Dissertation Abstract: Inferential Role Semantics for Natural Language

Peter Blouw

March, 2018

Introduction

Questions concerning the semantics of natural language are both theoretically foundational and deeply contested within the field of cognitive science. On the one hand, theories of meaning are often tasked with the job of explaining language-world connections in terms of reference relations and truth conditions (Lewis, 1970; Soames, 2010; Speaks, 2014). On the other hand, these theories are also tasked with the job of explaining various facts about language use (Brandom, 1994; Horwich, 1998; Wittgenstein, 1953) and the mental states of language users (Block, 1986; Harman, 1982). Given the diversity of these aims, one can legitimately wonder whether the various theories of meaning that have been proposed are really theories of the same sort (Block, 1986). And even if they are of the same sort, one can go on to wonder whether it is possible for a single theory to satisfy all of the aims in question (Horwich, 1998).

Against this backdrop of uncertainty, the goal of my thesis is to argue that "meanings" function mainly to codify certain implicit expectations regarding the effects of language use. Specifically, I propose that linguistic expressions acquire meanings by regulating social practices that involve "intentional interpretation" (Brandom, 1994; Dennett, 1987) wherein people explain and predict one another's behavior through linguistically specified mental state attributions. The purpose of a semantic theory is accordingly to account for how language is able to perform this predictive function in the context of practices that involve people adopting what Dennett (1987) calls "the intentional stance" towards one another. The key to achieving this purpose, I argue, is to formalize the tacit inferential relationships that hold amongst various linguistic expressions and non-linguistic perceptions and actions, and thus explain how tacit expectations are generated on the basis of these relationships. I therefore propose a semantic theory on which the meaning of a linguistic expression is primarily characterized in terms of the inferences it licenses, or its "inferential role."

My thesis develops this theory through a combination of (a) philosophical argumentation, (b) empirical justification, and (c) model specification. Specifically, Chapter 1 outlines the philosophical foundations of the theory and defends them on both methodological and empirical grounds. Chapters 2 and 3 consider the relationship between the theory and formal methods that are designed to induce the meanings of linguistic expressions from statistical regularities found in text corpora and various labeled datasets. Chapter 3 introduces a model that assigns inferential roles to arbitrary linguistic expressions by learning from examples of how sentences are distributed as premises and conclusions in a space of possible inferences. The model is evaluated experimentally, and on this basis argued to provide a promising illustration of the feasibility of a formally precise version of the theory under development.

Chapters 4 through 6 concern the implications of this work with respect to debates about the compositionality of language, the relationship between language and cognition, and the relationship between language and the world. With respect to compositionality, I argue that the debate is really about generalization in language use, and that the required sort of generalization can be achieved by "interpolating" between familiar examples of correct inferential transitions. With respect to the relationship between thought and language, I argue that it is a mistake to try to derive a theory of natural language semantics from a prior theory of mental representation because theories of mental representation invoke the sort of intentional interpretation at play in language use from the get-go. With respect to the relationship between language and the world, I argue that questions about truth conditions and reference relations are best thought of in terms of inferential norms the sustain successful predictions and actions on the part of language users.

Chapter 1

To provide a philosophical underpinning to the project, Chapter 1 introduces what I call the IPA framework for theorizing about language use. The framework characterizes linguistic expressions in terms of the inferences (I) they license, the behavioral predictions (P) that their uses thereby sustain, and the affordances (A) that they provide to language users in virtue of these inferential and predictive involvements. Initially, the framework is motivated by the commonplace observation that language use is a species of joint action (akin to a dance) in which two or more people co-ordinate their behavior in the service of some common purpose (Brandom, 2010; H. Clark, 1996), and by philosophical analyses of the relationship between meaning and use (Brandom, 1994; Sellars, 1953; Wittgenstein, 1953).

On an empirical level, the IPA framework is motivated by a range of evidence. One strand comes from research that emphasizes the importance of social cues and shared attention to early word learning (Bloom, 2001; Tomasello, 2003). A second strand comes from research on language acquisition more generally, which is increasingly being characterized as a kind of skill acquisition, wherein a learner develops the ability to process and use linguistic expressions correctly in the context of cooperative interactions (Christiansen & Chater, 2016; Harley, 2014; Seidenberg, 1997). A third strand of evidence comes from psycholinguistic research suggesting that the formation of predictions about the trajectory of a dialogue or speech event is a core component of language comprehension (Christiansen & Chater, 2016; Pickering & Garrod, 2007, 2013; Rohde, 2008).

Overall, the IPA framework integrates existing theoretical and empirical work to a motivate a theory that derives the meanings of linguistic expressions from their use as informal instruments of prediction when adopting the intentional stance. It accordingly sets the stage for a theory of our intuitive theories concerning communication. What is needed, then, is a more formal theory of how linguistic expressions function as primitives in our implicit theories of the social world.

Chapter 2

To work towards this goal, Chapter 2 examines and interprets a number of methods for mathematically characterizing the use-regularities that are evident in natural language. These methods involve encoding statistical patterns gleaned from linguistic corpora into the elements of real-valued vectors to produce "distributed representations" of linguistic expressions (Baroni, Bernardi, & Zamparelli, 2014; Mitchell & Lapata, 2010; Turney & Pantel, 2010). I examine additive (Landauer & Dumais, 1997), multiplicative (Smolensky & Legendre, 2006), sequential (Elman, 1990), and tree-structured (Socher, Huval, Manning, & Ng, 2012) models for producing such representations, and demarcate the limits within which they can be interpreted as encoding the meanings of the expressions they manipulate.

I then look at more recently developed models that learn to label sentence pairs with specific inferential relationships (see, e.g. Bowman, Angeli, Potts, & Manning, 2015). I show how these models can be used to construct rudimentary inferential roles for arbitrary linguistic expressions by filtering out the sentences that follow from a given input from a larger set of "candidate" sentences.

Chapter 3

To build further on these formal foundations, Chapter 3 introduces a neural network model that learns to generate sentences that are the inferential consequences of its inputs. The model functions by first encoding a sentence into a distributed representation, and then decoding this representation to produce a new sentence. The encoding procedure involves dynamically generating a tree-structured neural network, while the decoding procedure involves feeding the encoded result through an "inverse" tree-structured network to produce a predicted sentence.[1] Through iterations of this encoding-decoding procedure, the model is able to generate numerous further sentences that are the predicted inferential consequences of any input sentence it is provided with.

The rest of the chapter evaluates the model's ability to produce correct entailments for sentences unseen in its training data. I present experimentally produced plausibility ratings for a random collection of generated sentences that indicate that the inferential transitions generated by the model are seen to be nearly as good as gold-standard transitions generated by humans. Next,

[1] To train the model parameters (i.e., the network weights shared across different tree structures) I use the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015)

I provide a number of more qualitative evaluations of the model that involve (a) iterating its predictions to produce chains of inferences (Kolesnyk, Rocktäschel, & Reidel, 2016), (b) performing word-level input substitutions to determine the inferential significance of subsentential expressions (Brandom, 1994), and (c) conditioning the model's predictions on additional inputs in the form of simple prompts and questions.

Overall, the analysis of the model concludes the technical component of the thesis, which aims to demonstrate the formal viability of the inferential role semantics under development. Subsequent chapters deal with the broader conceptual consequences of the theory.

Chapter 4

One important consequence of this work is that it motivates a revision to standard characterizations of the principle of compositionality. In the theory outlined above, for example, the inferences that do the explanatory work during intentional interpretation involve sentences, not words, which makes it difficult to assign inferential roles to subsentential expressions and then use these to "build up" to the sentential level in a compositional manner (Brandom, 1994). The model described in Chapter 3, moreover, does not produce explicitly compositional representations comprised of parts and wholes.

To account for these observations, I propose to translate the question of how meanings compose into the question of how people are able to generalize to the use of expressions beyond those that they have had direct exposure to. The motivation for adopting this strategy is simple: existing arguments to the effect that natural languages are compositional largely take the form of inferences to the best explanation, in which the phenomenon to be explained is the evident generality of our capacities for linguistic comprehension (Szabó, 2012, 2013). Other explanations of this phenomenon are possible (Tomasello, 2003), which makes it inadvisable to treat the principle of compositionality as a presupposition of semantic theory.

I go on to discuss three kinds of generalization – similarity-based, syntactic, and procedural – that are potentially relevant to explanations of how people extrapolate on the basis of prior experience to the correct usage of novel linguistic expressions. I assess the degree to which each form of generalization is relevant to explanations of language use. On the basis of this assessment, I provide a positive account of linguistic generalization that is consistent with both the IPA framework and the formal properties of the model introduced in Chapter 3. The result is an account that emphasizes the explanatory priority of inferential relations amongst sentences, and that describes procedures through which these relations are determined in novel contexts via a kind of "interpolation" between familiar examples of correct inferential transitions. I conclude that only a weak, procedural notion of compositionality is suitable for the analysis of natural language.

Chapter 5

A second important consequence of the theory concerns the relationship between thought and language. Linguistic expressions are often treated as vehicles for transporting thoughts (Fodor, 1998; Korta & Perry, 2015), but this transportation model is arguably incompatible with an inferential role semantics (Brandom, 1994, 2010). The difficulty lies in the fact that an inferential role is not something that gets directly communicated by a linguistic expression; rather, it is something that regulates how the expression is used.

To argue in favor of the plausibility of a theory that avoids positing clear mappings between linguistic expressions and mental representations that encode their meanings, I rely on an insight from Dennett (1987). Namely, there is no reason to describe the states of a system in representational rather than non-representational terms unless one is committed to the applicability of the sort of "linguistic calculus" that is implicit in the use of intentional state attributions. I then draw an important lesson from this insight: one has to *start* with intentional interpretation when theorizing about mental representations, which means that one cannot use these representations to independently *derive* an account of how intentional interpretation works. As such, the order of intentional explanation is importantly "top-down," given that the use of intentional vocabulary is fundamentally rooted in the interpretation of linguistic practices rather than in the interpretation of mental states. Moreover, the point of using representational rather than, say, causal descriptions of a system's internal states is to hook up these descriptions to the linguistic calculus of intentional systems theory.

I then show that rather than conflicting with the methods of contemporary cognitive science, this "priority thesis" concerning intentional interpretation actually helps to explain why these methods are successful. After all, a main goal of cognitive science is to develop explanatory "ladders" that allow one to traverse between theories at different levels of abstraction. The intentional stance gives rise to a theoretical framework concerning overt linguistic behavior. Work in cognitive psychology and neuroscience, by comparison, gives rise to a theoretical framework concerning causal relations amongst states in cognitive systems. The act of labeling these states as representations does the crucial job of building a bridge between these two theoretical frameworks, so as to lessen the mystery of how thought and language are related to one another. On the basis of these philosophical considerations and related empirical considerations, I conclude that an inferential approach to semantics is cognitively quite plausible.

Chapter 6

A final consequence of the theory is that it mandates a slightly non-standard account of the relationship between linguistic expressions and the non-linguistic world. The purpose of this chapter is to characterize this relationship in terms of inferential transitions between linguistic expressions and perceptual or behavioral responses to the surrounding environment,[2] without sliding into a form of solipsism.

The basic idea is that language use is governed by certain socially instituted norms, and these norms are such that the inferences licensed by particular linguistic acts support a broad range of successful predictions and actions. Then, when a linguistic expression is used to make a claim, the claim "gets things right" (i.e., represents the world correctly) insofar as the inferences it licenses sustain success in prediction and action indefinitely.

The principle benefit of this approach to dealing with the intentionality of language is that it avoids the internal confusions found in theories based on more robust notions of the reference relation and the truth property. For one thing, these theories have difficulty accounting for linguistic expressions that do not refer or correspond to the world in a clear-cut manner. For another, they also problematically imply that one can grasp the fundamental

[2] Regarding concerns about whether it is plausible to analyze perception and action in terms of inferential transitions, there is a fair bit evidence to suggest that perception is inferential in the sense that top-down expectations or predictions shape how certain features of the environment are perceived – people essentially *infer* the state of the world from noisy and misleading perceptual data (A. Clark, 2013). So there is nothing obviously wrong with treating perceptions as "premises" from which certain "conclusions" are drawn.

structure of reality by grasping the meanings of linguistic expressions. Finally, the approach is also uniquely consistent with the notion that languages are fallible and continually evolving tools that aid complex organisms (i.e. people) in dealing with their surroundings effectively. I conclude that there is much to recommend an approach to the language-world relation that takes inference rather than reference as its starting point.

Conclusion

In summary, there are two main contributions offered in this thesis. The first, theoretical contribution is to describe an inferential approach to natural language semantics that is motivated by insights and empirical discoveries spanning the fields of philosophy, psychology, and linguistics. The second, technical contribution is to demonstrate how this approach can be formalized using computational models that learn to assign rudimentary inferential roles to arbitrary linguistic expressions. Together, these contributions offer a novel and plausible approach to thinking about the meaning of language, and should therefore be of interest to a large portion of the cognitive science community.

References

Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies*, *9*, 241-346.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, *10*(1), 615-678.

Bloom, P. (2001). Précis of *How Children Learning the Meanings of Words. Behavioral and Brain Sciences*, *24*, 1095-1134.

Bowman, S., Angeli, G., Potts, C., & Manning, C. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics.

Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment.* Harvard University Press.

Brandom, R. (2010). Inferentialism and some of its challenges. In B. Weiss & J. Wanderer (Eds.), *Reading Brandom: On Making it Explicit* (p. 159-180). Routledge.

Christiansen, M., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, 1-72.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181-253.

Clark, H. (1996). *Using language.* Cambridge University Press.

Dennett, D. (1987). *The intentional stance.* MIT Press.

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

Fodor, J. (1998). *Concepts: Where cognitive science went wrong.* New York, NY: Oxford University Press.

Harley, T. (2014). *Psychology of language: From data to theory* (4th ed.). Psychology Press.

Harman, G. (1982). Conceptual role semantics. *Notre Dame Journal of Formal Logic*, *23*, 242-256.

Horwich, P. (1998). *Meaning.* Oxford University Press.

Kolesnyk, V., Rocktäschel, T., & Reidel, S. (2016). Generating natural language inference chains. *arXiv preprint arXiv:1606.01404*.

Korta, K., & Perry, J. (2015). Pragmatics. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy.* CSLI Publications.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Lewis, D. (1970). General semantics. *Synthese*, *22*(1), 18-67.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*, 1388-1429.

Pickering, M., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Science*, *11*, 105-110.

Pickering, M., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*, 329-392.

Rohde, H. (2008). *Coherence driven effects in sentence and discourse processing* (Unpublished doctoral dissertation). University of San Diego.

Seidenberg, M. (1997). Language acquistion and use: Learning and applying probabilistic constraints. *Science*, *275*, 1599-1603.

Sellars, W. (1953). Inference and meaning. *Mind*, *62*(247), 313-338.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar* (Vol. 1). MIT Press.

Soames, S. (2010). *Philosophy of language.* Princeton University Press.

Socher, R., Huval, B., Manning, C., & Ng, A. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (p. 1201-1211). Association for Computational Linguistics.

Speaks, J. (2014). Theories of meaning. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy.* CSLI Publications.

Szabó, Z. (2012). The case for compositionality. In M. Werning, W. Hinzen, & E. Machery (Eds.), *The Oxford Handbook of Compositionality* (p. 64-80). Oxford University Press.

Szabó, Z. (2013). Compositionality. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy.* CSLI Publications.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Harvard University Press.

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141-188.

Wittgenstein, L. (1953). *Philosophical investigations* (P. Hacker & J. Shulte, Eds.). Wiley-Blackwell.