26 Key Issues and Future Directions: Models of Human Language and Speech Processing

WILLEM ZUIDEMA AND HARTMUT FITZ

Computer programming and mathematics are not typically major components of educational programs in linguistics, and modeling papers only constitute a small fraction of the scientific literature that addresses the nature of languages and the cognitive apparatus to learn and use them. Nevertheless, it is striking how central mathematical and computational models have been and continue to be in many of the big debates in the language sciences, including those about the nature of the cognitive representations underlying language (such as discussed, for instance, in the "past tense debate"; see Pinker & Ullman, 2002, and references therein), innateness (e.g., Elman, Bates, & Johnson, 1998), learnability (e.g., Johnson, 2004), language change (e.g., Gray & Atkinson, 2003) and language evolution (e.g., Fitch, de Boer, Mathur, & Ghazanfar, 2016).

This central role of modeling in many different debates raises a number of questions, including: (a) why are there so many alternative schools of thought in modeling? (b) What are the advantages and attractions of modeling approaches that enable them to be taken so seriously? And (c) why has modeling, despite these advantages, not led to more consensus? In this closing chapter of the part *Modeling Language*, we will briefly consider these questions. We start describing the great variety of modeling approaches, as evidenced in the previous chapters and elsewhere. We will then highlight the advantages of having such a rich modeling toolbox at our disposal and discuss some of the inherent and some of the more easily avoidable causes of disagreement. We end with a number of suggestions for ways forward, including more systematic research on model comparison and a focus on developing models of the neurobiological basis of language.

1. Why Are There So Many Different Modeling Paradigms?

The primary medium of language is speech. Spoken utterances arrive at the ears of the receiver as patterns of vibrating air. The inner ear converts the stream of input into spike trains in the auditory nerve, and the receiver's brain somehow discovers (or imposes) segments in the input, recognizes segments as members of a category (phonemes, syllables, words), recognizes relations between segments in the input, assigns meaning to them, and decides whether the input is well-formed, incomplete, or ungrammatical.

In modeling the neural and cognitive processes involved in interpreting a spoken utterance (and similarly, in production, acquisition, and evolution of written or signed languages), modelers have to make a series of choices and simplify the unorderly, complex reality. Do we focus on the physical reality of the speech signal, on the neural reality of processing in the brain, or on the psychological reality of understanding a received message? The different research traditions reviewed in the previous chapters—including symbolic and neural network modeling paradigms that are often presented as incompatible—start from different answers to these questions.

For instance, the models of syntax and sentence processing reviewed in Demberg and Keller (chapter 22 of this volume), as well as many of the models of language generation reviewed in Krahmer (chapter 25) take abstract syntactic categories and hierarchical structure as a starting point, while greatly simplifying the nature of the signals. The various neural network models discussed in Frank, Monaghan, and Tsoukala (chapter 21) and Zuidema and Le (chapter 23), on the other hand, aim to account for how the empirical observations of linguistic behavior, with what at least to some extent looks like discrete categories and hierarchy, might emerge from the interaction between nodes with continuous activation values in a network. The work discussed in Wehbe, Fyshe, and Mitchell (chapter 24), then, addresses explicitly the relation of these kind of models with detectable activity in the human brain. Finally, the models of speech production and the vocal

tract reviewed in de Boer (chapter 20), stay close to the physical signal.

These chapters already cover an enormous variety of approaches, but the variety of modeling paradigms is even larger than we have been able to represent in the chapters of part IV. One prominent framework not covered is that of Bayesian modeling.[1] Typically, Bayesian models are "rational" models, informing researchers about the *optimal strategy* under the assumed levels of uncertainty, without making direct claims about the actual cognitive processes that humans use. These models thus represent yet another idealization, orthogonal to the ones we already discussed (a good introduction of such Bayesian rational models, applied to the domain of language acquisition, can be found in Pearl & Goldwater, 2016).

We conclude that the great variety of modeling approaches that can be identified in language and speech research is to a large extent the unavoidable consequence of (a) the enormous complexity of language and speech processing, in combination with (b) the necessity of explanatory modeling to simplify. This richness could be an asset for research on language and speech, and rather than interpreting alternative models as rival accounts of language cognition, we would like to stress their complementary strengths. Differences among modeling paradigms are best viewed as different but defensible hard choices on simplifications necessary to reach a deeper understanding of how language and cognition work.

2. What Are the Advantages and Attractions of Modeling Approaches That Enable Them to Be Taken So Seriously?

Our second question concerns the advantages of models that have given them their prominent role in the cognitive and language sciences. An important part of the answer to this question is again the overwhelming complexity of speech and language. With so many interacting components, and so many crucial simplifications to keep track of, researchers need tools to clarify exactly what component in a theory plays what role, where, and when, and moreover they need tools to help derive (potentially counterintuitive) consequences of given assumptions.

The formalization and automation that computational models bring provide exactly these tools. There are endlessly many examples of models of speech and language that illustrate that formalization forces researchers to make theories much more precise, and as a result easier to criticize (which is a good thing). For instance, Demberg and Keller discuss garden path effects and related difficulties in language processing; the many models that have been developed to account for these difficulties have given increasingly precise characterizations of previously multi-interpretable notions such as "incompatibility," "reanalysis," and "semantic plausibility."

There is also a long tradition of using formal models to *automatically* generate predictions. In phonology, morphology, and syntax, the key paradigm of "generative linguistics" even reflects this feature in its name: it is named after a style of model building that is generative, that is, the automatic generation of new linguistic items that can be tested against native speaker intuition. A very different example can be found in Wehbe, Fyshe, and Mitchell, where neural networks are used to predict brain activation given a linguistic stimulus. Although the math for computing the activation of each individual node in these neural networks, and for updating the weights on connections between nodes, are well understood, obtaining the full prediction would be unfeasible without the automation that the implementation of computational models bring.

3. Why Has Modeling, Despite These Advantages, Not Led to More Consensus?

The answer to our third question, about continuing controversy, is more complex, and, ironically, more controversial—even the two authors of this chapter point to different causes for the ongoing controversies. Some of the disagreement about what the right modeling framework is, is unavoidable, due to the vastly different goals and interests of modelers. But some of the disagreements can be traced back to confusions among modelers themselves.

The first author of this chapter points to one particular cause of confusion that can be called "mission creep": The original goals of a modeling framework are gradually forgotten and simplifications that might have been defensible in earlier versions become problematic when models are repurposed to answer new questions. Important areas where we think mission creep plays a role, in many different subdomains of the language sciences, are the issues of how to deal with time (in particular whether and how to abstract out the time dimension), whether and how to convert the continuous physical signal into discrete objects, and how to represent structure in the input stream or sequence of units, including what could be called combinatorial structure, hierarchical structure, and/or slot-filler structure.

For instance, consider the practice of *abstracting out time* in neural network models, such as those discussed in the preceding chapters. These models are all so-called rate-coding networks: the "neurons" in these networks take continuous values, often between −1.0 and +1.0, representing the degree of activity of a single neuron or a group of neurons over a certain time window. The networks are typically organized in layers, and layers are updated one after the other, but all neurons in one layer in one go ("synchronous updating"; this implicitly imposes a global clock that regulates layer-by-layer updating). The exact timing of spikes, differences in oscillation phases, bursting, and other temporal relations between neurons are thus not represented in these models.

But time and timing are central to the study of language processing, which is reflected in many dependent measures used in psycholinguistics (e.g., reading and reaction times, fixation and regression in eye tracking, temporal characteristics of electroenceophalography and magnetoencephalography signals). To put time back into the models, a standard approach is to make the networks *recurrent* (Elman, 1990), but the relation between model time in recurrent networks and physical time in the systems they are meant to model is largely lost.

All this can still be a legitimate simplification, depending on the research questions. However, the simplification becomes problematic when the mission of modelers shifts away from providing an existence proof that neural networks can discover some structure in sentences (Elman, 1990) to questions that directly involve the way the human brain keeps track of time or to questions about supposed fundamental inabilities of neural networks. One example of the latter can be found in the literature on the "binding problem" in neural networks (see Kiela, 2011, for a discussion that requires more space than we have here); models that aim to solve the binding problem are forced to change the way time is represented, such that for instance binding-by-synchrony, or other uses of temporal information, become available.

The second author of this chapter emphasizes another source of confusion, which can be labeled "lack of constraints." Different computational modeling approaches—whether they are connectionist, symbolic, Bayesian, or other—have primarily been concerned with modeling the *output* of the language system, rather than the system itself. The aim of this approach is to redescribe speech and language processing data using some formalism or computational mechanism.

Clearly, however, any finite set of processing data can be captured by an infinity of formalisms, limited only be the imagination of the modeler. Thus, the choice between computational primitives, memory components, representational spaces, input encodings, learning algorithms, and such, is largely unconstrained. This toolbox approach to data recoding has yielded a vast array of distinct models that are based on different assumptions and abstractions. As a consequence, they are often difficult to compare, and rarely has it been possible to decide between architectures in a principled way. As we will discuss in section 4, apart from behavioral data, the neurobiology of the human language system provides a rich set of constraints that can inform the design of computational language models.

4. Ways Forward

The recommendations for the ways forward of the field "models of language" depend on the diagnosis. For the diagnosis "mission creep," the key issue is to continuously reevaluate the simplifications made, including implicit simplifications that modelers are hardly aware off. The remedy we recommend is a model comparison approach, where we develop very different models for the same phenomenon, using very different modeling paradigms and, through careful comparison, highlight the role of simplifications made.

For the diagnosis "lack of constraints," the key issue is to identify underappreciated sources of constraints that stem from characterizing the system itself. We argue that more attention should be paid to the neurobiological basis of language: models should be informed, first and foremost, by the properties of the neurobiological infrastructure that supports language and speech processing. We refer to this approach as *causal modeling* (Fitz et al., 2019).

4.1. Model Comparison Model comparison in the broad sense is standard practice in the modeling literature—every new version of a model is compared with an earlier version—but this broad sense is not what we have in mind here. Rather, we argue that modelers of language and speech should more often systematically compare models from very different modeling paradigms for the same phenomenon: spiking networks versus rate-coding networks, logic-based models of inference versus recursive neural networks (e.g., Bowman, Potts, & Manning, 2015), formal grammar-based models of grammaticality versus recurrent neural networks (e.g., Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018). Such cross-paradigm

comparisons require taking both paradigms seriously, but they shed light on the implicit and explicit assumptions that both paradigms embody and thus yield an opportunity to reevaluate those assumptions.

An early example of such cross-paradigm comparisons yielding important insights is the discovery that context-free grammars, push-down automata, and augmented transition networks are really alternative ways at processing context-free languages and can be translated into each other. Another is the series of discoveries made about the relation between rival syntactic frameworks, including the finding that minimalism, combinatory categorical grammar, and lexicalized tree adjoining grammar have essentially the same binding theory (Steedman & Baldridge, 2011). Zuidema (2003) provided another example, implementing the differential equation model of Nowak, Komarova, and Niyogi (2001) in an agent-based simulation. Detailed comparison of the two models revealed that Nowak et al.'s result of the necessity of a finite and small size of the number of possible grammars (the "size of Universal Grammar") was crucially dependent on the problematic assumption of a uniform probability distribution over possible grammars that the learner is exposed to.

All these examples illustrate the benefits of seriously comparing models from different modeling traditions, accounting for the same phenomenon: Some disagreements disappear, as the supposed contradictions disappear on closer inspection. Other disagreements might remain, but are traced back to differences in more fundamental assumptions.

4.2. Causal Modeling Human language is a neurobiological system implemented as a sparsely connected, recurrently coupled network of highly dynamic neurons and synapses. Existing computational models of language processing have made little or no contact with the detailed biophysical properties of this system and have focused mainly on the reproduction of behavioral data.

Models of behavior start at the computational or algorithmic level of description (in the terminology of David Marr) and attempt to reverse engineer the language system from input-output relations. Recent experimental work, however, has shown that it is difficult to reverse engineer even simple computational systems (e.g., a microprocessor) whose functionality is completely known (Jonas & Kording, 2017). This methodological issue is exacerbated with increased system complexity and the noisiness of measurements made from the human language system.

A causal modeling approach, on the other hand, starts out at the implementational level of description and attempts to synthesize language function in the brain from first principles. These include, for instance, the principle of state-dependence (Buonomano & Maass, 2009) and the principle of information processing as computation over high-dimensional transients (Rabinovich, Huerta, & Laurent, 2008). Moreover, causal modeling attempts to capture the characteristics of the biophysical system itself, rather than to reproduce a particular aspect of behavior. The aim is to understand, through simulation and theoretical insight, how neural computation and memory in cortical circuits support language behavior. The extent to which a causal model approximates the neurobiological infrastructure of the real system determines how humanlike it will behave. Reproduction of behavior, however, is viewed as an independent outcome, not the primary goal of modeling.

Causal models also differ from models of behavior in that parameters have physical units of measurement that need to fall within physiological bounds. This places strong constraints on the model space and reduces degrees of freedom. In addition, many of the arbitrary design choices in models of behavior turn into empirical questions (see, e.g., Bartol et al., 2015). While models of behavior often attempt to fit data with as few parameters as possible, the challenge for causal models is to deal with the abundance of parameters provided by the neurobiological system (e.g., $\sim 10^{14}$ synaptic conductances).

Also the issue of time, which we discussed in section 3, can be dealt with from the causal modeling perspective. Here, network time corresponds to real physical time, since it arises from biophysical models of spike generation and the dynamics of synaptic transmission (Gerstner, Kistler, Naud, & Paninski, 2014). Due to this nomological relation, causal models allow us, in principle, to investigate how language and speech processing unfold over time at any desired grain size (e.g., on a millisecond scale).

Taking a causal modeling approach to language does not mean that we ought to replicate the neurobiological substrate at all levels of detail. Even causal models can (and should) be high-level abstractions of the underlying physiology in that they are composed of mathematically reduced but phenomenologically adequate parts. Many such parts have been described in computational neuroscience, for instance, the adaptive-exponential neuron (Brette & Gerstner, 2005), models of short-term synaptic facilitation and depression (Markram, Wang, & Tsodyks, 1998), mechanisms of homeostatic plasticity

(Vogels, Sprekeler, Zenke, Clopath, & Gerstner, 2011), long-term potentiation (Clopath, Büsing, Vasilaki, & Gerstner, 2010), and principles of synaptic consolidation (Clopath, Ziegler, Vasilaki, Büsing, & Gerstner, 2008). Using these experimentally validated components ensures that models will gradually begin to approximate the dynamic properties of the human language system.

As this chapter has attempted to highlight, it has been difficult to unify computational approaches to language and speech processing into a coherent framework. On the view outlined here, we should strive to replace data modeling by models of the neurobiological language system itself. This would mark a much needed paradigm shift in computational language modeling. The long-term goal of this approach is not to describe processing in terms of interacting neurons and synapses, but to distill an algorithmic abstraction from the neurobiological substrate that characterizes cognitive function.

There are many challenges involved in causal language modeling. For instance, complex networks of spiking neurons with plastic synapses are computationally costly to simulate, and their behavior is often difficult to interpret. These challenges can be overcome through the development of suitable neuromorphic architectures, novel approaches such as event-driven simulation, and large-scale team efforts to analyze and understand the computational role of component parts through model comparisons. Eventually, this approach might be able to bridge descriptions of brain function across all levels of explanation (Carandini, 2012) and yield a causal, mechanistic understanding of the human capacity for language.

NOTE

1. Models in the Bayesian framework define probability distributions over potentially quite rich structures (ranging from phonetic structure, as in de Boer, to syntactic structure, as in Demberg & Keller). The probability distributions are defined indirectly, by setting up a stochastic process (a "generative model") that generates the linguistic structures of interest. Using a rich toolbox of computational techniques, the probability distribution over possible data is used to compute the probability of observed data given a model and, using Bayes' law, to find the model or models that make the data maximally probable.

REFERENCES

Bartol, T. M., Jr., Bromer, C., Kinney, J., Chirillo, M. A., Bourne, J. N., Harris, K. M., & Sejnowski, T. J. (2015). Nanoconnectomic upper bound on the variability of synaptic plasticity. *eLife*, *4*, e10778.

Bowman, S. R., Potts, C., & Manning, C. D. (2015). Recursive neural networks can learn logical semantics. In A. Allauzen, E. Grefenstette, K. M. Hermann, H. Larochelle, S. Wen-tau Yih (Eds.), *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)* (pp. 12–21). Stroudsburg, PA: Association for Computational Linguistics.

Brette, R., & Gerstner, W. (2005). Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of Neurophysiology*, *94*, 3637–3642.

Buonomano, D. V., & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, *10*, 113–125.

Carandini, M. (2012). From circuits to behavior: A bridge too far? *Nature Neuroscience*, *15*, 507–509.

Clopath, C., Büsing, L., Vasilaki, E., & Gerstner, W. (2010). Connectivity reflects coding: A model of voltage-based STDP with homeostasis. *Nature Neuroscience*, *11*, 344–352.

Clopath, C., Ziegler, L., Vasilaki, E., Büsing, L., & Gerstner, W. (2008). Tag-trigger-consolidation: A model of early and late long-term-potentiation and depression. *PLoS Computational Biology*, *4*, 1–14.

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

Elman, J. L., Bates, E. A., & Johnson, M. H. (1998*). Rethinking innateness: A connectionist perspective on development* (Vol. 10). Cambridge, MA: MIT Press.

Fitch, W. T., de Boer, B., Mathur, N., & Ghazanfar, A. A. (2016). Monkey vocal tracts are speech-ready. *Science Advances*, *2*(12), e1600723.

Fitz, H., van den Broek, D., Uhlmann, M., Duarte, R., Hagoort, P., & Petersson, K. M. (2019). Neuronal memory for language processing. *bioRxiv*. https://doi.org/10.1101/546325.

Gerstner, W., Kistler, W. M., Naud, R., & Paninski, L. (2014). *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge: Cambridge University Press.

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*(6965), 435.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni M. (2018). Colorless green recurrent networks dream hierarchically. In M. Walker, H. Ji & A. Stent (Eds.), *Proceedings NAACL-HLT 2018* (pp. 1195–1205). Stroudsburg, PA: Association for Computational Linguistics.

Johnson, K. (2004). Gold's theorem and cognitive science. *Philosophy of Science*, *71*(4), 571–592.

Jonas, E., & Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? *PLoS Computational Biology*, *13*, e1005268.

Kiela, D. (2011). Variable binding in biologically plausible neural networks (Unpublished master's thesis). University of Amsterdam. Retrieved from https://eprints.illc.uva.nl/id/document/1994.

Markram, H., Wang, Y., & Tsodyks, M. (1998). Differential signaling via the same axon of neocortical pyramidal neurons. *Proceedings of the National Academy of Sciences*, *95*, 5323–5328.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2001). Evolution of universal grammar. *Science*, *291*(5501), 114–118.

Pearl, L., & Goldwater, S. (2016). Statistical learning, inductive bias, and bayesian inference in language acquisition. In J. Lidz, W. Snyder, & J. Pater (Eds.), *Oxford handbook of*

developmental linguistics (pp. 664–695). Oxford: Oxford University Press.

Pinker, S., & Ullman, M. T. (2002). The past-tense debate: The past and future of and weaknesses of connectionist and rule-based models of language and cognition. *Trends in Cognitive Sciences*, *6*(11), 456–463.

Rabinovich, M., Huerta, R., & Laurent, G. (2008). Transient dynamics for neural processing. *Science*, *321*, 48–50.

Steedman, M., & Baldridge, J. (2011). Combinatory categorial grammar. In R. Borsley & K. Borjars (Eds.), *Non-transformational syntax: Formal and explicit models of grammar* (pp. 181–224). Hoboken, NJ: Wiley-Blackwell.

Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., & Gerstner, W. (2011). Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science*, *334*, 1569–1573.

Zuidema, W. H. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 51–58). Cambridge, MA: MIT Press.