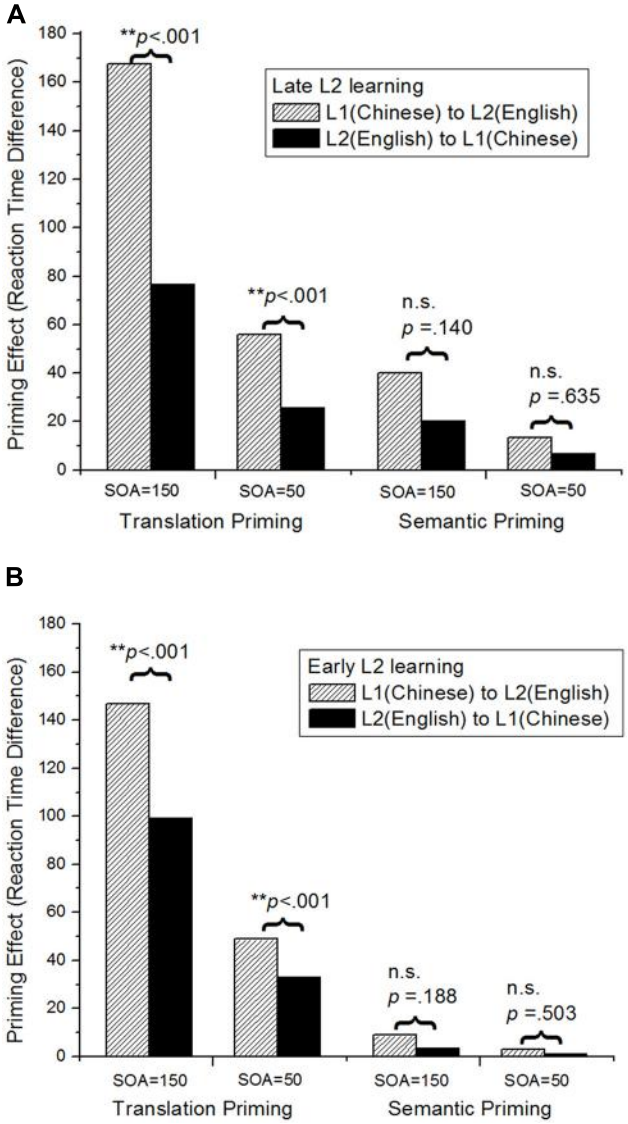



**FIGURE 5 | Cross-language priming and priming asymmetry effects shown in DevLex-II: (A) late L2 learning; (B) early L2 learning.** Priming effects are calculated by subtracting the RTs of related word pairs from the RTs of unrelated word pairs. SOA represents the stimulus onset asynchrony. The *p* values indicate the significance level of the priming asymmetry under the different conditions (paired-samples *t*-test of the 20 simulations under each condition: **significant priming asymmetry; n.s., not significant) (adapted from Zhao and Li, 2013, reproduced with permission from Cambridge University Press).

**Figure 5** presents the basic results of our simulations. The model clearly displayed both translation priming and semantic priming effects, although translation priming was always stronger than semantic priming, consistent with patterns from empirical studies (Basnight-Brown and Altarriba, 2007). Another important simulated pattern was the "priming asymmetry": in the empirical literature (see Dimitropoulou et al., 2011 for a review), it has been observed that priming effects are stronger if participants are presented with L1 words as primes and L2 words as targets (i.e., the L1-to-L2 direction of priming), as compared with the situation in which participants are presented with L2 words as primes and L1 words as targets (i.e., the L2-to-L1 direction of priming). As seen in **Figure 5**, the priming effects from L1 (Chinese) to L2 (English) were always larger than those from L2

[6]Zhao and Li (2013) also developed a mechanism to capture the time elapses in lexical decision tasks so that differences in reaction time (RT) could be modeled for cross-language priming (see technical details in that paper).

to L1. More interestingly, such "priming asymmetry" decreased as a function of the effect of AoA; for example, it was larger in the late L2 learning situation than in the early L2 learning situation.

DevLex-II provided a mechanistic account for this asymmetry, following the ideas discussed above regarding AoA effects, by reference to the richness of semantic representation of the L2 in our model (i.e., the number of activated semantic features that will lead to different degrees of priming from L2 primes to L1 targets). This account is particularly significant in light of the DevLex-II's emphasis on cross-language lexical competition. Specifically, the richness of semantic representation and the potential lexical competition are inversely related: the richer or more elaborated the representation of a word, the less competition (and hence less confusion) the learner may experience between the word and other lexical items in the memory. If the priming is from the L2 to L1 direction, due to the dense representation of L2 (thus strong competition), a brief exposure to L2 may not trigger initial activations strong enough to spread to the target L1 items not directly adjacent in the representation. In contrast, activations of L1 items could be much stronger given that they are more sparsely represented. If the priming is from the L1 to L2 direction, it will be always larger than the reverse, due to distinct (and richer) semantic representations of the L1 words (thus having less competition). Consequently, the amount of priming from L2 to L1 (and L1 to L2) may be enhanced or decreased, depending on a bilingual's L2 level as a function of AoA or proficiency, thereby giving rise to the different amount of "priming asymmetry." If the L2 is acquired at an early stage, its semantic representations are more enriched, and more distinct from L1 representations (rather than depending or being parasitic on L1 representations, as discussed earlier). In this case, the L2 to L1 priming effects are more comparable to the L1 to L2 priming effects given the similar representations of the two lexicons, thus cause a less salient priming asymmetry. Such an account has found empirical support in the semantic priming literature, in both the L1 and the L2 context (see discussions in Wang and Forster, 2010; Dimitropoulou et al., 2011).

## GENERAL DISCUSSION

In this article, we first reviewed previous connectionist models based on SOMs, and discussed the significant implications that SOM-based models have for understanding language representation and acquisition. We then described a specific model using SOM in the study of language acquisition, the DevLex-II model. We presented an overview of how this model can be successfully used to address a number of important issues in monolingual and bilingual language acquisition, and illustrated its properties and applications in several psycholinguistic domains, including the modeling of vocabulary spurt, aspect acquisition, AoA effects, and cross-language semantic priming. We demonstrated that DevLex-II is a scalable model that can account for a variety of linguistic patterns in child and adult language learning.

We can highlight here a few key features of DevLex-II for the study of language acquisition. First, in contrast to previous computational models, DevLex-II is based on unsupervised learning (specifically SOM) and Hebbian learning, two powerful and biologically plausible principles of computation. These principles have allowed us to simulate the dynamics underlying both monolingual and bilingual lexical representations and interactions. Second, DevLex-II relies on the use of large-scale realistic linguistic data as the input. By simulating actual lexical forms and meanings, we are able to achieve developmental and lexical realism in our models. Third, DevLex-II incorporates computational learning properties (e.g., self-adjustable neighborhood functions, spreading activation, lateral connection) against the context of realistic language learning so that it can be used to simulate both language acquisition and language processing, in both L1 and L2 contexts.

To scholars of monolingual language acquisition, connectionist learning models are no new beasts. The original Rumelhart and McClelland (1986) past tense model and the subsequent debates, the Elman et al. (1996) book on rethinking innateness, and the more recent special issue edited by MacWhinney (2010) have all popularized the utility of connectionist models. Most researchers in L1 studies can appreciate the distinct advantages of connectionist learning models in allowing us to manipulate variables of interest flexibly and to study their interactions in a more systematic way (e.g., input quantity and quality, word frequency, word length in affecting error patterns). However, to scholars of bilingual language acquisition, the utility of connectionist models has yet to be fully appreciated.

The most popular computational model in bilingualism, the Bilingual Interactive Activation (BIA) model (Dijkstra and van Heuven, 1998), was based on the interactive activation (IA) model of McClelland and Rumelhart (1981). IA-based models typically lack a learning mechanism, and as such, they tend to focus on capturing representation and processing states of mature bilingual speakers and listeners (which is important in its own right). Computationally implemented learning models of bilingualism, however, remain scarce. It is important for researchers to develop connectionist learning models to capture the acquisition and interaction of multiple languages. This is because through modeling we can systematically identify the interactive effects of the two languages in terms of L2 onset time, L2 input frequency, amount of L1 vs. L2 input, order of L1 vs. L2 learning, and how these variables may separately or jointly impact both the learning trajectory and the learning outcome.

In a recent special issue edited by Li (2013) on computational modeling of bilingualism, a number of studies have attempted to fill the gap by taking advantage of the features of connectionist models to study bilingual acquisition and processing (e.g., Cuppini et al., 2013; Monner et al., 2013; Zhao and Li, 2013). These studies not only attempted to address specific problems and disentangle the effects of entrenchment, proficiency, memory resources, and lexical semantic distances, but also provided mechanistic accounts of important theoretical issues. For example, Monner et al. (2013) tested specifically the "less is more" hypothesis (Johnson and Newport, 1989) in a connectionist model, in which the increase of working memory was simulated by the use of new cell assemblies in the model, whereas L1 entrenchment was simulated by the training of the network with variable-length exposure of L1 before the onset of L2. In this way, the modeling results allowed the researchers to dissociate effects due to the increase of memory

and the increase of age, which are confounded in natural learning settings.

Monner et al.'s (2013) model illustrates the important role that connectionist modeling can play in second language learning, and at the same time speaks to the possibility that age-related learning differences as prescribed by the critical period hypothesis may be accounted for by the interactive effects of entrenchment of L1 and computational resources, which is highly consistent with the simulation results from DevLex-II as discussed above (see also Hernandez et al., 2005). The degree of entrenchment is a result of how well established the network has the L1 representation structure: the more consolidated the representation (as in late L2 learning), the more resistant to change the topographic structure becomes in the model. New items from the L2 have to use existing structures built from the L1, and any further learning is only able to result in what we call "parasitic" representations. By contrast, when learning occurs early, fewer L1 words may have become fully consolidated in the representation and the network may be less committed to L1 representations, and therefore the system is still open to adaptation in the face of new input from L2 so as to be able to continually reorganize and restructure the L2 representations. It is important to note that timing itself is not the cause, but time of learning is accompanied by different dynamics of interactions between the two languages for learning. Simulations from Monner et al.'s (2013) model and from DevLex-II suggest that the nature of bilingual representation is the result of a highly dynamic and competitive process in which early learning constrains later development, therefore shaping the time course and structure of later language systems. To what extent early learning impacts later learning, and to what extent extensive later learning can soften or even reverse early-learned structure, will remain the key research questions in the years to come.

What would be the future for SOM-based connectionist language models, in particular the DevLex-II model? One issue to bear in mind as we move forward is that we must bridge computational modeling results with a variety of other behavioral, neuropsychological, and neuroimaging findings, especially given the neurally plausible architectures of multiple SOM models (e.g., DevLex-II or DISLEX). As discussed earlier, Kiran et al. (2013) provided an excellent example in this regard, in which the investigators constructed a model to simulate neuropsychological patterns of each of 17 bilingual patients following traumatic brain injury and subsequent treatment. A second dimension to explore SOM-based models for language acquisition is to further identify the relationship between map organizations developed at different stages of learning and the impact that these different organizations may have on the behavior of learning (e.g., in terms of speed and outcome of learning success). DevLex-II has made some efforts in this regard, for example, in simulating vocabulary spurt and cross-language semantic priming, by linking the representational structure and semantic richness of the representation to the performance (e.g., word learned or priming effects) in the model, but more needs to be done. A third dimension to extend SOM-based models of language may be to study how syntactic structures can be acquired in both L1 and L2. Given the status of syntax in linguistic theories, connectionist models have yet to demonstrate their utility in learning syntactic structures. Elman (1990) showed that the simple recurrent network (SRN) can learn the hierarchical recursive structure of sentences. One could consider to introduce mechanisms into SOMs to capture temporal order information in language representation by using recursive SOMs (see Tiňo et al., 2006 for an example).

As we think ahead we also must develop SOM models of language that can make distinct predictions in light of the simulations and empirical data. In some cases, the empirical data may have not yet been obtained, or cannot be obtained (e.g., as in the case of brain injury, one cannot go back to pre-lesion conditions), and this is the occasion where modeling results will be extremely helpful. Not only should computational modeling verify existing patterns of behavior on another platform, it should also inform theories of L1 and L2 acquisition by making distinct predictions under different hypotheses or conditions. In so doing, computational modeling will provide a new forum for generating novel ideas, inspiring new experiments, and helping formulate new theories (see McClelland, 2009 for a discussion of the role of modeling in cognitive science). Finally, computationally minded researchers in language science should follow a recent call by Addyman and French (2012) to make an effort to provide user-friendly interfaces and tools to non-modelers, so that many more students of language acquisition can test computational models without fearing the technical hurdles posed by programming languages, source codes, and simulating environments.

## ACKNOWLEDGMENTS

This research was supported by a grant from the National Science Foundation (BCS-0642586, and in part BCS-1057855) to Ping Li and by a Faculty Development Grant of Emmanuel College to Xiaowei Zhao. We thank Julien Mayor, Colin Davis, Michael Thomas, and Brian MacWhinney for comments and suggestions on an earlier draft of this article.

## REFERENCES

Addyman, C., and French, R. M. (2012). Computational modeling in cognitive science: a manifesto for change. *Top. Cogn. Sci.* 4, 332–341. doi: 10.1111/j.1756-8765.2012.01206.x

Basnight-Brown, D., and Altarriba, J. (2007). Differences in semantic and translation priming across languages: the role of language direction and language dominance. *Mem. Cogn.* 35, 953–965. doi: 10.3758/BF03193468

Bates, E. (1999). "Plasticity, localization and language development," in *The Changing Nervous System: Neurobehavioral Consequences of Early Brain Disorders*, eds S. Broman and J. M. Fletcher (New York: Oxford University Press), 214–253.

Bates, E., and Carnevale, G. (1993). New directions in research on language development. *Dev. Rev.* 13, 436–470. doi: 10.1006/drev.1993.1020

Benedict, H. (1979). Early lexical development: comprehension and production. *J. Child Lang.* 6, 183–200. doi: 10.1017/S0305000900002245

Bickerton, D. (1984). The language bioprogram hypothesis. *Behav. Brain Sci.* 7, 173–188. doi: 10.1017/S0140525X00044149

Bloom, K., Lifter, K., and Hafitz, J. (1980). Semantics of verbs and the development of verb inflection in child language. *Language* 56, 386–412.

Bowers, J. S. (2002). Challenging the widespread assumption that connectionism and distributed representations go hand-in-hand. *Cogn. Psychol.* 45, 413–445. doi: 10.1016/S0010-0285(02)00506-6

Clark, E. V. (2009). *First Language Acquisition*, 2nd Edn. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511806698

Clark, E. V., and Hecht, B. F. (1983). Comprehension, production, and language acquisition. *Annu. Rev. Psychol.* 34, 325–349. doi: 10.1146/annurev.ps.34.020183.001545

Cuppini, C., Magosso, E., and Ursino, M. (2013). Learning the lexical aspects of a second language at different proficiencies: a neural computational study. *Biling. Lang. Cogn.* 16, 266. doi: 10.1017/S1366728911000617

Dale, P. S., and Fenson, L. (1996). Lexical development norms for young children. *Behav. Res. Methods Instrum. Comput.* 28, 125–127. doi: 10.3758/BF03203646

Davis, C. (1999). *The Self-Organising Lexical Acquisition and Recognition (SOLAR) Model of Visual Word Recognition*. Unpublished doctoral dissertation, University of New South Wales, Kensington.

Dijkstra, T., and van Heuven, W. J. B. (1998). "The BIA model and bilingual word recognition," in *Localist Connectionist Approaches to Human Cognition*, eds J. Grainger and A. M. Jacobs (Mahwah, NJ: Erlbaum), 189–226.

Dimitropoulou, M., Duñabeitia, J. A., and Carreiras, M. (2011). Two words, one meaning: evidence of automatic co-activation of translation equivalents. *Front. Psychol.* 2:188. doi: 10.3389/fpsyg.2011.00188

Elman, J. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Elman, J., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Foster-Cohen, S. H. (1999). *An Introduction to Child Language Development*. London: Longman.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* 3, 128–135. doi: 10.1016/S1364-6613(99)01294-2

Goldfield, B. A., and Reznick, J. S. (1990). Early lexical acquisition: rate, content, and the vocabulary spurt. *J. Child Lang.* 17, 171–183. doi: 10.1017/S0305000900013167

Grossberg, S. (1976a). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biol. Cybern.* 23, 121–134. doi: 10.1007/BF00344744

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions. *Biol. Cybern.* 23, 187–202. doi: 10.1007/BF00340335

Guenther, F. H., and Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *J. Acoust. Soc. Am.* 100, 1111–1121. doi: 10.1121/1.416296

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*, 2nd Edn. Upper Saddle River, NJ: Prentice Hall.

Hebb, D. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York: Wiley.

Hernandez, A., and Li, P. (2007). Age of acquisition: its neural and computational mechanisms. *Psychol. Bull.* 133, 638. doi: 10.1037/0033-2909.133.4.638

Hernandez, A., Li, P., and MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends Cogn. Sci.* 9, 220–225. doi: 10.1016/j.tics.2005.03.003

Hinton, G. E., and Sejnowski, T. J. (1999). *Unsupervised Learning: Foundations of Neural Computation*. Cambridge, MA: The MIT press.

James, D., and Miikkulainen, R. (1995). "SARDNET: a self-organizing feature map for sequences," in *Advances in Neural Information Processing Systems*, Vol. 7, eds G. Tesauro, D. S. Touretzky, and T. K. Leen (Cambridge, MA: MIT Press), 577–584.

Johnson, J. S., and Newport, E. L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cogn. Psychol.* 21, 60–99. doi: 10.1016/0010-0285(89)90003-0

Kiran, S., Graesman, U., Sandberg, C., and Miikkulainen, R. (2013). A computational account of bilingual aphasia rehabilitation. *Biling. Lang. Cogn.* 16, 325. doi: 10.1017/S1366728912000533

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd Edn. Berlin: Springer. doi: 10.1007/978-3-642-56927-2

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Percept. Psychophys.* 50, 93–107. doi: 10.3758/BF03212211

Lenneberg, E. H. (1967). *Biological Foundations of Language*. New York, NY: Wiley.

Li, P. (2003). "Language acquisition in a self-organising neural network model," in *Connectionist Models of Development: Developmental Processes in Real and Artificial Neural Networks*, ed. P. Quinlan (Hove: Psychology Press), 115–149.

Li, P. (in press). "Bilingualism as a dynamic process," in *Handbook of Language Emergence*, eds B. MacWhinney and W. O'Grady (Boston: John Wiley & Sons, Inc.).

Li, P. (2013). Computational modeling of bilingualism. *Biling. Lang. Cogn.* 16, 241–366. doi: 10.1017/S1366728913000059

Li, P., Burgess, C., and Lund, K. (2000). "The acquisition of word meaning through global lexical co-occurrences," in *Proceedings of the Thirtieth Annual Child Language Research Forum*, ed. E. V. Clark (Stanford, CA: Center for the Study of Language and Information), 167–178.

Li, P., and Farkaš, I. (2002). A self-organizing connectionist model of bilingual processing. *Adv. Psychol.* 134, 59–85. doi: 10.1016/S0166-4115(02)80006-1

Li, P., Farkaš, I., and MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Netw.* 17, 1345–1362. doi: 10.1016/j.neunet.2004.07.004

Li, P., and MacWhinney, B. (2002). PatPho: a phonological pattern generator for neural networks. *Behav. Res. Methods Instrum. Comput.* 34, 408–415. doi: 10.3758/BF03195469

Li, P., and Shirai, Y. (2000). *The Acquisition of Lexical and Grammatical Aspect*. Berlin: Mouton de Gruyter.

Li, P., and Zhao, X. (2009). "Computational modeling of the expression of time," in *The Expression of Time*, eds W. Klein and P. Li (Berlin: Mouton de Gruyter), 241–271.

Li, P., and Zhao, X. (2012). "Connectionism," in *Oxford Bibliographies Online: Linguistics*, ed. M. Aronoff (New York, NY: Oxford University Press). Available at: http://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0010.xml

Li, P, Zhao, X., and MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cogn. Sci.* 31, 581–612. doi: 10.1080/15326900701399905

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (2010). Computational models of child language learning: an introduction. *J. Child Lang.* 37, 477–485. doi: 10.1017/S0305000910000139

MacWhinney, B. (2012). "The logic of the unified model," in *Routledge Handbook of Second Language Acquisition*, eds S. Gass and A. Mackey (New York, NY: Routledge), 211–227.

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua* 92, 199–227. doi: 10.1016/0024-3841(94)90342-5

Mayor, J., and Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychol. Rev.* 117, 1–31. doi: 10.1037/a0018130

McClelland, J. (2009). The place of modeling in cognitive science. *Top. Cogn. Sci.* 1, 11–28. doi: 10.1111/j.1756-8765.2008.01003.x

McClelland, J., and Rumelhart, D. (1981). An interactive activation model of context effects in letter perception: part 1. An account of basic findings. *Psychol. Rev.* 88, 375–407. doi: 10.1037/0033-295X.88.5.375

McClelland, J., Rumelhart, D., and the PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2. Cambridge, MA: MIT Press.

Menn, L., and Stoel-Gammon, C. (1993). "Phonological development: learning sounds and sound patterns," in *The Development of Language*, 3rd Edn, ed. J. B. Gleason (New York, NY: Macmillan), 65–113.

Mervis, C. B., and Bertrand, J. (1994). Acquisition of the novel name-nameless category (N3C) principle. *Child Dev.* 65, 1646–1663. doi: 10.2307/1131285

Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. Cambridge, MA: MIT Press.

Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self organizing feature map model of the lexicon. *Brain Lang.* 59, 334–366. doi: 10.1006/brln.1997.1820

Miikkulainen, R., Bednar, J. A., Choe, Y., and Sirosh, J. (2005). *Computational Maps in the Visual Cortex*. New York: Springer.

Miikkulainen, R., and Kiran, S. (2009). "Modeling the bilingual lexicon of an individual subject," in *Lecture Notes in Computer Science 5629: Proceedings of the Workshop on Self-Organizing Maps* (WSOM'09, St. Augustine, FL) (Berlin: Springer), PMC2767190.

Miller, G. A. (1990). WordNet: an on-line lexical database. *Int. J. Lexicogr.* 3, 235–312. doi: 10.1093/ijl/3.4.235

Monner, D., Vatz, K., Morini, G., Hwang, S., and DeKeyser, R. (2013). A neural network model of the effects of entrenchment and memory development on grammatical gender learning. *Biling. Lang. Cogn.* 16, 246. doi: 10.1017/S1366728912000454

Munakata, Y., and Pfaffly, J. (2004). Hebbian learning and development. *Dev. Sci.* 7. 141–148. doi: 10.1111/j.1467-7687.2004.00331.x

Neely, J. H., and Durgunoglu, A. (1985). Dissociative episodic and semantic priming effects in episodic recognition and lexical decision tasks. *J. Mem. Lang.* 24, 466–489. doi: 10.1016/0749-596X(85)90040-3

Pavlenko, A. (2009). "Conceptual representation in the bilingual lexicon and second language vocabulary learning," in *The Bilingual Mental Lexicon: Interdisciplinary Approaches*, ed. A. Pavlenko (Tonawanda, NY: Multilingual Matters), 125–160.

Reznick, J. S., and Goldfield, B. A. (1992). Rapid change in lexical development in comprehension and production. *Dev. Psychol.* 28, 406–413. doi: 10.1037/0012-1649.28.3.406

Richardson, F. M., and Thomas, M. S. (2008). Critical periods and catastrophic interference effects in the development of self-organizing feature maps. *Dev. Sci.* 11, 371–389. doi: 10.1111/j.1467-7687.2008.00682.x

Ritter, H., and Kohonen, T. (1989). Self-organizing semantic maps. *Biol. Cybern.* 61, 241–254. doi: 10.1007/BF00203171

Rumelhart, D., and McClelland, J. (1986). "On learning the past tenses of english verbs," in *Parallel Distributed Processing: Explorations, in the Microstructure of Cognition,* Vol. 2, *in Psychological and Biological Models*, eds L. J. McClelland, D. E. Rumelhart, and PDP Research Group (Cambridge: MIT Press), 216–271.

Saxton, M. (2010). *Child Language: Acquisition and Development.* London: SAGE Publications.

Shirai, Y. (1991). *Primacy of Aspect in Language Acquisition: Simplified Input and Prototype.* Ph.D. dissertation, Applied Linguistics, University of California at Los Angeles.

Silberman, Y., Bentin, S., and Miikkulainen, R. (2007). Semantic boost on episodic associations: an empirically-based computational model. *Cogn. Sci.* 31, 645–671. doi: 10.1080/15326900701399921

Spitzer, M. (1999). *The Mind within the Net: Models of Learning, Thinking, and Acting.* Cambridge, MA: MIT Press.

Sporns, O. (2010). *Networks of the Brain.* Cambridge: The MIT Press.

Thomas, M. S. C., and Johnson, M. H. (2008). New advances in understanding sensitive periods in brain development. *Curr. Dir. Psychol. Sci.* 17, 1–5. doi: 10.1111/j.1467-8721.2008.00537.x

Tiňo, P., Farkaš, I., and van Mourik, J. (2006). Dynamics and topographic organization of recursive self-organizing maps. *Neural Comput.* 18, 2529–2567. doi: 10.1162/neco.2006.18.10.2529

Wang, X., and Forster, K. I. (2010). Masked translation priming with semantic categorization: testing the sense model. *Biling. Lang. Cogn.* 13, 327–340. doi: 10.1017/S1366728909990502

Westermann, G., Ruh, N., and Plunkett, K. (2009). Connectionist approaches to language learning. *Linguistics* 47, 413–452. doi: 10.1515/LING.2009.015

Xing, H., Shu, H., and Li, P. (2004). The acquisition of Chinese characters: corpus analyses and connectionist simulations. *J. Cogn. Sci.* 5, 1–49.

Zhao, X., and Li, P. (2009). Acquisition of aspect in self-organizing connectionist models. *Linguist. Interdiscip. J. Lang. Sci.* 47. 1075–1112. doi: 10.1515/LING.2009.038

Zhao, X., and Li, P. (2010). Bilingual lexical interactions in an unsupervised neural network model. *Int. J. Biling. Educ. Biling.* 13, 505–524. doi: 10.1080/13670050.2010.488284

Zhao, X., and Li, P. (2013). Simulating cross-language priming with a dynamic computational model of the lexicon. *Biling. Lang. Cogn.* 16. 288–303. doi: 10.1017/S1366728912000624

Zhao, X., Li, P., and Kohonen, T. (2011). Contextual self-organizing map: software for constructing semantic representation. *Behav. Res. Methods* 43, 77–88. doi: 10.3758/s13428-010-0042-z

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 25 June 2013; paper pending published: 18 August 2013; accepted: 18 October 2013; published online: 19 November 2013.*

*Citation: Li P and Zhao X (2013) Self-organizing map models of language acquisition. Front. Psychol. **4**:828. doi: 10.3389/fpsyg.2013.00828*

*This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 20 November 2013
doi: 10.3389/fpsyg.2013.00857

# Context, cortex, and associations: a connectionist developmental approach to verbal analogies

***Pavlos Kollias*[1]* and *James L. McClelland*[2]**

[1] *Department of Psychology, Princeton University, Princeton, NJ, USA*
[2] *Department of Psychology, Center for Mind, Brain and Computation, Stanford University, Stanford, CA, USA*

***Edited by:***
*Gary Lupyan, University of Wisconsin, USA*

***Reviewed by:***
*Brian MacWhinney, Carnegie Mellon University, USA*
*Gary Lupyan, University of Wisconsin, USA*
*Arturo Hernandez, Maine Medical Center Research Institute, USA*

****Correspondence:***
*Pavlos Kollias, Department of Psychology, Green Hall, Princeton University, Princeton, NJ 08540-1010, USA*
*e-mail: kollias@princeton.edu*

We present a PDP model of binary choice verbal analogy problems (A:B as C:[D1|D2], where D1 and D2 represent choice alternatives). We train a recurrent neural network in item-relation-item triples and use this network to test performance on analogy questions. Without training on analogy problems *per se*, the model explains the developmental shift from associative to relational responding as an emergent consequence of learning upon the environment's statistics. Such learning allows gradual, item-specific acquisition of relational knowledge to overcome the influence of unbalanced association frequency, accounting for association effects of analogical reasoning seen in cognitive development. The network also captures the overall degradation in performance after anterior temporal damage by deleting a fraction of learned connections, while capturing the return of associative dominance after frontal damage by treating frontal structures as necessary for maintaining activation of A and B while seeking a relation between C and D. While our theory is still far from being complete it provides a unified explanation of findings that need to be considered together in any integrated account of analogical reasoning.

**Keywords: analogical reasoning, connectionist models, cognitive development, FTLD, cognitive control, word association**

## 1. INTRODUCTION

Analogical reasoning, the ability to detect and exploit patterns of relational similarity between domains of knowledge, has been argued to be at the core of human cognition (Hofstadter, 2001). Studies and models have focused on different aspects of analogical reasoning. According to the number of constituents that the two knowledge domains will have, the form of the questions that the task will assume, and other variables, different paradigms have been developed. Some studies have focused on the processing of analogous domains of knowledge and situations where many objects are related with each other (Duncker, 1945; Gick and Holyoak, 1983). In this case, the entities in the knowledge domains are assumed to have a form of structure that can be mapped with entities in an analogous domain as a result of analogical reasoning (Gentner, 1983). In some studies, the subjects are explicitly asked to solve an analogy problem, while in others their capability to spontaneously infer an analogy is tested, mainly for goal-directed problem solving tasks (Duncker, 1945; Gick and Holyoak, 1983; Holyoak et al., 1984).

In one important type of explicit analogy problems, participants see three items (A:B::C) and must select a fourth item to complete an analogy of the form "A is to B as C is to D" (Spearman, 1923; Sternberg and Nigro, 1980; Sternberg et al., 1982). Participants, commonly, are given a set of candidate D items and must choose the option that maximizes the similarity of the relation between A and B with the relation between C and the picked D. Such forced-choice verbal analogy problems are often used in standardized tests of mental ability, and researchers have examined performance of adults and children either using pictorial presentation of objects and scenes (Goswami and Brown, 1990; Kotovsky and Gentner, 1996) or verbally (Gentile et al., 1969, 1977; Sternberg and Nigro, 1980).

In this paper we seek to provide an integrated account of both developmental and neuropsychological findings from studies employing forced choice verbal analogy problems. The number of candidate options for the D item is not constant across studies. Depending on the variables of interest, studies have used, two (Morrison et al., 2004), four (Goswami and Brown, 1990) or more candidate responses for the D item.

For simplicity, we simulate performance in binary choice verbal analogy problems, where only two candidate D responses are provided (denoted henceforth as A:B::C:[D1|D2], where D1 and D2 are the two alternatives). This is sufficient for the scope of behavioral phenomena we consider. We believe this focus on a single type of problem, together with the integration of both developmental and neuropsychological constraints, is a good first step for the development of an account in which verbal (and perhaps other forms of) analogical reasoning is viewed as an emergent consequence of reliance on learning and distributed representations. As such, our model complements other approaches which aim to address a broader range on analogical reasoning processes within the framework of mechanisms specifically constructed to support analogical reasoning (Hummel and Holyoak, 1997; Morrison et al., 2004; Doumas et al., 2008). We study the development of analogical reasoning as a consequence of knowledge acquisition and examine the special role of word associations. We suggest that word-association statistics complement the role of learning in explaining developmental patterns such as the relational shift seen in cognitive development from associative responses to appropriate relational responses.

Also, we investigate the role of word-associations in performance following frontal or temporal damage, and explain how associative responding returns after frontal damage and the deterioration of cognitive control. Our theory is far from addressing all of the findings in the very broad analogical reasoning literature. However, we argue that we bring together findings from the more limited domain of forced-choice verbal analogy problems that have not been jointly considered before and provide an emergentist alternative to classical approaches to solving such analogy problems. Extensions to our framework will be required to address the full range of analogical reasoning paradigms.

In the rest of this introductory section we review the key findings that we consider to be important for the development and deterioration of performance in verbal analogies. Our model provides an integrated qualitative account of these findings. In section 2 we describe the architecture and representational assumptions of our model in detail. In addition, we explain the training process and testing of the model in analogy questions. In section 3 we demonstrate the results of our simulations. Finally, in section 4, we discuss the achievements and shortcomings of our model, compare it with other models in the literature and consider extensions to address a broader range of analogical reasoning situations.

## 1.1. KEY FINDINGS

### *1.1.1. The role of knowledge acquisition*

Early developmental theories of analogy-making attributed developmental changes in performance to a domain-general progression through a series of stages. Piaget et al. (1977) found uncertain evidence of analogical reasoning in children from 5- to 12- years old. These findings for incompetence of analogical reasoning at these ages were aligned to Piaget's more general account of the development of reasoning. Similarly Sternberg and Nigro (1980) suggested that children's strategies shift from associative responding in early ages to relational reasoning through domain-general changes. However, Goswami (1991) has argued that these theories underestimate children's analogical reasoning abilities and the influence of the environment.

Precursors of analogical reasoning have been noticed in children in early ages from infancy in simple problem solving studies (Crisafi and Brown, 1986; Brown, 1989). Additionally, children in the ages 3–6 show competence in analogical completion in traditional forced choice analogy studies (Goswami and Brown, 1989, 1990; Rattermann and Gentner, 1998), contradicting Piaget's earlier findings. In the Goswami studies, the materials were chosen to be familiar to children. Thus, the conclusion was that what guides analogical development is experience with the items and relations involved, instead of a change in a domain-general mechanism. The Goswami and Brown (1989, 1990) finding that the ability of children to complete analogies within familiar domains, compared to the incapacity in the Piagetian studies (Piaget et al., 1977), suggests that the capacity for analogical reasoning is not based on a domain-general capacity for formal operations, but depends on the amount of experience that children have within specific domains of knowledge.

### *1.1.2. Word associations and the relational shift*

A number of factors may affect children's responses in forced-choice analogy problems. Sternberg and Nigro (1980) suggested that children's preferences in problems of this type are initially associative. Achenbach (1970, 1971) designed a task to test individual preferences on relational versus associative strategies. In an A:B::C:[D1|D2] task used to distinguish analogical from associative responding, the candidate D choices contain, among others, the correct analogical choice and at least one choice that is more or less associated to C than the correct response. For example in the PIG:BOAR::DOG:[WOLF|CAT] analogy problem, the correct response would be the WOLF. But the foil CAT, which has higher semantic association with the word DOG than WOLF has, can be used to test the ability to respond analogically despite the presence of semantic distractors. Sternberg and Nigro (1980) showed that the response speed and errors of 9- and 12-year-olds depended on the degree of association between candidate D terms and C terms in the analogies, thus they concluded that younger children rely on associations while older children rely on relational matching.

Goswami and Brown (1989) has argued, though, that these relations were hard for the children to handle and proposed that children rely on associations when there is not enough knowledge of the domain. Taken together, the findings suggest that, despite the primary effect that domain-specific knowledge has, the role of word association should not be disregarded. Gentile et al. (1969) discovered that word pair association factors can explain a large portion of the variance in analogical responding of university students, who could also be primed to respond associatively. Recent studies have also highlighted the influence of word associations in analogy completion. Thibaut et al. (2011) have shown that analogies constructed with pairs of weakly semantically associated items were harder for children with inhibition problems. Also, in a neuroimaging study of verbal analogies, Bunge et al. (2005) showed that strong associations in the A:B part of the analogy significantly improved performance in the analogy completion task. This suggests that the more familiar the A:B relation the easier the comparison with the C:D term is.

In all cases semantic association seems to play an important role which either facilitates or inhibits correct analogical response, according to the relative strength of the association between the C term and correct vs. the incorrect alternative.

### *1.1.3. Neural basis of analogical reasoning*

Given the centrality and complexity of analogical reasoning it is unsurprising that several brain areas, associated with various cognitive processes, are involved in analogy-making. Specifically, cognitive control and semantic retrieval processes are involved. Bunge et al. (2005) showed activation of distinct cortical areas in association with component processes of analogical reasoning (semantic retrieval and relational integration).

A large number of neuropsychological (Stuss and Benson, 1984; Shallice and Burgess, 1991; Duncan et al., 1995; Waltz et al., 1999) and neuroimaging studies (Baker et al., 1996; Prabhakaran et al., 1997; Osherson et al., 1998) have implicated prefrontal cortex (PFC) in complex and high-level cognition such as reasoning.

Waltz et al. (1999) found that patients with frontal lobe damage had impaired performance in the more complex trials of the Ravens Progressive Matrices test (i.e., when more than one relation had to be integrated), a test that is cognitively similar to analogy tests. Mediation of the PFC has been found also in analogical reasoning tasks. Wharton et al. (2000) showed evidence for activation of the left dorsomedial prefrontal cortex (Brodmann's area 44 and 45) in geometric analogy problems.

Despite the evidence of activation of the PFC in analogical reasoning its exact role is still unknown. In non-analogy studies, Cohen and Servan-Schreiber (1992), by reviewing the deterioration of performance of schizophrenic patients in attentional (Abramczyk et al., 1987; Cornblatt et al., 1989) and linguistic (Chapman et al., 1964) tasks, suggested that the PFC plays an essential role in maintaining an internal context representation in a form that can constrain processing task-relevant input. Interestingly, Chapman et al. (1964) showed that schizophrenics could not interpret correctly a weak meaning of an ambiguous word even if the context of the sentence provided clear evidence for disambiguation. Instead, patients demonstrated meaning-frequency effects, preferring the more frequent meaning of a word over the contextually-appropriate meaning. Cohen and Servan-Schreiber (1992) provided simulations that captured this effect by lesioning a model component that corresponded to the prefrontal cortex. The PFC may play a similar role in allowing the correct alternative to be selected in A:B::C:[D1|D2] problems. Let us consider the concrete example we presented previously: In the PIG:BOAR::DOG:[WOLF|CAT] case the ability to pick the relationally appropriate response WOLF may depend on the PFC to maintain an internal representation of the A:B "context" to help override the strong association between DOG and CAT. The A:B part of the analogy is what provides the appropriate context for picking the analogically correct response.

In addition to the prefrontal cortex, temporal areas are argued to be important to verbal analogies, given their importance for semantic tasks (Hodges, 2000). The anterior temporal cortex (particularly in the left hemisphere) is argued to be important for verbally transmitted conceptual knowledge (Martin et al., 1996; Mummery et al., 1999).

The importance of these cognitive processes becomes apparent with the study of frontotemporal lobar degeneration (FTLD) patients. FTLD is a regional neurodegenerative etiology of dementia. A main classification of FTLD patients can be done, according to the primary locus of damage, which can be either in the frontal or in temporal areas and especially in the anterior temporal areas. Morrison et al. (2004) compared frontal and temporal FTLD patients' performance with that of control subjects. In a forced binary choice analogy task they found that both temporal and frontal damage patients made more errors than control participants. Since the choice was binary, one choice (called here D1) was correct and the other (called D2) was incorrect. The relative association of the C:D1 pair compared to that of the C:D2 was called the *Semantic Facilitation Index* (SFI). This Semantic Facilitation Index took positive, zero, and negative values. The sign of the index was based on an approximation of the difference between the C:D1 association and the C:D2. Frontal damage patients performed well with positive SFI problems (C:D1 association stronger than C:D2 association, where D1 is assumed to be the correct response), but their performance was impaired for equal SFI and especially negative SFI items, in which the incorrect choice had a higher association with the C term. Temporal patients, on the other hand showed overall depressed performance, and were less affected by SFI.

**Table 1** summarizes the key findings that we consider to be important for treatment within an integrated mechanistic account. We believe that any framework of analogical reasoning needs to follow a knowledge-acquisition approach in order to address the overall role of experience in solving verbal analogy problems. Specifically, we highlight here the role of association strength (the environment's statistics) on performance. We suggest that people's ability to use the context provided by the A:B pair depends in part on prefrontal integrity to maintain a representation of this context and in part on prior experience, and that a by-product of this experience dependence is that the retrieval process is either facilitated or inhibited by the relative association of the C item with the correct alternative as opposed to the incorrect response. Our model qualitatively integrates and simulates these findings.

### 1.2. MODELING FRAMEWORK AND DESIGN GOALS

Our model belongs to the Parallel Distributed Processing (PDP) tradition (McClelland et al., 1986; Rumelhart et al., 1986a). Connectionist networks embody characteristics that are appealing for relational representation such as gradience in representation, interactivity in a bidirectional manner between units allowing mutual satisfaction of constraints, nonlinearity, and adaptivity (McClelland, 1993). The principle of graded representations is essential for being able to represent a graded performance of analogical reasoning instead of an all-or-none approach where a comparison is or is not analogically appropriate. In addition, allows developmental explanations based on knowledge acquisition. Connectionist networks are accompanied by learning

**Table 1 | Key findings of verbal analogies.**

| DEVELOPMENT OF VERBAL ANALOGICAL REASONING | |
|---|---|
| 1. | Experience in relational knowledge is a key force behind the development of analogical reasoning (Goswami, 1991) |
| 2a. | There is a shift in analogical responding from associative strategies to relational reliance (Sternberg and Nigro, 1980) |
| 2b. | This relational shift must be supported by knowledge acquisition (Goswami and Brown, 1989) |
| **THE NEURAL BASIS OF VERBAL ANALOGIES** | |
| 3a. | The prefrontal cortex has been implicated in analogy and analogy-like imaging studies (Waltz et al., 1999; Wharton et al., 2000) |
| 3b. | A putative role of PFC is to maintain an internal representation of task context (Cohen and Servan-Schreiber, 1992) |
| 3c. | Frontal lobe lesions cause strong influence of associations in responding (Morrison et al., 2004) |
| 4. | Temporal lobe lesions cause general impairments in performance regardless of task relative associations (Morrison et al., 2004) |

algorithms that allow them to modify their weights, and hence their knowledge, with experience.

In line with the gradient character of our framework, one goal of our model was to address gradience in analogical reasoning. We argue that analogies can be drawn between pairs of items that have similar, instead of completely identical, relations. For example we suggest that the relation between DOG and PUPPY is more similar to the relation between CAT and KITTEN than it is to the relation between RIFLE and PISTOL, but also that both the DOG:PUPPY::CAT:KITTEN and DOG:PUPPY::RIFLE:PISTOL analogies can be valid, even though the relations vary in their degree of similarity. By using distributed relational representations we are able to solve relational problems even for cases where the relational representations are similar but not identical.

Second, our model is motivated by a desire to allow the relation retrieved between A and B to be affected by the rest of the analogy problem. As argued by French (2008) one has to consider both parts of the analogy to figure out which relation is the most appropriate. Considering our previous example, one cannot be a priori certain that the relevant relation between DOG and PUPPY is that of kinship, size-relation or anything else before the candidate relations are constrained by the C:D terms. Thus, we consider interactivity during relation-retrieval to be crucial for our theory.

Before turning to the details of our model, we note that our work builds on two previous modeling efforts. Leech et al. (2008) proposed a learning based model that served as one of the main inpirations for our approach, demonstrating how learning could explain aspects of development of analogical reasoning abilities. Morrison et al. (2004) offered a model of the pattern of neuropsychological deficits seen in FTLD within the LISA model of analogical reasoning (Hummel and Holyoak, 1997). Our model differs from both of these earlier models in several ways, and is the first to address both the developmental data and the neuropsychological findings within the same model. In the discussion we consider similarities and differences between the models in more detail.

## 2. MATERIALS AND METHODS

### 2.1. NEURAL NETWORK MODEL

#### *2.1.1. Architecture and representation*

Under our approach verbal analogies are a by-product of simple relational learning. A cognitive agent is exposed to item1-relation-item2 triples. It learns to associate the items with each other, such that the presentation of the two items tends to result in filling in the relation. This relation, in turn, may then work together with the presentation of one of the two items to constrain the retrieval of the other item. Our model draws on related early work by Hinton (1981). Hinton's effort embodied the same computational principles and the same psychological content. In an effort to implement semantic networks in parallel hardware, Hinton introduced a network very similar to the one proposed here, though at the time the learning machinery available for training such networks was more primitive.

Our network's training architecture is shown in **Figure 1**. Similarly to Hinton's network there are two visible pools for the role-fillers of a relational triple (A and B), a visible pool for the relation (R) and a hidden integrating pool. All three visible pools are connected with bidirectional projections with the hidden pool.

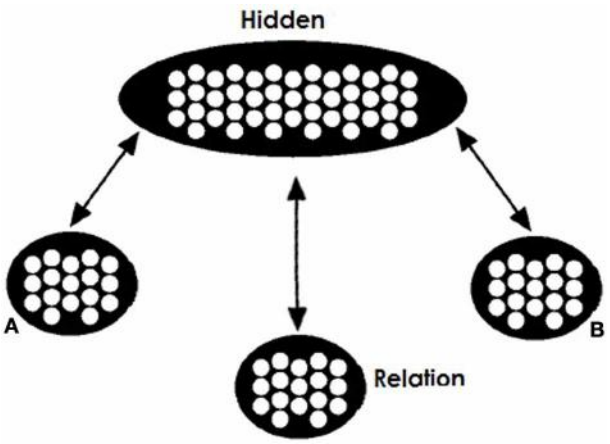



**FIGURE 1 | Model's training architecture.** The network consists of two visible pools (A and B) for the two concepts in a relation, a visible pool (Relation or R) for the relation between them, and a hidden pool. Connectivity between pools is bidirectional.

Objects in the A and B pools are represented in a localist manner. In contrast, representation in the Relation pool is distributed. We acknowledge that localist coding does not allow the network to capture the subtleties and effects of surface similarity between concepts, but reduces the complexity of learning for the network. For the Relation pool patterns of activation correspond to specific relations with similar patterns representing similar relations. Representations for a relation correspond to activations of 0 and 1. Each unit is assumed to correspond to semantic or visual features of the relation. In our simulations relations that are seemingly the same or can instantiate a valid analogy are assumed to come from a shared prototype pattern. For example the relation for the pair PIG:BOAR will be very similar to the relation for the pair DOG:WOLF since they both are distorted instances of a prototype pattern approximately corresponding to the relation "domesticated form of." Importantly, these two instances will be similar but not identical. Activation of appropriate representations in these three pools corresponds to a specific relational fact. We will call these facts *propositions* and henceforth denote them as A:R:B or A:B with the relations being implied. Of course we hold that both items and relations involve distributed representation—we use distributed relation representations to underscore that the relation (like items) are likely to vary across cases that might sometimes be labeled as the same, and to demonstrate that relations need not be identical for analogical reasoning to succeed.

#### *2.1.2. Training*

We train the network to complete relational propositions when given any 2 of a triple's elements as inputs. We use the backpropagation-through-time (Rumelhart et al., 1986b) learning algorithm as implemented in the pdptool simulation

environment [version 2.07, McClelland (2012)]. As it learns the associations between objects and relations, the model assigns to each input a stable pattern of activity across the hidden units. For each training epoch a set of propositions is presented to the network. This set of propositions corresponds to the network's environment. Each proposition (i.e., each A:R:B triple) appeared many times within each epoch. One third of the time, the A and B items were presented as input; in another third, the A and R items were presented as input; and in the final third, the B and R items were presented as input. In all three types of cases, the network had the task of filling in or completing the third member of the triple. Also each input combination (i.e., A:B, or A:R, or B:R) for a proposition can appear multiple times within an epoch. This number of times is called the proposition's *frequency*. We assume that the frequency of co-occurrence of items within propositions is an important contributor to the strength of their association, an idea well grounded in psycholinguistics (Spence and Owens, 1990). Of course we don't argue that associative value is captured only by co-occurrence frequency, but instead that it is being sufficiently approximated and on the same time allows us to address our questions on a very simple neural network. We leave the details of the model's environment for the Simulations section and a complete description is given in the Appendix.

#### 2.1.3. Testing

Our model is not trained on analogies *per se*, and we show that the ability to complete analogy problems can emerge from our training architecture. One way in which this might work would be to imagine that the network is first presented with A:_:B, and fills in the appropriate relation R; and that R or a trace of it persists after removing B and replacing A with C, allowing completion of the C:R:_ triple. Such an approach would be similar to the "relational priming" framework (Leech et al., 2008), which is grounded empirically on findings suggesting that analogies occur spontaneously (Goswami and Brown, 1989; Pauen and Wilkening, 1997; Tunteler and Resing, 2002). However, instead of using priming as a mechanism for retrieving the A:B relation first, and later use that to infer the D term, we suggest that the brain may possess the ability to simultaneously represent the A,B,C, and D terms of the analogy, and can use them all together to find a common relation that completes both the A:_:B triple and the C:_:D triple. Note that we do not claim that the brain's architecture has this capability solely to solve verbal analogies problems, but that, in general it possesses the capability of allowing mutual constraints to influence completion of neighboring propositions, just as mutual constraints can shape the perception of letters in visual letter perception (Rumelhart and McClelland, 1982). For present purposes we rely on this capability only for analogical reasoning, however.

The architecture we use is shown in **Figure 2**. The network consists of two copies of the trained network, sharing a common relation pool. The weights between A and Hidden1 (H1) are identical to the weights between C and H2; the weights between B and H1 are identical to the weights between D and H2; and the weights between H1 and R are the same as the weights between H2 and R. This way, the network takes

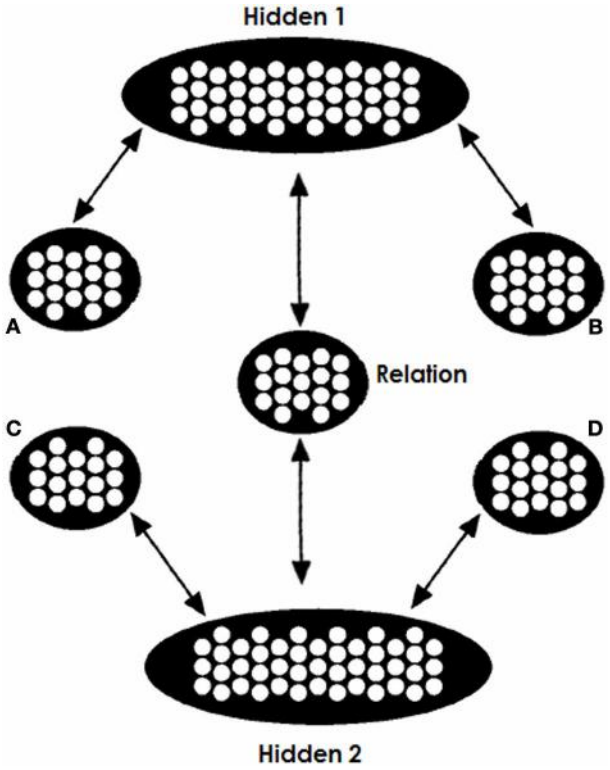



**FIGURE 2 | Model's testing architecture.** Two copies of the trained network share a common relation pool, so that both the A and B terms and the C and one candidate D term jointly constrain the search for a relation. While the A, B, and C items are clamped for the entire testing process the D items are clamped only at the beginning. We run two tests, one for each candidate D item; the D alternative with the strongest "echo" of activation at the end of testing is chosen.

advantage of its experience. Finally, one more refinement is required. In the thought experiment where one clamps to both the top and bottom part the same A:B inputs (denoted as A:B::A:B), the net input arriving at the Relation pool is double the input that it was trained to receive. What that means is that the net input that this pool receives departs from its experience and has essentially double the magnitude. For this reason we have halved the contribution of the H1-to-R and H2-to-R projections.

It is important to stress that we do not assume the brain literally contains two copies of the identical network, sharing the relation pool between them. We do, however, assume that both parts of an analogy problem can access connection-based knowledge at the same time and can mutually constrain each other, something that is made possible by this architecture. We assume that this ability is part of the general cognitive machinery that allows the interpretation of each of two items to be constrained by the other, even if one is presented first. A model with some relevant properties was previously proposed by McClelland (1986).

When the network is clamped with A,B,C, and D representations, the two parts of the network will try to fill in the R pool

the relations associated and learned for both of these two pairs of objects. Activation in the relation pool will depend upon the activation of the Hidden1 and Hidden2 pools. The Hidden1 pool will acquire a representation learned for the A:B input and Hidden2 will acquire a representation learned for the C:D input. Thus, Hidden1 will push the Relation pool toward representing the relation between items A and B and Hidden2 will push toward the relation between C and D. The more similar the two relations are, the greater the goodness of the network's state. Since all inputs are hard-clamped, the consistency of the Relation pool does not affect activation in the A,B,C nor D pools. However, when the D item is unclamped the consistency of the two relations and the goodness of the network's state will affect activation in pool D. If the two relation are similar, then the completed Relation with the C term will support activation of that D term. However, if the two relations were less similar, then the filled relation will not resemble the relation between C and D and thus the D item will not be supported by activation in the Hidden2 pool (we use below an example to make this idea clearer).

The test procedure we used is similar to one used by Dilkina et al. (2010) in a lexical decision task.[1]. For each analogy question we conduct two test trials. We clamp the A, B, and C items in their corresponding pools for the entire test. Each of the D alternatives is clamped on the D pool for some processing cycles, then removed for several more cycles, and the residual activation of the D unit that was initially activated is then recorded as a measure of the strength of the "echo" produced by that alternative. The alternative with the strongest echo is chosen as the network's response.

In summary, for a given analogy question A:B::C:[D1|D2], the process below is followed: The model calculates separately how good the A:B::C:D1 and A:B::C:D2 analogies are and compares their goodness to find the network's response. For each analogy we clamp the A, B, C, and D terms to their corresponding pools for a few processing cycles. During this phase, activation is spread over the network. Of interest is the fact that the A and B terms in the top part of the network push activation in the Hidden1 and the R pool as the network has learned to do in training. The same happens for the C and D terms. If the two pairs (A:B and C:D) share similar relations then the top and bottom part of the network will pattern-complete in the R pool similar representations. If they have different relations then the two parts of the network will push dissimilar representations in a resulting "meaningless" representation. After the first phase, we unclamp the D term (but keep all others clamped) and let the network process a few more cycles. By unclamping here we mean that we stop hard-coding input activation but keep the pool's state as it was without flushing it to zero. At this second phase, activation is still spread. Of interest is the activation in the D pool. The bottom part of the network has the C term clamped and now has a relation partially filled. Whether this filled relation was consistent (A:B and C:D similar) or inconsistent (A:B and C:D dissimilar) will determine how the bottom part of the network will allow the D activation to be maintained (echo measure). For consistent relations between the two parts the D term maintains higher activation. The model then chooses the D term with the higher maintained activation.

An example is given in **Figure 3**. As mentioned previously, for each analogy question we perform two separate tests on the network (one for D1 and one for D2). In our example, in one test we clamp PIG, BOAR, DOG, and WOLF in pools A,B,C, and D respectively; in the other test, CAT is clamped on the D pool instead of WOLF. The D item will be clamped for a few processing cycles. According to the degree of training that the network has received the activation in the top part will push the R pool's representations toward activating the relational pattern of the proposition PIG:BOAR. On the other hand activation on the bottom part will push the R pool toward the C:D relation. An approximation of the joint representations for the A:B and C:D pair is filled in the relation pool. In the case of DOG:WOLF the relation is very similar to the PIG:BOAR, opposed to the case of DOG:CAT. Thus, the R pattern completed in the DOG:WOLF case is more consistent with the DOG:WOLF proposition as it appears in the training set, opposed to the DOG:CAT proposition. This way, the WOLF unit is expected to maintain higher echo (activation at the end of processing).

However, it is important to note that such a behavior depends on the stored weights that the network has acquired and the extent to which acquired knowledge can support relational retrieval versus free association. D activation will also depend on the frequency and association of each of the D terms with the C term, given the presence of C. In our example, CAT is trained more frequently with the word DOG than is WOLF, thus the response CAT gains an advantage this way. Representational similarity facilitates analogical responding, but frequency facilitates associative responding, regardless of which alternative is the relationally correct response. Our expectation is that the interactions between these two forces will support the correct response in tasks where the correct is strongly associated with the C term and will prevent correct responding in tasks where the correct response is weakly associated, but that as training progresses, the network's encoding of both low and high frequency associations will become sufficiently robust that the relational similarity will allow correct responses, regardless of relative frequency.

#### *2.1.4. Effects of frontal and temporal damage in frontotemporal lobar degeneration (FTLD)*

***2.1.4.1. Frontal damage.*** Following the ideas of Cohen and Servan-Schreiber (1992) we assume that frontal damage diminishes an individual's ability to maintain context information—here, the representation of the A:B item—needed to constrain

[1]We explored a variety of different alternative decision criteria. The first was the echo of activation (residual activation after unclamping) as mentioned before. Alternatively, we left the D terms clamped for the entire processing and used the net-input at the end of processing as a decision criterion. Finally, we did not clamp any of the D items and selected the item with the highest activation at the end of processing. The network's responses were the same for all three different criteria, thus we do not report any simulations for the other two decision-criteria. We preferred the echo criterion because it expresses the close-ended nature of the task, without allowing the network to consider irrelevant alternative options. It also lies somewhere between the two alternative considered options allowing for both open-ended retrieval processes (echo after unclamping) and close-ended binary choice (necessary decision of higher echo among two alternatives).

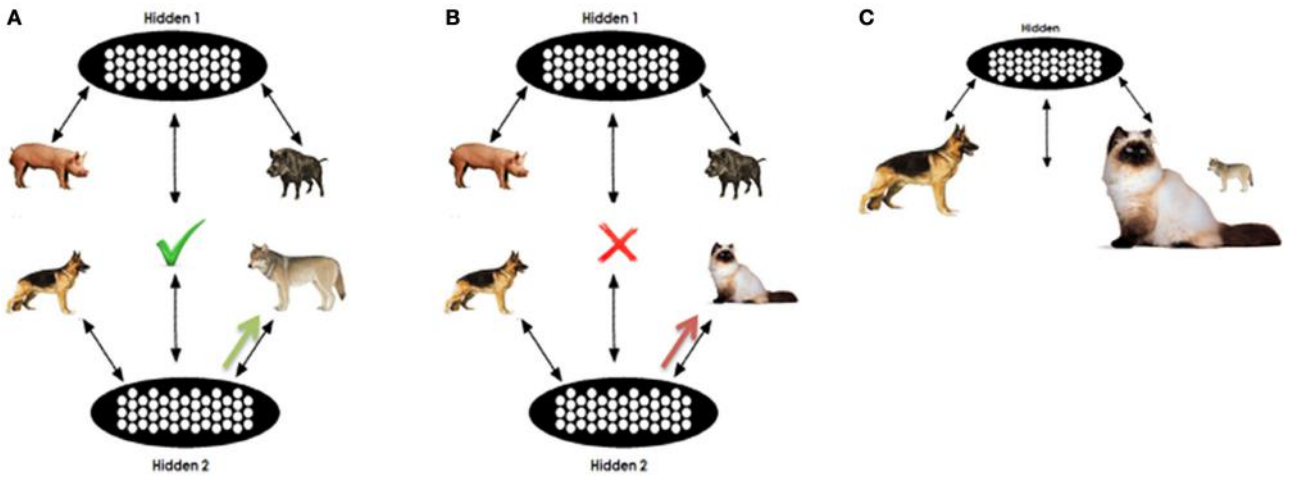



**FIGURE 3 | Illustrative analogy question. (A,B)** Sketch analogy testing in the question PIG:BOAR::DOG:[WOLF|CAT]. The concept WOLF will initially support consistent pattern completion in the R pool. When unclampled, the partially completed pattern will support residual activation in the D term opposed to the CAT. However, as shown in **(C)** the higher association of CAT with DOG provides an associative advantage when incomplete or insufficient information is clamped in the relation pool.

the C:[D1|D2] decision. One way PFC might do this is to regulate the overall activation of the hidden units mediating the A:B association. Accordingly, we treated frontal damage as reducing an overall biasing input to the hidden units in the A:B part of the network. This leads to a reduction of activation in H1 pool, impairing the ability of the A:B pair to influence the pattern of activation on the relation units, thereby causing the network to operate approximately as in **Figure 3C**. There are other possible ways in which PFC damage might reduce the contribution of the A:B association to constraining the specification of the relation between C and D which would likely have similar effects, and it is possible that different frontal syndromes (e.g., FTLD, schizophrenia) might produce such an effect in slightly different ways.

***2.1.4.2. Temporal damage.*** Anterior temporal damage in the network is much more straightforward. The role of the anterior temporal lobe is to allow the completion of propositions—in our case, the filling-in of the missing relations between the presented items. This mechanism is mapped to the pattern completion processes of the two parts of the network. Since all projections in the network are essential for pattern completion we will assume that random loss of connections corresponds to anterior temporal damage. The approach of randomly removing connections (set their weights to 0) has been followed by Rogers et al. (2004) for lesioning a model of semantic memory. Relying on our assumption that both parts of the analogy draw on the same underlying connection-based knowledge, we removed connections from one part of the network (A:B part) at random according to a specified probability, then copied the projections from the lesioned part of the network to the complementary (C:D) part, so that the lesion was identical for both parts of the network.

## 2.2. SIMULATIONS

We ran two simulations. The first was intended to demonstrate how the relational shift emerges within a single-purpose learning network. For our second simulation we used the trained networks of the first simulation and applied lesion to demonstrate how our model accounts for frontotemporal lobar degeneration. The two simulations use the same training set, which we describe in the following section. For each simulation we trained five networks with randomly initialized connection weights and their own randomly generated training environments.

### *2.2.1. Relational patterns*

The relational pattern representations were generated as follows. Relations in the training set come from 8 different relational prototypes, consisting of 16 active units out of the full set of 128 relation units. $\mathbf{R_i}$ refers to one of the relational prototypes. Two different prototypes (thought of as corresponding to very dissimilar relations) have no overlap at all on their set of active units. However, relations generated from the same prototype have 12 units in common. Specifically, an instance of a relation is obtained by turning off two of the units of the active units of the prototype. The turned off units are necessarily different across instances (See **Figure A1C** for two instances of the $\mathbf{R_2}$ prototype).

### *2.2.2. Training environment and parameters*

The training environment for each network consisted of blocks of propositions called *cells*. Each cell was designed to provide three analogy questions. One with positive SFI, one with neutral, and one with negative (recall that SFI is defined as the relative association of the correct response with the C term compared to the association of the incorrect response with the C term). For satisfying such a constraint, each cell should minimally have the format seen in **Figure 4**. In this format we have a basic *source proposition* A:R1:B, one *relational target* proposition C:R1′:D1, and three *foil target* propositions C:R2:D2, C:R3:D3, and C:R4:D4, one *weak*, one *moderate*, and one *strong* (Note that each relation in the above propositions is an exemplar from a different prototype, except that R1 and R1' are exemplars form the same prototype). This set gives three analogy tests (SFI>0: A:B::C:[D1|D2], SFI=0:

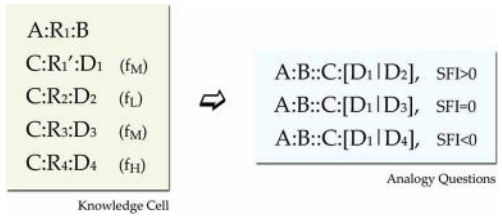



**FIGURE 4 | Training environment design.** Minimal design for a knowledge cell. The source proposition provides the A:B part of the analogy question, the C:D1 provides the correct response and there are three foils with various associations (shown in parentheses as Low, Medium, and High training frequency). Further details of the environment design are presented in the Appendix.

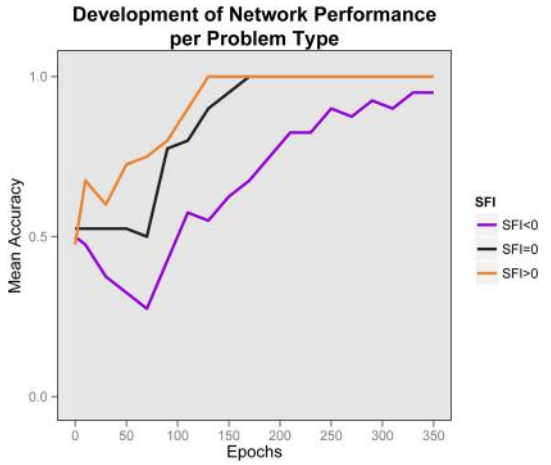



**FIGURE 5 | Development of performance by problem type.** Average performance of networks, by problem type in time. The relational shift is apparent by the early bifurcation of the plotted lines according to problem type combined with the convergence to correct responses later. Chance is at 0.5. Early in training, performance is below chance for the negative SFI problems and higher SFI problems are learned faster. Later, performance is improved for all problem types.

A:B::C:[D1|D3], and SFI<0: A:B::C:[D1|D4]). The three kinds of test within a cell are a result of the frequency variation that reflects association variation of a word (C) with several other words (D1, D2, D3, D4). Such a minimal design confounds frequency of association with frequency imbalances in the rates of occurrences of relations and items. Seemingly, a word D4 that has a high-frequency of co-occurrence with the word C for example, seems to have overall higher frequency of occurrence. To fix that, by keeping the structure behind this basic design that yields 3 analogy questions, we counterbalanced global frequency of training by reusing items and relations in propositions of various frequencies across cells (See Appendix for details on the imbalances and our scheme for counterbalancing).

## 3. RESULTS

### 3.1. SIMULATION 1: RELATIONAL SHIFT

We trained 5 randomly initialized networks with 5 randomly generated training sets (as described in the Appendix) for 350 epochs. The average error measure was almost zero at the end of training. However, what is important is not performance on the relational propositions, but on the analogy questions (higher echo of D1 vs. foils for each cell). At 350 epochs the networks had an average correct performance in the analogy tests of 0.97.

One important focus of interest is the development of this performance through time. We sampled performance every 10 epochs (initially all networks had performance at chance). In **Figure 5** we show the development of performance by problem-type. It is obvious that for all problem-types performance improves with knowledge-acquisition. Importantly, the networks learn to solve problems with higher SFI before other problem types, and is impaired early on for problems with lower SFI. The fact that performance is significantly below chance in the negative SFI condition indicates that responding is primarily determined by differences in association strength early on. Higher association of the correct response facilitates performance, while for the negative case, lower association inhibits performance and the networks fall below chance even after 10 epochs of training. Performance in the negative SFI condition starts to show improvement at approximately 80 epochs of training and the networks became relational even for the negative SFI case after 130 epochs of training, when the acquired knowledge allows the testing process to overcome the prepotency of the high-association foils. This pattern can be described as a relational shift, since the early bifurcation can be attributed to a reliance on word-association, while later performance relies on relational knowledge. Note that this occurred, even though the network was never trained to carry our analogical reasoning. Once both parts of the analogy are highly familiar, their mutual constraint outweighs associative responding.

### 3.2. SIMULATION 2: FTLD

Our simulations aim to capture the same key characteristics that Morrison et al. (2004) classified as important (**Figure 6**-Left):

1. control participants showed good performance at all levels of SFI,
2. frontal lobe patients showed depressed performance for lower SFI problems,
3. temporal patients showed depressed performance, and
4. both frontal and temporal patients exhibit a SFI effect, which is bigger for the frontal patients.

We lesioned the 5 networks of Simulation 1 at 350 epochs of training. We applied either a frontal lesion as a reduced bias in H1 pool or a temporal lesion as random loss of connections. The bias in the H1 units was reduced from −2 to −6.5 for the frontal lesion and connections were removed with .42 probability for the temporal lesion. Since the temporal lesion was randomly applied, we generated 5 lesioned versions of each network resulting in 25 networks with temporal lesions in total. Our network accounts for the interaction of lesion-type with problem-type. It is obvious that the negative SFI-problems are impaired compared to neutral and zero after frontal damage and

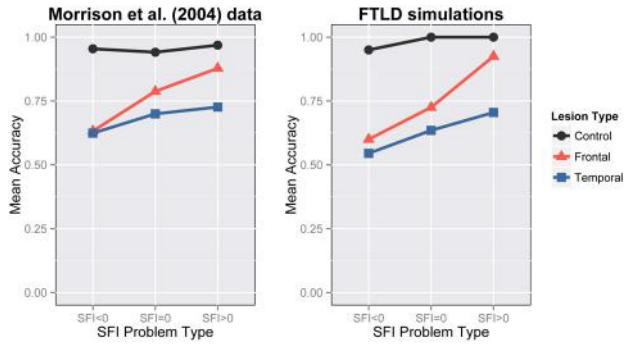



**FIGURE 6 | Frontotemporal lobar degeneration simulations.** Performance on analogy problems per problem-type for various lesions. **Left**: FTLD data from Morrison et al., 2004 (Reprinted with author's permission). **Right**: Simulation 2 results.

there is an overall degradation of performance after temporal damage (**Figure 6**). When frontal damage is applied to the model, evidently the model loses part of its ability to respond relationally and moves to associative strategies. This happened because the H1 pool loses its ability to maintain a representation of the A:B proposition. Hence the relation pool is influenced predominantly by the C:D part of the analogy. In the bottom part of the network there was only a constant input coming from the C pool. Thus, given the lack of other constraints, the most natural reaction is to complete the patterns that are more frequently trained with C. This is how associative responding emerges in our network. We consider this a very important implication of our model that is consistent with previous work on the prefrontal cortex (Cohen and Servan-Schreiber, 1992). In the semantic case, the network lost its ability to successfully complete the patterns. Hence, the whole process can be considered a noisy version of the control test.

As mentioned previously, there is a SFI effect in the experimental data for both frontal and temporal groups—the effect is larger for frontal than for temporal lesions, and this is captured in the simulation results as well. In both frontal and temporal cases the SFI effect is larger in the simulations than in the experimental data. We believe that the size of the SFI effect empirically will likely depend on a range of factors, including the degree of asymmetry of the word associations—our model appears to show a larger asymmetry effect overall than the experimental data. A possible reason for that could be the fact that the actual associations used in the experiment were less asymmetric than our model assumed.

## 4. DISCUSSION

We aimed to provide a model of key findings of verbal analogical reasoning. Despite our apparent focus on a specific class of analogy problems we unified disparate findings related to normal performance, development, and deterioration of verbal analogical problem solving within a learning system that learns relationally-mediated associations. Our simulations were qualitative and aimed to explain key phenomena at an abstract level, however, the basic pattern of the findings were robust and consistent with the basic patterns seen in developmental and neuropsychological data.

We showed how a neural network trained solely on relational propositions can solve analogy questions by allowing both halves of the analogy problem to mutually constrain the selection of a relation. In addition we showed in accordance with Goswami's theories how knowledge-acquisition can drive the improvement in performance during development—we do not require the invocation of a qualitative change in processing but only the gradual buildup of relation-mediated associations as the basis for the so-called relational shift. Importantly, we showed how knowledge acquisition can interact with the environment's statistics in a complementary manner to explain the behavioral patterns observed during development. Specifically, we explained the shift of children's reasoning from associative to relational as a by-product of learning and pattern completion on a given architecture. The architecture assumes cognitive control components that attempt to use acquired contextual information for overriding prepotency of incorrect responses. We showed that early in training top-down contextual information was not enough for overcoming the prepotency of strong foil responses. However, after training, without any changes on the system's parameters or architecture, this phenomenon is diminished and the network learns to yield relationally appropriate responses. The same network, trained on the same training environment, explained the overall degradation of performance after temporal damage as the loss of connections responsible for pattern completion processes and explained the return of associative responding after frontal damage as the loss of capacity to maintain context-related information that guides the retrieval of the appropriate target representations.

### 4.1. MODEL LIMITATIONS

Of course the interaction between frontal lobe development and knowledge acquisition is of great interest. It is important to note that our theory does not exclude frontal lobe development as a causal factor behind the behavioral pattern of analogical reasoning during development. On the contrary, executive-functions skills, attentional switching, and inhibitory control play very important and specialized roles in the development of analogy-making (Richland et al., 2006; Morrison et al., 2011; Richland and Burchinal, 2013). However, we argue that the developing frontal lobe synergistically with background knowledge cause the observed relational shift. Even if frontal-lobe development has its own trajectory we argue that it needs to exploit not only changes in frontal control functions but also acquired knowledge.

Our goal was not to provide a mapping from model components to brain areas. We do not believe that the two separate network parts reside in different brain areas, but instead that our architecture provides a neurocognitive explanation of the role of top-down contextual biasing. The frontal lobe, however, has a dual role in such a task. The one is to actively maintain goal-relevant contextual information (Cohen and Servan-Schreiber, 1992) for top-down biasing. The other is to guide attentional switching between what we model as two different networks (Hummel and Holyoak, 1997; Doumas et al., 2008; Morrison

et al., 2011). While these functions are potentially somewhat different in nature, the extent of their separability is unclear and they may potentially share the same underlying neural basis. Damage in the frontal lobe in FTLD patients probably causes severe impairments in both functions. Our theory, however, deals with impairments only in the former. We acknowledge that impairments in attentional switching functions as well (a component function not explicitly included in our model) could play a role in the associative effects found in frontal patients. Moving toward neurally grounded models will help us understand how the plausibility-driven constraint of interactivity is actually implemented in the brain and how it is deteriorated with frontal damage. As a first step, we argue that our model is not incompatible with the switching function. In terms of the architecture shown in **Figure 2** the top-down function would focus on actively maintaining the retrieved relation for the A:B pair of the analogy providing bias in the Hidden1 layer, as explained through the paper. We tried other forms of lesioning the top-down biasing function (e.g., impaired clamping in the A and B layers) and all had similar results, showcasing the importance of active maintenance of information in the A:B part of the network. Then switching control would be used to map this relation to the one from the C:D pair. This mapping could be done interactively by means of several continuous rapid switches of attention from one pair to another.

Our model did not aim to provide a fitted quantitative match for the data in the literature. While this is a goal for future work, we suggest that our approach is a useful, and perhaps necessary, first step. We were very concerned with potential confounds caused by stimulus-frequency constraints, so we prioritized counterbalancing. This allows us to be sure that our results depend on the co-occurrence frequency factors and not on the frequency of the items and relations that enter into these associations. Our design allowed us to counterbalance overall frequency of training of each word or relation. Of course, we don't argue that such counterbalancing is plausible and we believe that frequency of a specific item indeed plays a crucial role in analogy making (both in reality and in our theory). However, such questions were considered to be out of the scope of our current model. In future, it will be important to consider how such a framework can be extended to process more plausible data-sets.

A final limitation of our model is that it lacks an explanation of how the relational representations are learned and developed. We believe that relation representations change as a function of experience, but out current model lacks this property. Even if the environment provides invariants for many visual relationships (Doumas et al., 2008), we think it may be inappropriate to assume that a cognitive agent has available learned representations of complex relations like "is among the strongest" or "is the favorite student of". Instead, a complete model of analogical reasoning should consider how these representations are learned and shaped by their exemplars. Our model provides a suggestive initial framework for capturing the interacting complementary role between a learning agent and it's environment, and provides a base on which further work can proceed to address this issue.

### 4.2. COMPARISON WITH OTHER MODELS

As mentioned earlier, our model is related to and inspired by the neural network of Leech et al. (2008). We believe our approach advances these author's relational priming approach in several different ways, some of which were circumstantial to the relational priming model and some of which were intrinsic to it. The nature of the theory for "relations as transformations" addresses intuitively only a small class of relational propositions, namely propositions that express causal transformations. Such causal relationships were used by Goswami and Brown (1989) in pictorial analogy tasks. However, our use of distributed representations allows for a more flexible representation of relations, giving us the flexibility to address a wider range of relation types and corresponding findings. As discussed in the peer-commentary of the relational priming Leech et al. (2008) paper, transformation upon relations that operate in a linear manner on item representations (as implemented on the relational priming approach) suffers from non-transitivity. Our flexible representation of relations does not come cost-free, however, since our theory lacks a complete description of how such representations develop.

Also, importantly, as argued by French (2008) a priming-based approach does not cover the need to consider both parts of the analogy before settling to the correct response. Our network interactively considers both parts of the analogy for completing the shared relation. Although it is likely that this interactivity was not necessary to account for the data we simulated, we nevertheless agree with French (2008) that such interactivity has a role to play in analogical reasoning. Our simulations also used similarity, rather than strict relational identity, as a basis for analogical reasoning. While we did not explore effects of variation in relational similarity, pilot results from preliminary simulations revealed that, indeed, a higher number of shared relational "microfeatures" (Hinton, 1981) is associated with higher levels of activation. We set as a future goal the implementation of more complete simulations that will more fully exploit the interactive and similarity-based features of our architecture.

On the other hand our model differs from the LISA approach significantly. Morrison et al. (2011) have recently considered the relational shift within the LISA theory. Admittedly, the LISA theory provides a much more complete framework for a vast array of findings related directly or indirectly to analogical reasoning. We believe our approach has an important benefit. Our cognitive control explanation (both for the relational shift and for the frontal lesion) is fundamentally different than that proposed in the LISA model. In our case, cognitive control is expressed as a top-down influence in the network's operation that does not directly inhibit irrelevant information. Instead, inhibition of alternatives naturally arises as a consequence of competition among alternatives; PFC serves primarily to maintain a representation of context, so that the mutual constraint between the A:B and C:D pairs can proceed, and the relational shift emerges within the network without any hard-coded domain-general changes but as a simple consequence of learning. In contrast to this, Morrison et al. (2011) argue that the relational shift is a result of the domain-general maturation of the inhibitory system that parametrically is hard-coded to change values during development.

While maturation of the PFC is likely to play a role in maintaining context representations, we emphasize that experience is also likely to contribute to the relational shift during development. Our approach is emergentist and such a suggestion is very important in cognitive science (McClelland, 2010) as it eliminates the need for assuming specialized systems or hard-coded components. In addition, our framework provides unified simulations of both the developmental and neuropsychological findings, operating upon the same training set, something that is important for the plausibility of the theory. We believe and hope that the two classes of models will both contribute to the further development of mechanistic accounts of analogical reasoning.

### 4.3. FUTURE WORK

We consider the potential of addressing findings in the cognitively related field of metaphor comprehension. Metaphor comprehension can be seen as the process of filling out the A and D terms of an analogy in which the B and C are given Turney and Littman (2003). Consider the metaphor "demolish an argument." Comprehension of such a sentence can be seen as the process of inferring an analogy between two domains which the metaphor links. One can think of the analogy CRITICIZE:ARGUMENT::DEMOLISH:BUILDING. The statistical pattern completion properties of neural networks are appealing for such a task. Clamping the known B and C terms might lead to completing the unknown terms, given the constraints that the R pool will pose. Such a prospect further justifies our modeling choice for an interactive architecture.

Finally, we note that an extended version of the model might some day be applicable to non-verbal analogies problems of the type found on the Raven's progressive matrix test. Experience with propositional relationships expressed in verbal form is likely to be of relatively little importance for such problems. However, there is still a potentially important role for an interactive architecture such as ours, in which selection among alternative visuospatial relationships rather than verbal relationships is mutually constrained by the given items in the specified cells of the matrix and the alternative choices provided for the completion of the missing cell. In such a model we would expect that we would observe, and be able to simulate, a role for factors similar to those at work in the current model, including relative familiarity of the correct relation and extent of cognitive control needed to allow a relation common to different rows or columns of the matrix to win out in competition with others.

## ACKNOWLEDGMENTS

This research was supported by a scholarship from the A.G. Leventis Foundation and was conducted at Stanford University. We thank Gary Lupyan and our anonymous reviewers for helpful feedback and constructive comments.

## REFERENCES

Abramczyk, R., Jordan, D., and Hegel, M. (1987). Reverse stroop effect in the performance of schizophrenics. *Percept. Motor Skills* 56, 99–106. doi: 10.2466/pms.1983.56.1.99

Achenbach, T. (1970). Standardisation of a research instrument for identifying associative responding in children. *Dev. Psychol.* 2, 283–291. doi: 10.1037/h0028748

Achenbach, T. (1971). The childrens associative responding test: a two-year follow-up. *J. Edu. Psychol.* 61, 340–348. doi: 10.1037/h0029898

Baker, S., Rogers, R., Owen, A., Frith, C., Dolan, R., Frackowiak, R., et al. (1996). Neural systems engaged by planning: a pet study of the tower of london task. *Neuropsychologia* 34, 515–526. doi: 10.1016/0028-3932(95)00133-6

Brown, A. (1989). "Analogical reasoning and transfer: what develops?" in *Similarity and Analogical Reasoning*, eds S. Vosniadou and A. Ortony (New York, NY: Cambridge University Press), 199–241.

Bunge, S., Wendelken, C., Badre, D., and Wagner, A. (2005). Analogical reasoning and prefrontal cortex: evidence for separable retrieval and integration mechanisms. *Cereb. Cortex* 15, 239–249. doi: 10.1093/cercor/bhh126

Chapman, L., Chapman, J., and Miller, G. (1964). "A theory of verbal behavior in schizophrenia," in *Progress in Experimental Personality Research, Vol 1*, ed B. Maher (San Diego, CA: Academic Press), 136–167.

Cohen, J., and Servan-Schreiber, D. (1992). Context, cortex and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychol. Rev.* 99, 45–77. doi: 10.1037/0033-295X.99.1.45

Cornblatt, B., Lenzenweger, M., and Erlenmeyer-Kimling, L. (1989). A continuous performance test, identical pairs version: Ii. contrasting attentional profiles in schizophrenic and depressed patients. *Psychiatry Res.* 29, 65–85. doi: 10.1016/0165-1781(89)90188-1

Crisafi, M., and Brown, A. (1986). Analogical transfer in very young children: combining two separately learned solutions to reach a goal. *Child Dev.* 57, 573–576. doi: 10.2307/1130371

Dilkina, K., McClelland, J., and Plaut, D. (2010). Are there mental lexicons? the role of semantics in lexical decision. *Brain Res.* 1365, 66–81. doi: 10.1016/j.brainres.2010.09.057

Doumas, L., Hummel, J., and Sandhofer, C. (2008). A theory of the discovery and predication of relational concepts. *Psychol. Rev.* 115, 1–43. doi: 10.1037/0033-295X.115.1.1

Duncan, J., Burgess, P., and Emslie, H. (1995). Fluid intelligence after frontal lobe lesions. *Neuropsychologia* 33, 261–268. doi: 10.1016/0028-3932(94)00124-8

Duncker, K. (1945). On problem solving. *Psychol. Monogr.* 58, i–113. doi: 10.1037/h0093599

French, R. (2002). The computational modeling of analogy-making. *Trends Cogn. Sci.* 6, 200–205. doi: 10.1016/S1364-6613(02)01882-X

French, R. (2008). Relational priming is to analogy-making as one-bal juggling is to seven-ball juggling. *Behav. Brain Sci.* 31, 384–385. doi: 10.1017/S0140525X0800455X

Gentile, J., Kessler, D., and Gentile, P. (1969). Process of solving analogy items. *J. Educ. Psychol.* 60, 494–502. doi: 10.1037/h0028379

Gentile, J., Tedesco-Stratton, L., and Davis, E. (1977). Associative responding versus analogical reasoning by children. *Intelligence* 1, 369–380. doi: 10.1016/0160-2896(77)90019-8

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cogn. Sci.* 7, 155–170. doi: 10.1016/S0364-0213(83)80009-3

Gentner, D. (1988). Metaphor as structure mapping: the relational shift. *Child Dev.* 59, 47–59. doi: 10.2307/1130388

Gentner, D., Holyoak, K., and Kokinov, B. (2001). *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge: MIT Press.

Gentner, D., and Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cogn. Sci.* 10, 277–300. doi: 10.1207/s15516709cog1003_2

Gick, M., and Holyoak, K. (1983). Schema induction and analogical transfer. *Cogn. Psychol.* 15, 1–38. doi: 10.1016/0010-0285(83)90002-6

Goswami, U. (1991). Analogical reasoning: What develops? a review of research and theory. *Child Dev.* 62, 1–22. doi: 10.2307/1130701

Goswami, U., and Brown, A. (1989). Melting chocolate and melting snowmen: analogical reasoning and causal relations. *Cognition* 35, 69–95. doi: 10.1016/0010-0277(90)90037-K

Goswami, U., and Brown, A. (1990). Higher-order structure and relational reasoning: contrasting analogical and thematic relations. *Cognition* 36, 207–226. doi: 10.1016/0010-0277(90)90057-Q

Hinton, G. (1981). "Implementing semantic networks in parallel hardware," in *Parallel Models of Associative Memory*. eds G. Hinton and J. Anderson. (Hillsdale, NJ: Erlbaum), 191–217.

Hodges, J. (2000). "Memory in the dementias," in *The Oxford Handbook of Memory*, eds E. Tulving and F. Craik (New York, NY: Oxford University Press), 441–459.

Hofstadter, D. (2001). "Analogy as the core of cognition," in *The Analogical Mind: Perspectives from Cognitive Science*, eds D. Gentner, K. Holyoak, and B. Kokinov, (Cambridge, MA: MIT Press), 499–538.

Holyoak, K. (2005). "Analogy" in *The Cambridge Handbook of Thinking and Reasoning*, (New York, NY: Cambridge University Press), 117–142.

Holyoak, K., Junn, E., and Billman, D. (1984). Development of analogical problem-solving skill. *Child Dev.* 55, 2042–2055. doi: 10.2307/1129778

Holyoak, K., and Thagard, P. (1995). *Mental Leaps: Analogy in Creative Thought*. Cambridge: MIT Press.

Hummel, J., and Holyoak, K. (1997). Distributed representations of structure: a theory of analogical access and mapping. *Psychol. Rev.* 104, 427–466. doi: 10.1037/0033-295X.104.3.427

Kotovsky, L., and Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Dev.* 67, 2797–2822. doi: 10.2307/1131753

Leech, R., Mareschal, D., and Cooper, R. (2008). Analogy as relational priming: a developmental and computational perspective on the origins of a complex cognitive skill. *Behav. Brain Sci.* 31, 357–414. doi: 10.1017/S0140525X08004469

Markman, A., and Gentner, D. (1993). Structural alignment during similarity comparisons. *Cogn. Psychol.* 23, 431–467. doi: 10.1006/cogp.1993.1011

Martin, A., Wiggs, C., Ungerleider, L., and Haxby, J. (1996). Neural correlates of category-specific knowledge. *Nature* 379, 649–652. doi: 10.1038/379649a0

McClelland, J. L. (1986). "The Programmable blackboard model of reading," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol II.*, eds McClelland, J. L., Rumelhart, D. E., and the PDP research group. (Cambridge, MA: MIT Press), 123–169.

McClelland, J. (1993). "The GRAIN model: a framework for modeling the dynamics of information processing," in *Attention and Performance xiv: synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, eds D. Meyer and S. Kornblum 655–688).

McClelland, J. (2010). Emergence in cognitive science. *Topics Cogn. Sci.* 2, 751–770. doi: 10.1111/j.1756-8765.2010.01116.x

McClelland, J. (2012). *PDP Software.* Available online at:http://psych.stanford.edu/ jlm/software.html

McClelland, J., Rumelhart, D., and the PDP research group (Eds.). (1986). *Parallel Distributed Processing: Explorations in the Mictostructure of Cognition. Vol. 2: Psychological and Biological Models*. MIT Press.

Morrison, R., Doumas, L., and Richland, L. (2011). A computational account of childrens analogical reasoning: balancing inhibitory control in working memory and relational representation. *Dev. Sci.* 14, 516–529. doi: 10.1111/j.1467-7687.2010.00999.x

Morrison, R., Krawczyk, D., Holyoak, K., Hummel, J., Chow, T., Miller, B., et al. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *J. Cogn. Neurosci.* 16, 260–271. doi: 10.1162/089892904322984553

Mummery, C., Patterson, K., Wise, R., Vandenbergh, R., Price, C., and Hodges, J. (1999). Disrupted temporal lobe connections in semantic dementia. *Brain* 122, 61–73. doi: 10.1093/brain/122.1.61

Osherson, D., Perani, D., Cappa, S., Schnur, T., Grassi, F., and Fazio, F. (1998). Distinct brain loci in deductive versus probabilistic reasoning. *Neuropsychologia* 36, 369–376. doi: 10.1016/S0028-3932(97)00099-7

Pauen, S., and Wilkening, F. (1997). Children's analogical reasoning about natural phenomena. *J. Exp. Psychol.* 67, 90–113. doi: 10.1006/jecp.1997.2394

Piaget, J., Montangero, J., and Billeter, J. (1977). *La formation des correlats*. Paris: Presses Universitaires de France.

Prabhakaran, V., Smith, J., Desmond, J., Glover, G., and Gabrieli, J. (1997). Neural subtrates of fluid reasoning: an fmri study of neocortical activation during performance of the ravens progressive matrices test. *Cogn. Psychol.* 33, 43–63. doi: 10.1006/cogp.1997.0659

Rattermann, M., and Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Childrens performance on a causal-mapping task. *Cogn. Dev.* 13, 453–478. doi: 10.1016/S0885-2014(98)90003-X

Richland, L., Morrison, R., and Holyoak, K. (2006). Childrens development of analogical reasoning: insights from scene analogy problems. *J. Exp. Child Psychol.* 94, 249–273. doi: 10.1016/j.jecp.2006.02.002

Richland, L., and Burchinal, M. (2013). Early executive function predicts reasoning development. *Psychol. Sci.* 24, 87–92. doi: 10.1177/0956797612450883

Rogers, T., Lambon Ralph, M., Garrad, P., Bozeat, S., McClelland, J., Hodges, J., and Patterson, K. (2004). The structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychol. Rev.* 111, 205–235. doi: 10.1037/0033-295X.111.1.205

Rumelhart, D. E., and McClelland, J. L., (1982). An interactive activation model of context effects in letter perception: part 2. The context enhancement effect and some tests and extensions of the model. *Psychol. Rev.* 89, 60–94. doi: 10.1037/0033-295X.89.1.60

Rumelhart, D., McClelland, J., and the PDP research group (Eds.). (1986a). *Parallel Distributed Processing: Explorations in the Mictostructure of Cognition: Foundations*. Vol. 1, (Cambridge, MA: Bradford Books/MIT Press).

Rumelhart, D., Hinton, G., and Williams, R. (1986b). "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Mictostructure of Cognition: Foundations*. Vol. 1, eds D. Rumelhart, J. McClelland, and the PDP research group (Cambridge, MA: Bradford Books/MIT Press), 1–34.

Shallice, T., and Burgess, P. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain* 114, 727–741. doi: 10.1093/brain/114.2.727

Spearman, C. (1923). *The Nature of Intelligence and the Principles of Cognition*. London: Macmillan.

Spence, D., and Owens, K. (1990). Lexical co-occurrence and association strength. *J. Psycholinguist. Res.* 19, 317–330. doi: 10.1007/BF01074363

Sternberg, R. (ed.). (1982). "Reasoning, problem solving, and intelligence," in *Handbook of Human Intelligence*. (Cambridge: Cambridge University Press), 227–351.

Sternberg, R., and Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. *Child Dev.* 51, 27–38. doi: 10.2307/1129586

Stuss, D., and Benson, D. (1984). Neuropsychological studies of the frontal lobes. *Psychoogical Bulletin* 95, 3–28. doi: 10.1037/0033-2909.95.1.3

Thibaut, J., French, R., Vezneva, M., Gerard, Y., and Glady, Y. (2011). "Semantic analogies by young children: testing the role of inhibition," in *European Perspectives on Cognitive Science*, eds B. Kokinov, A. Karmiloff-Smith, and N. J. Nersessian. New Bulgarian University Press.

Tunteler, E., and Resing, W. (2002). Spontaneous analogical transfer in 4-year-olds: a microgenetic study. *J. Exp. Child Psychol.* 83, 149–166. doi: 10.1016/S0022-0965(02)00125-X

Turney, P., and Littman, M. (2003). *Learning Analogies and Semantic Relations* (Tech. Rep. No. ERB- 1103 (NRC #46488)). National Research Council, Institute for Information Technology. Available online at: http://arxiv.org/abs/cs/0307055

Waltz, J., Knowlton, B., Holyoak, K., Boone, K., Mishkin, F., Menezes Santos, M. de, et al. (1999). A system for relational reasoning in human prefrontal cortex. *Psychol. Sci.* 10, 119–125. doi: 10.1111/1467-9280.00118

Wharton, C., Grafman, J., Flitman, S., Hansen, E., Brauner, J., Marks, A., et al. (2000). Toward neuroanatomical models of analogy: a positron emission tomography study of analogical mapping. *Cogn. Psychol.* 40, 173–197. doi: 10.1006/cogp.1999.0726

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 19 June 2013; accepted: 28 October 2013; published online: 20 November 2013.*

*Citation: Kollias P and McClelland JL (2013) Context, cortex, and associations: a connectionist developmental approach to verbal analogies. Front. Psychol. **4**:857. doi: 10.3389/fpsyg.2013.00857*

*This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*

## APPENDIX

### TRAINING SET DESIGN

As mentioned in the Materials and Methods section we used *cells* of knowledge. Each network's environment consists of 8 cells of knowledge. The 8 cells have the same basic structure (as shown in **Figure A1A** and in the main text), but they are augmented with additional propositions. The complete design is explained here. The network's world consists of 32 items and 8 relational *prototypes* (i.e., types of relations). Among the 32 items, 16 appear in the first slot of a relational proposition (8 of them act as A items in propositions and 8 act as C items) and the other 16 are fillers of the second slot (8 act as B and 8 act as D). These 32 items and 8 relations are used in such a way across 8 cells of knowledge in the training set that will provide necessary frequency counterbalancing. The cell, as shown in **Figure A1B**, is divided in two parts, one which contains A:B propositions and one which contains C:D propositions. The two types of propositions have no qualitative differences. They are labeled differently because they serve different roles in the analogy questions we created. In each of the two parts, there are four propositions two of which have a frequency of 3, one with a frequency of 1 and one with a frequency of 5. In summary, a cell has four A:B propositions that have frequencies 3, 1, 3, and 5 and four C:D propositions that again have frequencies 3, 1, 3, and 5. The first A:B proposition is designated to be the source part of the analogy question for that cell. The first C:D proposition is assumed to be the correct response and the other 3 C:D propositions are put as incorrect responses in analogy questions to generate various SFI-type problems.

Since, the first propositions in the two sets are the ones that provide a correct analogy, they share the same relational prototype (but with different instances), while all others have different relations. Analogy tests are obtained by taking the first proposition in the A:B part and the four propositions in the C:D part. The other A:B propositions exist for counterbalancing purposes. Now we need to clarify how counterbalancing occurs. Counterbalancing follows a simple rule. We will try to make all As, Bs, Cs, Ds, and Rs to appear in all possible propositions-rows, so that in total they will all have the same frequency.

**Figure A1B** shows the general format of a cell. All A:B and C:D propositions in cell $j$ use the same $A_j$ and $C_j$. Thus, all A and C items have the same frequency in the training set (frequency of 12). Within each cell there is one main relation, and six additional ones. The main relation used in cell $j$ is relation $Rj$. Except for the first A:B proposition and the first C:D proposition that are assigned relation $R_j$, all other propositions are assigned a sequence of relations that starts from $R_{j+1}$ ($R_8$ is followed by $R_1$). This way each relation appears in all possible row-propositions and provides necessary relational counterbalancing. For example relation $R_1$ would appear in the first and fourth proposition in cell 1, in the second proposition in cell 8, in the third proposition in cell 7, and so forth, and would not appear at all in cell 2. Finally, each cell has a subset of the 8 B and the 8 D items. This subset is a pseudo-random permutation of 4 integers from 1 to 8. These permutations were constrained. One constraint is that the B (or D) items in a single cell have all to be different. The item $B_i$ cannot appear twice in a cell. Also, each $B_i$ (and $D_i$) had to appear four times in the entire training set (appear in four cells) appearing in each cell in a different proposition-row. This way all B and D items appear once and only once in all four different row types through the training set. A final constraint is that if a B (and D) and a specific relation R were associated in a given proposition, they should not be associated in a different proposition elsewhere in the training set. That is for controlling for the conditional probability of an item given another item. The number of possible B and D permutation assignments gives us the freedom to create different training sets and test the network in a number of random training environments that obey the same principles, giving us this way more reliable results.

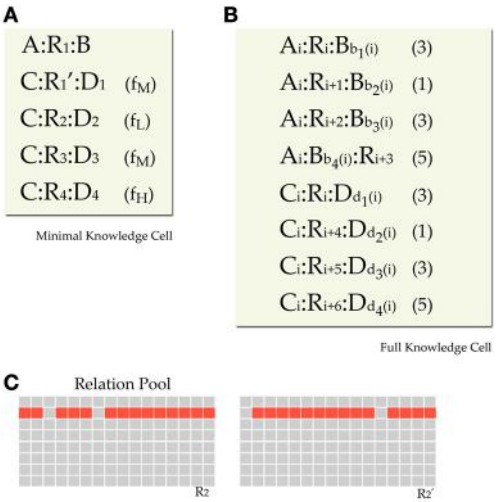



**FIGURE A1 | Model details. (A)** Minimal cell design as described in Materials and Methods. **(B)** Final design of a knowledge cell augmented with propositions that serve counterbalancing purpose. Each cell has 8 propositions (one source, one correct target, three foils, and three additional). All A, B, C, and D items are reused in various propositions in various cells for counterbalancing frequencies. All A items have the same frequency between them by being used in all first 4 propositions of a cell. Similarly for C items in bottom 4 propositions. All relational prototypes appear in all 8 rows as 8 different instances. A prototype appears in two rows (source and target) in its respective cell and shift in other cells. Items B and D appear in their respective 4 rows in different cells. We constrain them so as to appear in different rows in different cells and not reappear associated with the same A/C or R items. **(C)** Activation of two instances ($R_{2_1}$ and $R_{2_2}$)derived from the same prototype **R2**. Each 16 units are dedicated to a relational prototype. For each instance we turn off 2 randomly chosen unique units.

### RELATIONAL REPRESENTATIONS

We also need to clarify our assumptions for relational patterns representations. As we said before relations in the training set come from relational prototypes. For two different relations the prototypes have no overlap at all on their set of active units. But not all occurrences of a prototype are the same. Each relational representation appears in relational instances instead of prototypes. As described above, there are 8 instances of a prototype in the training set. Instances have the same inactive units and have high correlation of active units. An instance is obtained by turning off two specific units of the active units of the prototype. The turned off units are necessarily different across instances. Each

prototype has 16 units on, and two of them are turned off in each instance. So, two distinct instances of a prototype have an overlap of 12 active units. **Figure A1C** shows examples of relational prototypes and instances.

#### NETWORK PARAMETERS

The A and C pools of the network had 16 units (one for each A and C item in the training set) and the B and D pools had 16 units as well (one for each B and D item). The hidden pools had 110 units and the R pool had 128 units. As noted earlier patterns in the A, B, C, and D pools were localist and in the R pool distributed. Weights were initialized to have random values between −0.25 and 0.25. Activation at each time-step was computed by the logistic function of the net input. We trained 5 networks for 350 epochs with the backpropagation through time algorithm. We used 7 intervals and 4 ticks. For training, input was clamped during the entire processing, while the target error was computed for the last 2 intervals. Training consisted of all three possible input combinations (A:_:B, A:R:_, and _:B:R). We used a learning rate of 0.001 and weight decay of 0.000001. In each training session noise was added in a relational instance so that in expectation one active prototype unit would be turned off in the target and one inactive would be turned on in the target. We used cross-entropy as an error measure. For testing we clamped the A, B, C, patterns for the entire processing while the D1 and D2 patterns were clamped only for 3 intervals. We used the relative echo of the D1 and D2 units as a response decision-criterion.

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 26 November 2013
doi: 10.3389/fpsyg.2013.00871

# Beyond modeling abstractions: learning nouns over developmental time in atypical populations and individuals

***Clare E. Sims[1]*, Savannah M. Schilling[2] and Eliana Colunga[1]***

[1] *Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO, USA*
[2] *Department of Electrical, Computer, and Energy Engineering, University of Colorado Boulder, Boulder, CO, USA*

***Edited by:***
*Gary Lupyan, University of Wisconsin–Madison, USA*

***Reviewed by:***
*Michael S. Vitevitch, University of Kansas, USA*
*Caitlin Fausey, Indiana University, USA*
*Maria Sera, University of Minnesota, USA*

****Correspondence:***
*Clare E. Sims, Department of Psychology and Neuroscience, University of Colorado Boulder, 345 UCB, Boulder, CO 80309-0345, USA*
*e-mail: clare.holtpatrick@colorado.edu*

Connectionist models that capture developmental change over time have much to offer in the field of language development research. Several models in the literature have made good contact with developmental data, effectively captured behavioral tasks, and accurately represented linguistic input available to young children. However, fewer models of language development have truly captured the process of developmental change over time. In this review paper, we discuss several prominent connectionist models of early word learning, focusing on semantic development, as well as our recent work modeling the emergence of word learning biases in different populations. We also discuss the potential of these kinds of models to capture children's language development at the individual level. We argue that a modeling approach that truly captures change over time has the potential to inform theory, guide research, and lead to innovations in early language intervention.

**Keywords: word learning, computational models of development, language development, neural networks (computer), language disorders**

## INTRODUCTION

At the core of connectionist models is the idea of modeling change over time. Nowhere is this feature more critical than in the modeling of developmental processes, which by definition occur in time. In this review we focus on the domain of semantic development, specifically early word learning, and highlight the characteristics of the connectionist approach that make it well-suited for modeling developmental processes. We illustrate these characteristics by reviewing several prominent connectionist models of word learning. We argue that, however, most of these models do not fully take advantage of the strengths of connectionist models in capturing the temporality of development. We then turn to our own work modeling developmental trajectories in typically developing children and late talkers. Our approach to modeling word learning has captured intriguing patterns of behavior, produced novel predictions, and has promise for exciting future applications. Throughout the paper, we will explore how computational models of word learning add insight to what is known about this developmental process as well as guide further discoveries.

Connectionist models have made significant contributions to our understanding of various phenomena observed in young children (see Munakata et al., 2008 for a review). In the domain of language development, connectionist models have been used to help explain behavioral data, to test mechanistic accounts of language learning, and to inform big theoretical debates (e.g., Smith et al., 2010; Elman, 2011; Seidenberg and Plaut, in press). In general, connectionist models are well suited to model the time-course and emergent properties of processes. This is because learning in connectionist models is incremental and representations are often under-determined in the beginning and learned as a way to solve a particular task. The current review includes only connectionist models. Connectionist models have the ability to capture processes of change over time as well as to capture multiple timescales of learning, all of which, as we argue in this review, makes them a good candidate model for development. Although connectionism is not necessarily the only way to model these aspects of development (e.g., see Yu et al., 2005; Kemp et al., 2007; Xu and Tenenbaum, 2007; Frank et al., 2009), current research suggests that this is an especially promising approach. We will return to this point in the discussion and touch on other developmental modeling approaches.

To assess the current state of the field, we use four criteria to guide our discussion of prior work (see **Table 1**), and to make the comparisons more informative, we focus on the domain of early word learning rather than attempt to do a comprehensive review of connectionist models of language development. The first three criteria we use have been previously established by Christiansen and Chater (2001), who applied them to a review of psycholinguistic models. These criteria are: data contact, task veridicality, and input representativeness. Data contact refers to how well a model matches empirical data and is able to make novel predictions. Task veridicality involves matching the tasks given to the model to tasks used in the behavioral studies which the model aims to capture. Input representativeness is how well the input to the model captures the input available to the person. In addition to these three criteria, we propose one additional point that is crucial to consider in assessing models of development: temporality. This is a model's ability to capture continuous change. These four criteria will guide our review of connectionist models of early word learning.

### DATA CONTACT

The first criterion we will apply to models of early word learning is the ability to make contact with empirical data. A good model should accurately capture the phenomenon of interest in order to make meaningful conclusions about what may be driving or supporting that phenomenon. We further propose that making

**Table 1 | The four criteria used to assess computational models of early word learning in the current paper.**

| Criterion | Description |
|---|---|
| Data contact | The degree to which the model captures the data and makes predictions that can guide and be tested by empirical research |
| Task veridicality | The match between the task given to the model and the behavioral task used with children |
| Input representativeness | The match between information given to the model and information available to children in the linguistic environment |
| Temporality | The ability to capture continuous changes in phenomena |

contact with data entails making informative predictions that can guide and be tested in subsequent research. Connectionist models of language development have satisfied this criterion with varying degrees of success.

One prior model of language development that has successfully met this criterion is the word learning model of McMurray et al. (2012). This model learned to map word forms to object referents in an unsupervised learning paradigm. The authors used their model to make contact with a variety of behavioral phenomena. For example, the model showed a pattern of comprehension preceding production of word-referent mappings, a preference for novel referents for novel word forms (consistent with mutual exclusivity), as well as graded object familiarity effects in novel word-referent mapping.

Importantly, McMurray et al. (2012) also demonstrated that their model provided novel insights and predictions. For example, the model was able to effectively learn words even when many object referents were present for a single given word. This suggests that associative learning is sufficient to support learning in highly ambiguous contexts, which young children arguably face when learning new words. The model also showed word learning dynamically unfolding in different ways at different timescales. At a shorter timescale, the model made initial connections between a single word form and a single object referent. At a longer timescale, the model created more efficient and long-lasting representations of word-referent links. From the model, the authors proposed that shorter timescale learning, including processes of word-referent mapping and word recognition, is supported in the moment by competition dynamics. On the other hand, longer timescale learning, the retention and refinement of initial mappings, is driven by slower associative learning dynamics.

Li et al. (2004) also made contact with data in their model of semantics and phonology in lexical development. In this model, phonological word form and semantic word meaning representations were formed initially, and were then organized and linked together through associative learning. Among other results, this model captured age of acquisition effects in word learning, showing that learning time was positively correlated with vocabulary size once the lexicon had reached a certain size. In terms of insights and predictions, the authors used their model to show that lexical category representations need not be innate. Mappings between phonological and semantic categories can be learned given the kind of input that is available in the linguistic environment of young children.

Finally, Yu (2005) presented a model of how category learning may interact with word learning early in development. This model was set up to explore a proposed feedback loop between perceptual features of objects and linguistic labels in children's category learning. Although Yu accurately captured the reinforcing relationship between category and language learning in children, the model did not clearly demonstrate the dynamics of the bidirectional relationship in question. The model results demonstrated that learning was improved by the presence of word representations compared to when they were removed, though further testing would be needed to strengthen the claim of bidirectionality. In particular, this model would benefit from tests of interactions over time, a point we will return to later when we discuss the fourth criterion of models of language development.

### TASK VERIDICALITY

The next criterion we will explore with respect to connectionist models of semantic development is the match between the task given to the model and the behavioral task used with people, and in this case, children. The need for a model to capture realistic components of experimental tasks must also be balanced with the need to isolate specific processes that may be at work. That is, modelers must decide which aspects of a given task must be included in a model and which are superfluous in terms of explaining phenomena. Although the ultimate goal would be to construct a model that could capture many different tasks, along the lines of constructing a unified theory, adding complexity does not always make for better explanatory value. For example, a hypothetical model that captures visual, auditory, and semantic processing in children's word learning may reproduce behavior more completely, but may not give much insight into each specific process. Various models of language development have struck this balance in different ways.

Regier's (2005) model achieved veridicality in both the training and testing tasks implemented in the model. In this model of word learning, word forms and word meanings were presented and organized into clusters of exemplars, and associative links between these clusters were learned. Over time, the dynamics of the network adjusted the weightings of various dimensions of form and meaning, simulating the dynamics of selective attention to features in word learning. The training task in this model, in which word forms and meanings were presented simultaneously, captured the typical situation of a child receiving simultaneous visual and label input as their parent teaches them new words. To test the model, Regier simulated a typical forced choice word learning task. After exposure to a novel word pattern, the model was presented with the target word form and had to correctly activate the target meaning from among multiple distractor patterns. This simulation is a good match to a common behavioral task administered to children.

Another example of task veridicality can be seen in Mayor and Plunkett's (2010) model of word learning. This model did a particularly good job of isolating specific processes that seem to be especially important in word learning. In an early stage of learning, the model was presented with visual object and acoustic language input, and each type of input was processed separately. Each type of input became organized into similarity-based categories, simulating a child learning perceptual patterns in the environment in an unsupervised manner, without explicit teaching signals or feedback. In a subsequent stage of learning, visual, and auditory input were presented simultaneously and became linked through associative learning, simulating supervised learning of word-object pairs. In this way this model set out to test the idea that specific, distinct learning processes drive language development at different times.

Li et al. (2004) achieved good task veridicality in the training scheme for their model. Phonological word form and semantic word meaning representations were presented simultaneously and interacted bidirectionally over learning. However, the veridicality of the testing tasks used in this model is not as clear. For example, in a test of comprehension the model was first given a phonological word form representation to process, which then fed forward to semantic processing, and finally produced a word meaning. The model was tested for production in a similar way, beginning with word meaning inputs. It is questionable whether performance on real comprehension and production tasks proceeds in this feed forward fashion. A more realistic task may instead include bidirectional interactions at the time of testing as well as training, with partial activations of word forms and meanings mutually influencing each other.

Overall, several models of early word learning have shown strong task veridicality, helping them in turn make contact with behavioral data. Yet another important component of such models that goes hand in hand with incorporating realistic training and testing tasks is using plausible input patterns. That is, a well-designed task simulation is no longer as realistic and meaningful if the input to that task differs dramatically from information that is actually available to young children learning language. Therefore, the need for input that accurately captures realistic and important information available in a child's linguistic environment is the next criterion we will turn to.

## INPUT REPRESENTATIVENESS

Christiansen and Chater (2001) defined input representativeness as the match between information given to the model and information available to the person. In the case of models of semantic development, this means designing inputs for the model that capture realistic patterns of information that are available in the linguistic environments of young children. Like the previous criterion discussed, input representativeness is also related to the idea of isolating specific processes using a model. To guide the design of input that is both simplified and representative, it is helpful to focus on the information that is most relevant to a process and to exclude irrelevant information. For example, in a model of visual processing, it would be important to capture information such as form, orientation, lighting, and contrast. However, while ultra violet light is technically a piece of information present in the system, it is not relevant to human visual processing and therefore would be irrelevant information for such a model. In this same way, it is important in models of language development to determine what information, such as semantic, perceptual, social, or phonological information, is relevant input to the particular phenomenon of interest.

One example of good input representativeness can be seen in Yu's (2005) model of word and category learning. To create input for the model, Yu collected visual and acoustic data from adult subjects. Multiple subjects were recorded while reading a storybook as if they were narrating to a young child. This method captured realistic co-occurrences between the visual objects that were seen on a page and information that was narrated in speech. Importantly, this input captured not only real information in an environment that would be experienced by a young child, but also the temporal order of this information. The combination of visual and acoustic information yielded model input that was highly representative of information available to young children learning language.

As discussed earlier, Mayor and Plunkett (2010) presented a model that learned word-object associations through an unsupervised followed by a supervised phase of learning. The authors designed input patterns that represented the kinds of information that young children would actually get in these two kinds of learning contexts. Initially, during unsupervised learning, the model was given uncorrelated visual object and acoustic word token representations. Later, during supervised learning, the model was given more structured input with simultaneous presentations of a word with its corresponding object representation. The authors referred to this second stage as joint attention, further showing the link between the input and the specific task that was simulated at that point in the model. In this case, the authors achieved input representativeness by matching the characteristics of the input to the specific learning task that was implemented at a given point in time.

Finally, another model discussed earlier demonstrates the balance between input representativeness and isolating specific processes. In their model of word learning, McMurray et al. (2012) designed the input such that auditory word forms and visual object categories were represented locally, by single units in the network. Learned associations between these units were represented in a hidden layer of lexical units. The hidden layer contained many more lexical units than either the word or category layers in order to better capture learning. Altogether, this input was somewhat removed from the level of information that would be readily available in the environment of a young child. The authors' use of localist representations does not allow them to capture certain finer details that real children use in word learning, such as visual scene variations and similarities that support object categorization. However, the authors chose to use localist rather than distributed representations because they offered certain advantages. This input allows the authors to isolate specific learning mechanisms, such as competition between potential lexical representations in referent selection. As the authors discussed, their simplifications in the model helped strengthen their theoretical point about learning mechanisms that may be crucial in early language development.

Of note, both the criteria of input representativeness and task veridicality are important for using a model to identify meaningful theoretical implications. The tasks that are simulated and the input presented to a model must represent at least some characteristics of the tasks and information encountered by children learning language. At the same time, both of these factors must be balanced with the isolation of specific processes. Isolating processes that are theorized to be key to language development allows researchers to conduct targeted tests of theory within their models. Models of language development must strike this balance between accurately representing the context of learning while targeting specific variables and processes that underlie and support the specific developmental phenomena of interest. We now turn to a final proposed criterion for evaluating connectionist models of development.

## TEMPORALITY

The criteria discussed so far are important to consider for any connectionist model. We propose a final criterion that sets developmental models apart: temporality, or the ability to capture continuous changes and the processes that drive that change. Rather than modeling discrete developmental stages, models that account for temporality capture an ongoing process of change. These changes can be characterized as occurring over time, but also could be, more generally, the sequence of developmental milestones reached, or any other continuous, sequential measure. The key is that the processes of change posited by the model can drive change through the appropriate series of milestones. Connectionist models are particularly well-suited to incorporate temporality and have been used to explore learning over multiple timescales. For example, such models can be used to investigate the formation of connections over time as they emerge and develop. However, many models in the domain of early word learning have not fully captured development as a continuous process. Here we will evaluate the connectionist models discussed above with respect to the final criterion of temporality.

First, although Yu (2005) modeled word learning, the model was not evaluated in a way that measured changes over time. The model results only represented the end point of learning in different conditions. Although Yu used the model to explore the idea of a developmental feedback loop between word and object category learning, the dynamics of this relationship were not explored over time. This model did capture behavioral results observed in young children's word learning, but as presented, it did not demonstrate how word learning unfolds as a developmental process.

The other connectionist models discussed above captured developmental processes of language learning more directly by modeling specific changes that take place over multiple time points of learning. However, in two of these models the developmental change was built into the model a priori. For example, in one model there was a major developmental change built into the input (Mayor and Plunkett, 2010). As discussed earlier, Mayor and Plunkett implemented two stages in their model of early lexical learning. An early, unsupervised learning stage was meant to capture the emergence and refinement of perceptual categories in infancy, and a subsequent supervised learning stage was meant to capture word learning through joint attentional events. Although these stages are theoretically grounded and development within each stage was explored, the process of transitioning between these two stages was not captured by the model. Instead, a qualitative change in word learning was represented by an abrupt change in input and training regime, which likely happens as more of a gradual transition in real children.

Another model that does not fully meet the criterion of temporality is that of Li et al. (2004). In this model, the authors also posited two stages of learning: an initial stage in which learning helps establish a rough topography of lexical categories in similarity space and another stage in which learning fine-tunes these representations. The change from one stage to the next was modeled as a gradual transition that unfolded over time, however the parameters guiding this transition were specified a priori in the model. Similarly to Mayor and Plunkett (2010), Li et al. (2004) did investigate continuous developmental changes taking place across stages and throughout the transition period. However, the developmental transition between those stages did not emerge from the modeled processes alone. Therefore, this model captured some extent of temporality, but ultimately resorted to an a priori change in parameters to capture an important part of the developmental process.

The final two models that we have focused on thus far more fully meet the criterion of temporality. These models captured continuous change over time through emergent dynamics rather than changes to input or parameters during the course of learning. For example, Regier (2005) actually made a theoretical point of using a single mechanism to model several word learning phenomena. Some researchers have posited a mechanistic shift from associative to referential learning to explain developmental changes in behavioral patterns of word learning. Regier's model demonstrated that behavioral patterns previously considered evidence for this shift can actually be explained by the dynamics of a single mechanism over time.

Another example of good temporality in a model can be seen in that of McMurray et al. (2012). In their model of word learning, the authors captured continuous developmental change over two time scales. Importantly, the mechanisms at work at each time scale emerged from the network rather than being implemented through explicit changes in the input or architecture. This model showed that immediate, short-term, "situation time" learning was driven by competition dynamics whereas slower, long-term learning and retention were driven by associative dynamics. These dynamics were continuously at play and interacted with each other over development in the model, resulting in temporality.

Taking this kind of developmental perspective in modeling, that is, striving to meet the criterion of temporality, could have important implications and applications. For example, capturing change over time could help to leverage the information we have about children at one point in time to predict how they will learn at a later time point and their future outcomes. This approach could perhaps even provide new opportunities to intervene and improve the learning process for children. In current, ongoing work in our lab we aim to do just this with a developmental model of word learning.

## OUR APPROACH

Our work builds on prior connectionist models of language development. We are interested in exploring how skilled word learning continuously develops and may emerge from general domain processes. This perspective is in line with other connectionist work that demonstrates how complex, smart behavior can emerge from simple learning rules acting over distributed representations (e.g., Rogers and McClelland, 2006; Elman, 2011; McMurray et al., 2012).

The phenomenon of interest is this: children become skilled learners, at least in part, because they know about the different kinds of properties that are relevant for categorizing different kinds of things. Typically developing children show *word learning biases*: they generalize names for solid objects by shape and names for non-solid substances by material (e.g., Landau et al., 1988; Jones et al., 1991; Soja et al., 1991; Soja, 1992; Samuelson and Smith, 1999; Colunga and Smith, 2008). These are termed the shape and material bias, respectively. The evidence suggests that children learn how to learn nouns – and specifically learn how different kinds of properties are relevant for different kinds of things – as a consequence of learning names for things. Each noun the child learns appears to teach the child something general about how to learn new nouns that name things of that same kind, and critically, at the same time, this learned general knowledge constrains and facilitates the types of nouns the child will learn next. This self-constructing developmental loop involving word learning and category learning (see **Figure 1**) has been partially implemented as a connectionist model. A simple neural network trained using contrastive Hebbian learning on a vocabulary structured like that of the average 2-year-old will show attentional biases akin to those of the average 2-year-old when learning new words (Colunga and Smith, 2005).

This relationship between vocabulary structure and word learning biases has been typically characterized in one of two ways: abstract knowledge guides, facilitates, and indeed allows word learning, or the words that have been learned give rise to, create, and in fact constitute generalized knowledge about word learning. Connectionist models implement a version of the latter account – without being given abstract, or rule-like knowledge, the networks acquire different biases for solids and non-solids as they learn individual categories of solids and non-solids instance by instance. Importantly, this modeling approach gives the power to test proposed causal accounts of word learning biases, a point we will return to when we discuss recent results from our lab.

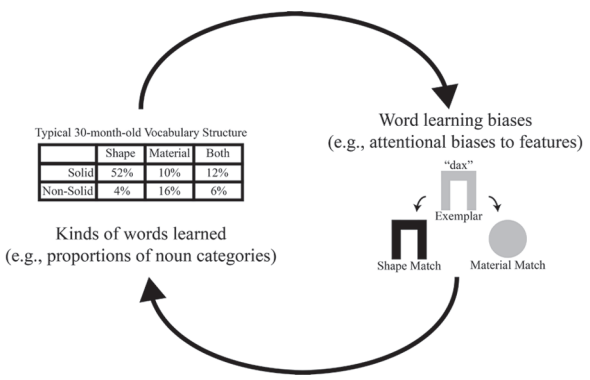



**FIGURE 1 | Schematic of the hypothesized developmental feedback loop.** On the left side are the kinds of words that children learn (represented here as proportions of different categories of nouns, as we used in our model). On the right side are attentional biases to features in word learning (represented here with a classic novel noun generalization task).

The work reviewed here extends these previous findings, and speaks to the criterion of temporality, in two important ways. First, we look at the relationship between vocabulary structure and word learning biases not only *after* the vocabulary of the average 2-year-old has been learned, but *while* this vocabulary is acquired. Second, we look at the relationship between vocabulary structure and word learning biases for children who are not average, but rather late or early talkers relative to their peers. Finally, we look at this relationship based on the vocabulary structure and word learning behavior of individual children between the ages of 18 and 30 months of age.

In the remainder of this paper we first review the evidence for this interactive link between vocabulary growth and the emergence of word learning biases. Then we will introduce our modeling approach and review some results of this approach, both from our neural network model and corresponding behavioral studies of young children. Finally, we will discuss implications and future directions for this developmental approach to modeling word learning.

## WORD LEARNING BIASES

Although there is some debate over the origin of word learning biases (e.g., see Samuelson and Bloom, 2008) some researchers have linked their emergence to the developmental process of vocabulary acquisition. By looking at the shape bias over time, research shows a larger developmental story involving an interplay between attentional biases and vocabulary learning. One study on the emergence of the shape bias tested children longitudinally on their attention to shape in generalizing a novel label (Gershkoff-Stowe and Smith, 2004). These researchers also collected diaries from parents tracking children's vocabulary growth. The results showed that children's attention to shape increased concurrently with increases in the number of nouns in their vocabularies. This suggests that the shape bias emerges in part due to the process of vocabulary growth itself. Another study provides evidence that the emergence of the shape bias can also influence subsequent vocabulary growth. Smith et al. (2002) intensively trained 17-month-old children on labels for novel shape-based categories of objects. The children exposed to this training not only developed a shape bias earlier than is typically seen, they also showed a dramatic increase in vocabulary size over the course of the study compared to a control group. These results suggest that the development of the shape bias accelerates children's learning of object names outside of the lab. Together these studies suggest that (1) the shape bias emerges out of language development, specifically word learning, and (2) as the shape bias emerges it can in turn exert an influence on further vocabulary growth.

Connectionist models of word learning have also helped contribute to understanding of how children may acquire attentional

biases in noun learning. For example, Colunga and Smith (2005) trained a connectionist model on input patterns structured like a typical early child noun vocabulary. These input patterns represented solid objects and non-solid substances which varied systematically in the key features of shape and material. The network input was designed to capture the correlations between solidity and feature for different types of words observed in children's early noun vocabularies (Samuelson and Smith, 1999). Colunga and Smith (2005) then tested the network for generalization to novel test patterns, and found output patterns consistent with shape and material biases. That is, in the generalization test the networks represented novel solid patterns based on similarity in shape rather than material, and represented novel non-solid patterns based on similarity in material rather than shape. This work shows that given input with the correlational patterns found in a typical early child noun vocabulary, a neural network model can similarly acquire selective attentional biases like those observed in toddlers.

In sum, prior research on the origins of the shape bias in toddlers suggests that attentional biases and vocabulary acquisition interact and build on each other over time. That is, selective attention and word learning are both key components of a self-constructing developmental feedback loop in children's early noun learning. Connectionist models of word learning may be a particularly useful way to further investigate and guide empirical studies of this loop. For example, the model of Colunga and Smith (2005) captured part of the feedback loop, showing that the typical early child vocabulary composition has a structure that is sufficient to support the development of attentional biases in generalization. In more recent work in our lab, we have used this modeling approach to further explore the developmental feedback loop in noun learning in a few different ways.

## MODELING THE EMERGENCE OF BIASES

In our work, we use a connectionist model that simulates how children learn words via selective attention to object features, or biases. We focus on a developmental feedback loop in word learning between the kinds of words a child knows and how they learn new words. Importantly, we implement this is in a temporal way, by stopping the network at multiple points during training and testing its performance to capture the trajectory of bias emergence and interactions within the loop. This methodology speaks to our fourth criterion of a good developmental model of language acquisition. In this way, we aim to capture the interactions between different kinds of attentional biases and different types of words which could occur in children's learning as their vocabularies grow. Indeed we have tested predictions made by the networks in a longitudinal study of 18- to 30-month-old children.

We use our model primarily to address the point of temporality, investigating the emergence of word learning biases as vocabulary grows over time. But how does our model measure up against the other key criteria of models of language development? As we will discuss shortly, our model is low in input representativeness. In order to focus on specific processes in learning we must greatly reduce the level of detail of the linguistic information that real children encounter. We do this in principled ways that we believe help us get at our key theoretical questions. By including minimal information in our input patterns, we are able to eliminate other possible factors which could affect word learning and focus specifically on the effect of vocabulary structure on word learning. However, the input to our model represents a subset of the words that a typical child is expected to learn within the first few years of life; therefore, this does not represent all of the linguistic input that children are truly exposed to, as children hear more words than they learn. In terms of task veridicality, we strive to meet this criterion by implementing a simulated version of a common word learning task that is given to children. Finally, we believe our model makes good contact with the behavioral data, as we will discuss in reviewing results from our lab.

Our model is a neural network implemented in the software package Emergent (O'Reilly et al., 2012). It uses the Leabra algorithm (Local, Error-driven and Associative, Biologically Realistic Algorithm), which combines both Hebbian and error-driven learning. The network architecture is adapted from Colunga and Smith (2005) and is shown in **Figure 2**. The word layer represents word labels in a localist way. Previously, we discussed another developmental word learning models' use of localist rather than distributed representations of labels with respect to our third criterion: input representativeness. A distributed pattern of representation provides a model with more information about a given word and how it is similar to other words. This kind of information could be phonological or semantic properties, for example, which are arguably useful in word learning. However, in our model we argue that this kind of information is not necessary to form attentional biases and we focus relevant input to only certain perceptual features. As seen in **Figure 2**, the only distributed patterns of representation in our model are those of the perceptual features, the shape and material, of an item. Solidity is represented discretely, with one unit representing solid and one unit representing non-solid. All of these layers are connected together by a hidden layer which allows the network to form associations between the different perceptual features of the word and the word label itself.

The network is trained with input structured like the noun vocabulary of a typical 30-month-old child. This input structure represents the endpoint of a learning process that we observe over time in the model. This is analogous to a longitudinal study

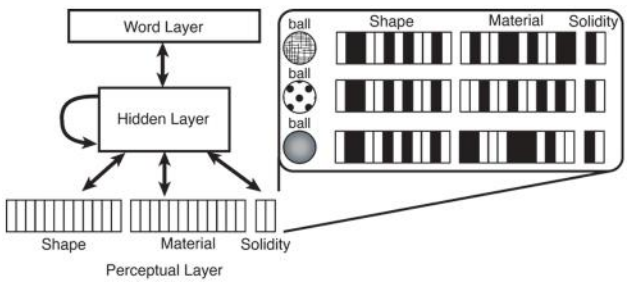



**FIGURE 2 | Network architecture of our word learning model and example input patterns for a solid shape-based noun category.**

of vocabulary growth in children which ends at 30 months of age. The main difference is that in our model we must specify the range of words that are to be learned over time, whereas in children this learning takes place naturalistically within the linguistic environment. To capture this typical early vocabulary structure, we used the 30-month-old vocabulary norms of the MacArthur-Bates Communicative Development Inventory (MCDI; Fenson et al., 1993). We divided this vocabulary into six categories based on solidity (solid or non-solid) and characteristic feature (shape, material, or both); see example words in each category in **Figure 3**. These category divisions were adapted from those used in Colunga and Smith (2005) and were based on adult judgments of solidity and characteristic feature for words in the MCDI[1]. These categories were then transformed into percentages, as shown in **Table 2**. These percentages represent the typical structure of an early child vocabulary, and can be used to create an input vocabulary for the model; in our model we created a 100 word vocabulary input containing the six categories of interest in the proportions shown in **Table 2**.

During training, a word, such as *ball*, is paired with a pattern of features across the perceptual layer (see sample input patterns in **Figure 2**). To simulate learning words for categories of items, each word is presented multiple times and feature patterns along the perceptual layer are manipulated in specific ways. *Ball*, for example, is a word for a solid item characterized by shape; therefore each instance of the word *ball* is represented as the same shape but can vary in material. To implement this computationally, each time the network sees the word *ball*, the pattern along the shape layer (representing, e.g., a round shape) remains the same, but the pattern along the material layer is randomly varied. This is done for each of 100 words in the typical 30-month-old vocabulary structure input presented at each epoch.

In order to capture the developmental trajectory of word learning in the model, we stopped the network at multiple points during word learning and measured its performance on a virtual novel noun generalization (NNG) task. The network was tested after it had learned a certain number of words: at 5 words learned, 10 words learned, and so on. This is a vital feature for a suitable developmental model. By tracking the progress of learning at different time points, based on amount of words learned, we can analyze the emergence and development of word learning biases. The key component which helps our model meet the criterion of temporality is that we not only track development in a temporal manner, but we model it without changing any network parameters. Rather than trying to represent development as discrete stages, we model it as a continuous process and thus focus on the emergence of word learning biases resulting solely from the structure present in the vocabulary input.

Testing was implemented by simultaneously presenting the network with a triad of novel patterns: one exemplar pattern, one pattern matching the exemplar in shape, and one pattern matching the exemplar in material. This virtual NNG task was implemented with both solid and non-solid patterns in order to see whether the network preferred shape or material in the context of each kind of item. In this way, for both the training and testing of the network, we attempt to achieve meaningful task veridicality. Training of the network is similar to a child's word learning in the real world: they are presented with objects (perceptual features) and corresponding labels multiple times as they learn new words. For testing, the task we have implemented is directly analogous to a forced choice NNG task, as the network is presented with an exemplar, and then has to determine which of the two different kinds of feature matches is most similar to the exemplar. Both of these tasks are representative of the behavioral tasks which we aim to model, therefore we achieved good task veridicality in our model.

The network's feature preference (i.e., its attentional bias) was measured based on similarities of hidden layer activations between the exemplar and the two matches. If the hidden layer activations of the exemplar and the shape match were more similar than those of the exemplar and material match, then the network was considered to have a shape bias. If the reverse was true, then the network had a material bias. In this way, we obtained a measure of the extent of attention to each feature in the network over time and we were able to pinpoint the particular point of bias emergence throughout the course of word learning. Using this approach, we have recently explored how different feature biases emerge and develop.

[1]For judgments of solidity, Colunga and Smith (2005) asked adult subjects to answer three questions about each word: (1) Do items named by the word change shape when pressed? (2) Do they return to their original shape after being pressed? (3) Do they take the shape of their container? Words were counted as solid if all questions were answered with "no" and as non-solid is all were answered with "yes." Judgments of characteristic feature were originally gathered by Samuelson and Smith (1999) by asking adult subjects to "indicate for each word which perceptible properties were characteristic across instances of the named category" (p. 5–6)

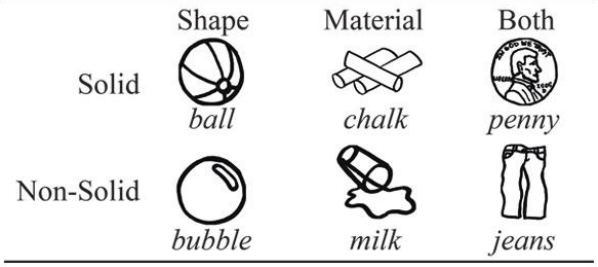



**FIGURE 3 | Noun categories based on adult judgments of solidity (solid or non-solid) and characteristic feature (shape, material, or both) for nouns in the MCDI vocabulary checklist.** Examples of words in the MCDI that fit into each of the six categories of interest are shown.

**Table 2 | Noun categories based on adult judgments of solidity (solid or non-solid) and characteristic feature (shape, material, or both) for nouns in the MCDI vocabulary checklist.**

| | Shape | Material | Both |
|---|---|---|---|
| Solid | 52% | 10% | 12% |
| Non-solid | 4% | 16% | 6% |

*Percentages indicate category representation for a typical 30-month-old vocabulary. An example noun from each category is also provided.*

In Schilling et al. (2012), we used this method to study the interactions between two different kinds of biases, the shape bias and the material bias. We ran 10 instances of the network as described above, then identified the point in the course of word learning where the shape bias emerged for each individual network. We found that, as the shape bias emerged for solid items, the network's attention to material for non-solid items diminished. This finding predicts that children must focus their attention on certain features, such as shape, when developing ways to learn new words and concurrently pay less attention to other features, such as material.

One of our criteria for a developmental language acquisition model was data contact. The results for Schilling et al. (2012) make an important prediction about how children shift and focus their attention to object features in learning new words and we can test this prediction in real children. In Sims et al. (2012), we did this in a longitudinal study of 18- to 30-month-old children. We recruited 20 participants for a monthly, yearlong study beginning at 18 months old age. At each visit, each child was administered a NNG task for both solid objects and non-solid substances to measure their extent of attention of object features and thus their bias development. We also measured their vocabulary growth with the parent-completed MCDI vocabulary checklist. We found that the network predictions were confirmed in children; as children's attention to shape on solid NNG tasks increased around the emergence of the shape bias, their attention to material on non-solid tasks decreased. As a model of word learning bias development, our model satisfies the criterion of data contact and provides novel, meaningful predictions about child language learning.

The aforementioned work makes useful conclusions about one side of the developmental feedback loop, attentional word learning biases, but what about the other side of the loop: vocabulary development? Could there be meaningful interactions in the kinds of words a child learns around the pivotal point of the emergence of the shape bias? These are questions which we hope to address in future research. Recently, in Sims et al. (2013), we have begun to use our model to investigate changes in vocabulary structure around the emergence of the shape bias for solid objects. This work predicts that as attention to shape increases for solid objects, the number of shape-based words which the network learns increases and at a relatively faster rate than that of the material-based words. These results hint at the possibility that certain types of words are learned better or worse at specific moments in development depending on how attention is deployed to specific object features. The state of the vocabulary structure analysis in children is currently inconclusive, but it is the topic of ongoing research in our lab.

From this work, we see that there may be interactions between both shifts in attention and the kinds of words that a child learns. Our neural network model is an important tool for data analysis because it allows us to make predictions about empirical data and to guide behavioral data analysis. Additionally, our model gives us some insight into the potential mechanisms of bias emergence. The model is given only the input of the structure of a child's vocabulary sans any phonological or semantic information and it learns word learning biases just as a child would. This is important because it supports the notion that word learning biases need not be an innate mechanism, but rather a phenomenon which emerges from the structure of a child's noun learning environment. The combination of our model and behavioral data provides useful insight into the developmental trajectory of word learning in toddlers.

## MODELING DIFFERENT POPULATIONS OF CHILDREN

So far we've reviewed how our connectionist network can capture the developmental trajectory of vocabulary growth and the emergence of word learning biases. The next question is, can we use this method to model different kinds of developmental trajectories? This approach may be useful for capturing and explaining meaningful differences among populations of children. Of specific interest are children who fall at two ends of a language endowment spectrum: late and early talkers. These are children who score on the lower and upper ends, respectively, of normative language production measures. These two groups of children differ significantly; a 2-year-old in the bottom 10th percentile may produce around 10 words whereas a 2-year-old in the top 10th percentile will produce well over 300 (Fenson et al., 1993). Late talkers in particular are children who are delayed in vocabulary development, but otherwise show no cognitive or neurological deficits. While some of these children catch up in vocabulary development, others are later diagnosed with Specific Language Impairment, and vocabulary measure norms are not sufficient to predict which children will catch up and which will lag behind (Thal et al., 1997; Rescorla, 2002; Desmarais et al., 2008). It is not yet clear why late and early talkers differ so much in language production, nor why individual late talkers can have such varied outcomes. It may be the case that these populations of children can be characterized by different approaches to word learning, a possibility that we explore with our model.

Variations from the typical trajectory of language development may be due to an interruption in the developmental feedback loop. Referring back to **Figure 1**, we see that the developmental feedback loop demonstrates a relationship between vocabulary structure and word learning biases, so a disruption in either of these factors can cascade and affect word learning. For example, late talkers have relatively small vocabularies and therefore may have less varied and potentially atypical vocabulary structures. Because of this, late talkers can miss out on useful correlational patterns present in the structure of larger, more typical early child vocabularies. For example, say a late talker knows just 13 words, as shown in **Table 3**. This hypothetical late talker knows 10 solid words, four of which are based on shape, another three based on material, and the last three characterized by both shape and material. With this information, there is not enough of a difference in frequency in the kinds of words this child has been exposed to. This child lacks information which typically developing children have (a vocabulary in which most solid words are characterized by shape) thus this late talker has no basis to support the development of a shape bias for solids (or any other bias for that matter). This shows that late talkers can have a deficit in one piece of the loop, vocabulary structure. Subsequently, this gives children who are late talkers less of a basis on which to build

**Table 3 | Example of possible vocabulary proportions for a late talker toddler.**

| Example late talker vocab (no. of words) | | | |
|---|---|---|---|
| | **Shape** | **Material** | **Both** |
| Solid | 4 | 3 | 3 |
| Non-solid | 0 | 0 | 3 |

*For solid words, this child knows approximately equal proportions of words characterized by shape, material, and both shape and material. If a child were developing a bias for solid words based on this vocabulary structure, it would be difficult for them to discern what is an important feature for identifying novel solid words.*

on the other piece of the loop, developing helpful word learning biases. In this case, it is the left side of the developmental feedback loop, the vocabulary acquisition, that is disrupted. Alternatively, the problem could be in the arrow from vocabulary structure to word learning biases; late talkers may struggle with leveraging the correlational structure in the words they already know to abstract higher level attentional biases. We investigated these possibilities in studies of early talker and late talker toddlers and networks.

In one study from our lab, we compared the vocabulary structures of early and late talker children (Colunga and Sims, 2011). We examined age-matched groups of early talker (above the 75th percentile on the MCDI) and late talker (below the 25th percentile) 18- to 30-month-old toddlers. Children's vocabulary structures were analyzed by sorting known nouns into the six categories described earlier, with solidity (solid object or non-solid substance) crossed with characteristic feature (shape-based, material-based, or both shape- and material-based). **Figure 4** shows an example of how raw vocabulary, the words a child knows, is used to create network input patterns. Once the words in the child's vocabulary are grouped into the six categories of interest, the vocabulary is then re-represented as word type proportions with respect to the total number of words in the vocabulary. These proportions are then scaled to 100 word units to create the network input patterns. In this way, it does not matter if the child is an early talker and knows 200 words or a late talker and knows just 10 words. Our model's input patterns are proportions of kinds of words and therefore capture the structure present in children's vocabularies while controlling for vocabulary size differences. Although both late talkers and early talkers knew more words for solid objects characterized by shape than any other category, the vocabularies of these two groups differed qualitatively. The most striking difference was seen in the variability among each group; late talkers showed greater variability in their vocabulary structures compared to early talkers. While early talkers' vocabularies tended to have the same structure as that of the 30-month MCDI norms, late talkers' vocabulary structures took on different forms.

We next used network simulations to explore possible ramifications of these different vocabulary structures (Colunga and Sims, 2011). Each individual early and late talker child's vocabulary structure was given as input to our word learning network described earlier. The networks trained on early talker vocabularies

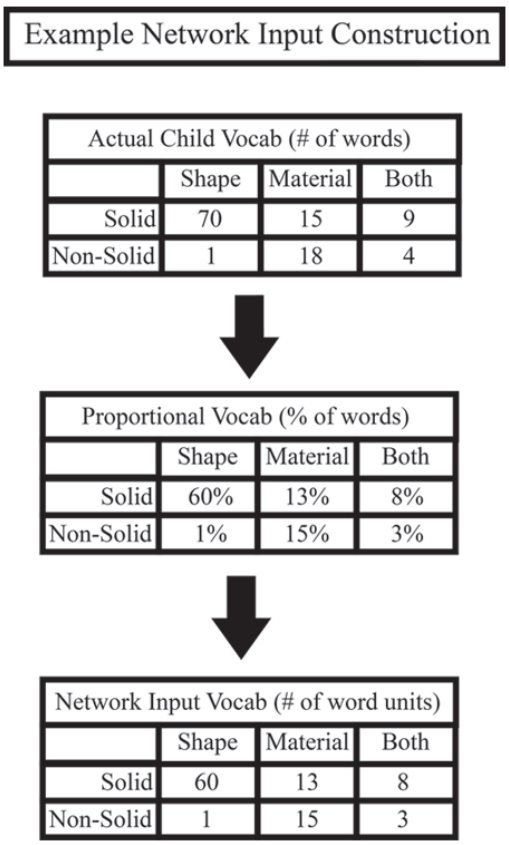



**FIGURE 4 | Example of how our network input is constructed from children's vocabulary data.** First, the nouns that children know from the MCDI are broken into the six categories of interest. These raw numbers are then turned into percentages, normalized by the total number of nouns a child knows. This intermediate representation best illustrates what we mean by vocabulary structure. Finally, these percentages are used to create 100 noun representations that are used as training input for the model.

all developed a shape bias for solids, and the majority also developed a material bias for non-solids. That is, these networks correctly extracted the kinds of attentional biases that have been shown to be helpful in young children's word learning. On the other hand, most but not all of the networks trained on late talker vocabularies developed a shape bias for solids, very few developed a material bias for non-solids, and several actually showed an overgeneralized shape bias for non-solids. The predicted generalization patterns for networks trained on late talker vocabularies significantly differed from those of networks trained on early talker vocabularies.

This first exploration of the developmental feedback loop in different populations of children showed that these children do indeed know qualitatively different kinds of words. Further, the network simulations suggested that these vocabulary differences may carry through and impact the kinds of word learning biases that different groups of children develop. But are differences in vocabulary structure linked to qualitative differences in word learning biases in real children? To answer this question, we brought a sample of early and late talker 18- to 22-month-old

toddlers to the lab to test the predictions of our network (Colunga and Sims, 2012). We tested children on two versions of the NNG task, one involving solid objects and one involving non-solid substances. Early talker toddlers showed a robust shape bias for solids as well as a robust material bias for non-solids. Late talkers also showed a robust shape bias for solids. While late talkers as a group did not show any consistent bias for non-solids, four out of nine of the children in this group showed an overgeneralized shape bias for non-solids, as predicted by the network simulations.

These results provide further evidence for the link between vocabulary composition and word learning biases, that is, the developmental feedback loop. This link has previously been suggested and supported by other research, but the work from our lab makes some new contributions. First, this work uses the powerful approach of modeling language development to make predictions and guide analysis of behavioral data. We use this approach in a novel way, helping to fill in the developmental picture of the relationship between vocabulary structure, that is, the kinds of words that children know, and attentional biases in different populations of children. This approach helps us to isolate the specific role of vocabulary structure and explore how differences in it can result in different attentional biases in early and late talkers. As confirmed by behavioral data, our model showed that early talkers develop helpful word learning biases earlier than the typical population of children, and that late talkers actually do exhibit an early (and sometimes overgeneralized) shape bias.

Second, this work and our modeling approach have provided insight into possible mechanisms that may be driving differences in ability along the language endowment spectrum. Our model and behavioral studies show that early and late talkers exhibit intriguing patterns of differences in both the vocabulary composition and attentional bias components of the developmental feedback loop. By isolating these processes and focusing on a specific piece of children's linguistic environments, the model results suggest that both parts of the self-constructing loop are disrupted in late talkers relative to early talkers and typically developing children.

It is important to note that, thus far, our work in modeling different populations of children has not incorporated temporal analysis. The aforementioned work in both networks and children has focused on the presence or absence of word learning biases at one point in time rather than interactions in bias emergence which occur as attention shifts throughout the trajectory of word learning. In future work, we plan to incorporate temporality into our models of late and early talkers. It is possible that different groups of children exhibit different kinds of interactions of word learning biases and vocabulary structure which could be predictive of future outcome. Exploring how the trajectories of learning differ for children at different points along the language endowment spectrum has the potential to guide diagnosis and intervention. Identifying differences in word learning interactions in early and late talkers could lead to intervention techniques and even early diagnosis of persistent late talkers (i.e., children with Specific Language Impairment). In the next section we will further discuss potential extensions and applications of this work.

## WHAT'S NEXT?

The use of computational models has deepened our understanding of the processes that drive language development. Our work reviewed here shows that a simple connectionist network, embedded in a structured environment, can capture critical characteristics of the developmental trajectories of the emergence of word learning biases in typically developing children as well as in late talkers. One direction that we are pursuing with our model is further exploring the full developmental feedback loop between vocabulary growth and selective attentional biases. So far we have good evidence, both from our model and longitudinal behavioral data, for the emergence of and interactions between the shape and material biases in early word learning. This supports one part of the developmental loop: as vocabulary structure emerges, the development of attentional biases is supported and unfolds dynamically over time. That is, our work both supports and establishes a detailed developmental picture of how word learning leads to the emergence of attentional biases to object features.

Yet this proposed developmental loop is also characterized by a complementary process through which attentional biases guide and influence the course of subsequent vocabulary growth. As discussed earlier, we have begun to explore this part of the loop with our model. So far, results indicate that the emergence of the shape bias for solid items leads to an increase in the rate of shape-based word learning in particular (Sims et al., 2013). An open question is how the later emergence of the material bias for non-solid items influences the learning rate of different kinds of words. Further, once we have established predicted patterns of vocabulary growth in our model, we will test them in our longitudinal study of toddlers. Once completed, this work will have important implications for theories of word learning and the cognitive mechanisms that support this particular part of language development.

Our work focuses on modeling an entire trajectory of word learning and tracking the development of vocabulary and word learning biases at each step along the way. This method is powerful in that it allows us to look past children's current learning and make predictions about language learning outcome. This direction has especially meaningful implications for work with different populations of children, such as those who are developmentally delayed. Thomas et al. (2009) have emphasized the importance of investigating trajectories of learning when studying developmental disorders. Through studies of children with Williams syndrome, Down syndrome, and autism spectrum disorder, these researchers argue that a trajectories approach is "descriptively powerful" for identifying developmental delays and factors that contribute to symptoms. As this work and others have done, we use the powerful approach of studying trajectories to focus on impaired development, particularly in the language development of late talkers. So far we have found evidence to support the idea that the vocabulary structures of late talkers as a group lead to the development of word learning biases that differ from those of early talkers and typically developing children. Next we must explore how these different trajectories unfold over time. That is, we want to go beyond modeling word learning biases at one arbitrary point in time, and instead to model the emergence of and interactions between attentional biases over developmental time among different populations. This work, and subsequent behavioral data analyses, will help to further

elucidate when and how different populations of children diverge from one another along the trajectory of word learning.

All children are not the same, though they are often studied as a single population. In our investigations of late and early talkers, we attempt to target specific groups and identify useful differences in their learning styles. This research aims to separate children based on specific qualities, but what about going even further? Can we model children on an individual level? Work in other fields has also aimed for this goal. For example, Dell et al. (1997) fit a model of lexical networks to individual aphasic patient data. These authors adjusted connection weights and decay rate in each model to match performance levels of each patient. These individualized models were able to make predictions about performance on various speech processing and production tasks. Importantly these models allowed for predictions about patients on an individual level, which could be useful for diagnosis and intervention with specific patients. Similarly, Ziegler et al. (2008) developed a model of dyslexia which they fit to individuals by adjusting levels of noise. These simulations were able to both capture group level dyslexia profiles in the literature and account fairly well for individual reading patterns. As with Dell and colleagues' work with aphasic patients, this kind of modeling work opens up the possibility of targeted intervention. In our own future work, we aim to pair this technique of modeling on the individual level with the study of trajectories of word learning. If we are able to model the trajectories of individual children as they learn new words and grow, we may be able to predict whether or not late talkers in particular will catch up with their peers or what specific intervention techniques could lead to this recovery. We want to utilize the information we have about a child, specifically the words that they know, at one point in time to predict how they will develop word learning biases and subsequently learn new words later in time. While we are confident in our current model's ability to do this at a group level, we may need to further develop the model in order to explore these dynamics and make predictions at the level of individual children.

An additional benefit of modeling developmental trajectories is that it allows us to test predictions which would be difficult or impossible to test in children. Vitevitch and Storkel (2013) demonstrate this point in their model of phonological word learning. These authors implemented manipulations such as reducing cognitive resources and exposing the model to learning environments that might retard typical development. The effects of such manipulations on word learning would be unethical to implement in an experimental study of children, but can be investigated with modeling techniques. Similarly in our future work, we could implement unfavorable word learning environments or induce language impairment in our simulations. These models would allow us to more efficiently test intervention techniques before implementing them with real children. In this way, modeling trajectories of word learning with connectionist models could results in improved techniques of intervention and a proliferation of information on the causes of word learning deficits.

Looking at trajectories of word learning at both the group and individual level, and among different populations of children, will be vital for informing early language interventions. Our hope in investigating the dynamics of the developmental feedback loop in word learning is to pinpoint when and how the processes in this loop may be most receptive to intervention. Models of early, typical, and late talkers will help reveal when in developmental time and at what point in the developmental loop these populations differ in their word learning trajectories. Models of individual trajectories will help to demonstrate how different vocabulary structures lead to different word learning and attentional bias outcomes. Putting this information together, it may be possible to identify an ideal point in development at which to introduce certain types of words or to train certain types of biases in order to facilitate learning for children who would otherwise struggle. Our developmental model of word learning will be a crucial tool in guiding this work and in the creation of intervention strategies that we can test longitudinally with children.

## CONCLUSION

Connectionist models of language that aim to capture and explain developmental processes represent a powerful approach to language research. As reviewed here, such models have satisfied three modeling criteria; they have made good contact with data from real children, captured psychologically valid tasks, and accurately simulated characteristics of young children's linguistic environments. Some of these models have also worked toward our fourth criteria for developmental models, temporality. They capture change in phenomena over time, providing novel insights into the dynamic processes that move children from one point in development to the next. This quality of temporality is important to strive for, particularly because of the potential to better understand and predict the course of development and eventual outcomes. In our model of word learning we have shown that there is continuous, dynamic interplay between the kinds of words learned and selective attentional biases to object features. Patterns of word learning and attentional biases may also provide signatures of learning differences among varied populations and possibly even between individual children. A connectionist modeling approach may help us better understand individual trajectories of language development and, more importantly, design personalized interventions for children who are struggling. Connectionist models of language development are an innovative tool for understanding, diagnosis, and intervention in children's language learning.

## ACKNOWLEDGMENTS

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award number R01 HD067315. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## REFERENCES

Christiansen, M. H., and Chater, N. (2001). Connectionist psycholinguistics: capturing the empirical data. *Trends Cogn. Sci.* 5, 82–88. doi: 10.1016/S1364-6613(00)01600-4

Colunga, E., and Sims, C. E. (2011). "Early talkers and late talkers know nouns that license different word learning biases," in *Proceedings of the 33th Annual*

*Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 2550–2555.

Colunga, E., and Sims, C. E. (2012). "Early-talker and late-talker toddlers and networks show different word learning biases," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 246–251.

Colunga, E., and Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychol. Rev.* 112, 347–382. doi: 10.1037/0033-295X.112.2.347

Colunga, E., and Smith, L. B. (2008). Knowledge embedded in process: the self-organization of skilled noun learning. *Dev. Sci.* 11, 195–203. doi: 10.1111/j.1467-7687.2007.00665.x

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., and Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychol. Rev.* 104, 801. doi: 10.1037/0033-295X.104.4.801

Desmarais, C., Sylvestre, A., Meyer, F., Bairati, I., and Rouleau, N. (2008). Systematic review of the literature on characteristics of late-talking toddlers. *Int. J. Lang. Commun. Disord.* 43, 361–389. doi: 10.1080/13682820701546854

Elman, J. L. (2011). Lexical knowledge without a mental lexicon? *Ment. Lex.* 60, 1–33. doi: 10.1075/ml.6.1.01elm

Fenson, L., Dale, P., Reznick, J. S., Thal, D., Bates, E., Hartung, J., et al. (1993). *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual.* Baltimore, MD: Paul H. Brokes Publishing Co.

Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychol. Sci.* 20, 578–585. doi: 10.1111/j.1467-9280.2009.02335.x

Gershkoff-Stowe, L., and Smith, L. B. (2004). Shape and the first hundred nouns. *Child Dev.* 75, 1098–1114. doi: 10.1111/j.1467-8624.2004.00728.x

Jones, S. S., Smith, L. B., and Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Dev.* 62, 499–516. doi: 10.1111/j.1467-8624.1991.tb01547.x

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Dev. Sci.* 10, 307–321. doi: 10.1111/j.1467-7687.2007.00585.x

Landau, B., Smith, L. B., and Jones, S. S. (1988). The importance of shape in early lexical learning. *Cogn. Dev.* 3, 299–321. doi: 10.1016/0885-2014(88)90014-7

Li, P., Farkas, I., and MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Netw.* 17, 1345–1362. doi: 10.1016/j.neunet.2004.07.004

Mayor, J., and Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychol. Rev.* 117, 1–31. doi: 10.1037/a0018130

McMurray, B., Horst, J. S., and Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychol. Rev.* 119, 831–837. doi: 10.1037/a0029872

Munakata, Y., Stedron, J. M., Chatham, C. H., and Kharitonova, M. (2008). "Neural network models of cognitive development," in *Handbook of Developmental Cognitive Neuroscience*, 2nd Edn, eds C. A. Nelson and M. Luciana (Cambridge, MA: MIT Press), 367–382.

O'Reilly, R. C., Munakata, Y., Frank, M. J., and Hazy, T. E. (2012). *Computational Cognitive Neuroscience.* 1st Edn, Wiki Book. Available at: http://ccnbook.colorado.edu

Regier, T. (2005). The emergence of words: attentional learning in form and meaning. *Cogn. Sci.* 29, 819–865. doi: 10.1207/s15516709cog0000_31

Rescorla, L. (2002). Language and reading outcomes to age 9 in late-talking toddlers. *J. Speech Lang. Hear Res.* 45, 360–371. doi: 10.1044/1092-4388(2002/028)

Rogers, T. T., and McClelland, J. L. (2006). *Semantic Cognition: A Parallel Distributed Processing Approach.* Cambridge, MA: The MIT Press.

Samuelson, L. K., and Bloom, P. (2008). The shape of controversy: what counts as an explanation of development? Introduction to the special section. *Dev. Sci.* 11, 183–184. doi: 10.1111/j.1467-7687.2007.00663.x

Samuelson, L. K., and Smith, L. B. (1999). Early noun vocabularies: do ontology, category structure and syntax correspond? *Cognition* 73, 1–33. doi: 10.1016/S0010-0277(99)00034-7

Schilling, S. M., Sims, C. E., and Colunga, E. (2012). "Taking development seriously: modeling the interactions in emergence of different word learning biases," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 246–251.

Seidenberg, M. S., and Plaut, D. C. (in press). Quasiregularity and its discontents: the legacy of the past tense debate. *Cogn. Sci.*

Sims, C. E., Schilling, S. M., and Colunga, E. (2012). "Interactions in the development of skilled word learning in neural networks and toddlers," in *Proceedings of the 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics* (San Diego, CA), 1–6. doi: 10.1109/DevLrn.2012.6400868

Sims, C. E., Schilling, S. M., and Colunga, E. (2013). "Exploring the developmental feedback loop: word learning in neural networks and toddlers," in *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society).

Smith, L. B., Colunga, E., and Yoshida, H. (2010). Knowledge as process: cued attention and children's novel noun generalizations. *Cogn. Sci.* 34, 1287–1314. doi: 10.1111/j.1551-6709.2010.01130.x

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., and Samuelson, L. K. (2002). Object name learning provides on-the-job training for attention. *Psychol. Sci.* 13, 13–19. doi: 10.1111/1467-9280.00403

Soja, N. N. (1992). Inferences about the meaning of nouns: the relationship between perception and syntax. *Cogn. Dev.* 7, 29–45. doi: 10.1016/0885-2014(92)90003-A

Soja, N. N., Carey, S., and Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meanings: object terms and substance terms. *Cognition* 38, 179–211. doi: 10.1016/0010-0277(91)90051-5

Thal, D. J., Bates, E., Goodman, J., and Jahn-Samillo, J. (1997). Continuity of language abilities: an exploratory study of late- and early-talking toddlers. *Dev. Neuropsychol.* 13, 239–211. doi: 10.1080/87565649709540681

Thomas, M. S., Annaz, D., Ansari, D., Scerif, G., Jarrold, C., and Karmiloff-Smith, A. (2009). Using developmental trajectories to understand developmental disorders. *J. Speech Lang. Hear. Res.* 52, 336. doi: 10.1044/1092-4388(2009/07–0144)

Vitevitch, M. S., and Storkel, H. L. (2013). Examining the acquisition of phonological word forms with computational experiments. *Lang. Speech* doi: 10.1177/0023830912460513

Xu, F., and Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychol. Rev.* 114, 245–272. doi: 10.1037/0033-295X.114.2.245

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: a computational study. *Connect. Sci.* 17, 381–397. doi: 10.1080/09540090500281554

Yu, C., Ballard, D. H., and Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cogn. Sci.* 29, 961–1005. doi: 10.1207/s15516709cog0000_40

Ziegler, J. C., Castel, C., Pech-Georgel, C., George, F., Alario, F., and Perry, C. (2008). Developmental dyslexia and the dual route model of reading: simulating individual differences and subtypes. *Cognition* 107, 151–178. doi: 10.1016/j.cognition.2007.09.004

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 10 July 2013; accepted: 31 October 2013; published online: 26 November 2013.*

*Citation: Sims CE, Schilling SM and Colunga E (2013) Beyond modeling abstractions: learning nouns over developmental time in atypical populations and individuals. Front. Psychol. **4**:871. doi: 10.3389/fpsyg.2013.00871*

*This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*