A case for deep learning in semantics*

Christopher Potts

Stanford Linguistics

1 Introduction

Pater's target article builds a persuasive case for establishing stronger ties between theoretical linguistics and connectionism – or, as it's recently been rebranded, DEEP LEARNING (DL; LeCun et al. 2015). In this commentary, I seek to further support his arguments by extending them to semantics. It's an exciting time for such discussions: DL is ascendant, and some of its breakthroughs in natural language processing (NLP) have come from melding assumptions and techniques from semantics with those of machine learning. Unfortunately, linguistic semantics has, to date, been much less influenced by DL research. I am concerned by this. There is now a large, vibrant, well-funded community of DL researchers working on compositional semantics, and semantic theory will suffer if semanticists don't engage with, and help shape, their research agenda.

When considering a role for DL in semantics, it is worth revisiting what Partee (1995) calls LEWIS'S ADVICE: 'In order to say what a meaning IS, we may first ask what a meaning DOES, and then find something that does that' (Lewis 1970:22). Lewis and his contemporaries proposed that INTENSIONAL FUNCTIONS do what meanings do, and higher-order logic in turn became the field's most important toolkit. Since then, semanticists have mostly not revisited this decision, though it has profoundly influenced how we delimit the field of semantics, which problems receive attention, and what we regard as an explanation.

A DL-based semantic theory would also try to follow Lewis's Advice, but it would replace intensional functions with n-dimensional arrays of numbers, and machine learning would likely replace logic as the most-used toolkit for the field. On this basis, one can build theories with many of the same properties as those of Lewis 1970. However, as with intensional functions, this foundational choice has far-reaching effects on the research agenda. After sketching what a DL-based semantics would be like, this commentary tries to identify these effects, both good and bad, with an eye towards a synthesis of intensional and DL semantics.[1]

*My thanks to Sam Bowman, Ignacio Cases, Chris Manning, Louise McNally, Joe Pater, and Nicholas Tomlin for incisive comments. All views are my own.

[1]For reasons of space and alignment with Pater's article, I focus on DL-based computational models. In NLP, SEMANTIC PARSING MODELS learn to predict logical forms or denotations using weighted symbolic semantic grammars. Compared to DL, this is a less radical way to combine semantic theory with machine learning. For discussion, see Liang & Potts 2015.

2 Semantics with functions or arrays

Current semantic theories adhere to the PRINCIPLE OF COMPOSITIONALITY, which says that the meaning of a non-terminal syntactic node is determined by the meanings of its child nodes and the semantic rule used to combine them. If a child node is a lexical item (terminal node), then its meaning is simply retrieved from the semantic lexicon.

This formulation says very little about the nature of the meanings or the combination operations. As I noted above, though, it is standard to model the meanings as intensional functions (elements in an intensional model) and, in modern type-driven semantic theories (Klein & Sag 1985), the mode of combination is almost always FUNCTION APPLICATION. Thus, let $[\![\cdot]\!]$ be the interpretation function, mapping syntactic nodes to intensional functions, and define $\mathbf{fa}(a, b) = a(b)$ or $b(a)$, whichever is well-formed. Then the tree in (1a) would be interpreted as in (1b).

(1) a.

A
- B
- C
  - D
  - E

b.

$[\![\mathrm{A}]\!] = \mathbf{fa}([\![\mathrm{B}]\!], [\![\mathrm{C}]\!])$
- $[\![\mathrm{B}]\!]$
- $[\![\mathrm{C}]\!] = \mathbf{fa}([\![\mathrm{D}]\!], [\![\mathrm{E}]\!])$
  - $[\![\mathrm{D}]\!]$
  - $[\![\mathrm{E}]\!]$

A DL-based compositional semantic theory could look very similar to this. The building blocks for all DL models are n-dimensional arrays of numbers: vectors (one-dimensional), matrices (two-dimensional), and higher-order tensors. Let $[\![\cdot]\!]^{\mathrm{DL}}$ be our DL-interpretation function, mapping syntactic nodes to n-dimensional arrays. A baseline model in this vein would say that lexical items denote vectors of dimension m and, in place of function application, we would use the following combination function:

(2) $$f([\![\mathrm{X}]\!]^{\mathrm{DL}}, [\![\mathrm{Y}]\!]^{\mathrm{DL}}) = \mathbf{g}\left(\left[[\![\mathrm{X}]\!]^{\mathrm{DL}}; [\![\mathrm{Y}]\!]^{\mathrm{DL}}\right] W\right)$$

Here, W is a matrix of dimension $2m \times m$, and $[a; b]$ is the concatenation of vectors a and b, which yields a new vector of dimension $2m$. The heart of the combination rule is the multiplication of $[a; b]$ with W, which yields a new vector x of dimension m, defined so that $x_j = \sum_{i=1}^{2m} [a; b]_i W_{ij}$ Each x_j has a transformation $\mathbf{g}$ applied to it. For instance, with $\mathbf{g} = \tanh$, each of the values is compressed into the range $(-1, 1)$. This element-wise non-linear transformation gives the network the power to model very complex functions (Cybenko 1989; Hornik 1992). Since the result of all this is another m-dimensional vector, the theory has the desired recursive property:

(3)

$[\![\mathrm{A}]\!]^{\mathrm{DL}} = \mathbf{g}\left(\left[[\![\mathrm{B}]\!]^{\mathrm{DL}}; [\![\mathrm{C}]\!]^{\mathrm{DL}}\right] W\right)$
- $[\![\mathrm{B}]\!]^{\mathrm{DL}}$
- $[\![\mathrm{C}]\!]^{\mathrm{DL}} = \mathbf{g}\left(\left[[\![\mathrm{D}]\!]^{\mathrm{DL}}; [\![\mathrm{E}]\!]^{\mathrm{DL}}\right] W\right)$
  - $[\![\mathrm{D}]\!]^{\mathrm{DL}}$
  - $[\![\mathrm{E}]\!]^{\mathrm{DL}}$

To illustrate, suppose we're working in a simplified two-dimensional semantic space in which the first dimension encodes sentiment valence and the second encodes intensity. With the lexical items in (4a–b) and the weight matrix W as in (4c), we achieve the effect that *not terrible* is less negative and less intense than *terrible* alone, as in (4d).

(4) a. $[\![\text{terrible}]\!]^{\text{DL}} = \begin{bmatrix}-1.00 & 1.00\end{bmatrix}$

b. $[\![\text{not}]\!]^{\text{DL}} = \begin{bmatrix}-0.10 & 0.10\end{bmatrix}$

c. $W = \begin{bmatrix} 0.06 & 0.32 \\ -0.14 & -0.53 \\ 1.24 & 0.00 \\ 0.02 & 1.06 \end{bmatrix}$

d. $[\![\text{not terrible}]\!]^{\text{DL}} = \tanh\left(\left[[\![\text{not}]\!]^{\text{DL}}; [\![\text{terrible}]\!]^{\text{DL}}\right] W\right) = \begin{bmatrix}-0.85 & 0.75\end{bmatrix}$

For toy examples like (4), we can invent values that achieve the desired outcome. However, such hand-construction is infeasible for large, complex semantic systems. In practice, the lexical representations in (4a–b) and the weight matrix W in (4c) are model parameters that are learned from data. For the most part, this is a SUPERVISED LEARNING process (Hastie et al. 2009): individual examples are labeled, and the optimization process finds parameter values that accurately predict those labels. To scale up (4), for example, we might collect real-valued polarity and intensity judgments for a wide range of adverb–adjective pairs, learn lexical entries and weights to predict those labels, and then evaluate the model by how accurately it can predict the labels for combinations it didn't see during optimization. In DL, these labels can take many forms and come from many different kinds of information; the ambition of the field is that they will ultimately be dense, high-dimensional representations of complex environments (which could be physical spaces, simulated worlds, large multi-media databases, and so forth). The recent surge in interest in DL derives in large part from advances in machine learning and computing power that together make it possible to optimize even very elaborate models effectively.

Socher et al. (2011) pioneered the use of tree-structured neural models for semantic tasks (see also Socher et al. 2012, 2013), building on early work on recursive connectionist architectures (Pollack 1990; Smolensky 1990; Plate 1994; Goller & Küchler 1996) and on compositional distributional semantics (Mitchell & Lapata 2010; Baroni & Zamparelli 2010; for an overview, see Baroni et al. 2014). These proposals are explicitly guided by the principle of compositionality and the usual assumptions and practices of formal semantics. Recurrent neural networks (RNNs; Elman 1990) are closely related variants in which the tree structure is a sequence (a strictly right-branching tree; see Pater 2018:§4.2). RNNs make fewer assumptions about what the incoming data are like than tree-structured architectures, but they still have the capacity to model aspects of semantic composition because each nonterminal node is represented in part by a lexical item and in part by a representation of the preceding sequence. RNNs and their variants are currently the most widely used DL architectures for language tasks.[2]

Those are the basics. I henceforth use the labels INTENSIONAL SEMANTICS and DL SEMANTICS to refer to the two general theoretical frameworks exemplified in the above sketches.

[2] I see two angles on the preference for RNNs over more richly structured models. It could derive from what Pater calls the 'emergentist tradition' of the field, which favors powerful models that make few assumptions about the data prior to learning. However, I suspect DL researchers would eagerly adopt tree-structured models if they showed consistent benefits, but so far they have not. It would be hasty to conclude that this tells us language isn't tree-structured, though. It's safer to conclude that the tree structures we're assuming are simply incorrect enough that they get in the way. Data-driven techniques like those of DL could help us discover the right trees.

These labels are unfairly reductionist, but I like how the first emphasizes how meanings are reconstructed and the second emphasizes the role of machine learning. The frameworks differ along other dimensions (e.g., logical vs. computational, symbolic vs. numerical), but I think these differences are largely incidental by comparison.

3 Learning and usage

Learning is the crux of Pater's (2018) arguments in favor of combining DL and generative approaches to language. Those issues might loom even larger in the context of semantics, which has long wrestled with the tension that Partee (1980:1) identifies in the following passage (Montague 1974 is a foundational collection, the basis for what came to be known as 'Montague Grammar'):

> The view that semantics is a branch of psychology is a part of the Chomskyan view that linguistics as a whole is a branch of psychology. [...] The contrasting view is ascribed to (and endorsed) by Thomason in his introduction to Montague 1974: 'Many linguists may not realize at first glance how fundamentally Montague's approach differs from current linguistic conceptions. According to Montague the syntax, semantics, and pragmatics of natural languages are branches of mathematics, not of psychology.'

The fact that DL models learn lexical entries and combination rules from data has profound effects on the resulting theories of meaning. With only a very few exceptions (like learning from dictionaries), the data set used to optimize the system will be, in one way or another, a record of UTTERANCES rather than idealized linguistic objects. As a result, it will reflect many aspects of language use: biases in word frequency, preferences for certain readings, pragmatic refinements of lexical items, and so forth. All of these factors will likely make their way into the final theory, in the sense that the learned model will reproduce these usage patterns. In other words, it will not only REPRESENT meanings, it will also make predictions about the interpretive choices listeners are likely to make when the language is used for some purpose (whatever purpose guided the data collection and labeling).

It seems clear that these consequences of learning compromise another methodological edict of Lewis (1970:19):

> I distinguish two topics: first, the description of possible languages or grammars as abstract semantic systems whereby symbols are associated with aspects of the world; and second, the description of the psychological and sociological facts whereby a particular one of these abstract semantic systems is the one used by a person or population. Only confusion comes of mixing these two topics.

Now, this edict COULD lead researchers to seek out innovative ways in which to abstract out an idealized semantics from the usage patterns encoded in a DL model's learned parameters. This would provide rich new perspectives on lexicography, on how children acquire semantic content from experience (Frank et al. 2009), and on why some aspects of semantic content are themselves highly variable and uncertain (Clark 1997; Wilson & Carston 2007; Potts & Levy 2015).

However, the dominant reaction from the NLP community has been to accept that their systems blur meaning and use together. If one wants one's system to do interesting things in the real world, it is generally desirable to have it, for example, resolve ambiguities rather than simply representing them. A DL semantics will likely only further encourage the blurring of these boundaries, since DL makes it easy to combine diverse representations – of language, of the physical environment, of others' mental states – and learn from how these representations interact. In this setting, it can be hard to discern a motivation for the restrictive view that Lewis advocates above. A DL semantics will naturally encourage holistic exploration of the full significance of utterances, with narrow semantic characterizations employed only where they prove useful for meeting these broader goals. Certain insights and goals might be subverted in the process.

To take one simple example, Socher et al. (2013) motivate a tree-structured recursive neural tensor network, which is essentially an elaboration of the model in (2) with a greater capacity to capture the relationships between the two child vectors. Their experiments are conducted on a corpus of sentences from movie reviews labeled at the phrase- and sentence-level for their evaluative sentiment. One of the case studies they report concerns coordination with *but*. They make the case that the learned model reflects the generalization, due to Lakoff (1971), that *A but B* concedes that A and argues that B, because the model consistently predicts that the sentiment of *A but B* is largely determined by the sentiment of *B*. This strikes me as an innovative way to use a learned DL model to support a nuanced generalization about meaning. However, there is no clear sense in which the model also captures what semanticists might regard as the essence of this lexical item: it is a coordinator that conveys a secondary meaning (perhaps a CONVENTIONAL IMPLICATURE) that has its own compositional properties.

All of this makes salient the role that intensional functions and their attendant logical apparatus play in encouraging the sort of division that Lewis advocates for above. Logics are the paradigm cases of closed, self-contained formal systems defined independently of particular users or instances of use. When we embed our semantic theories in these systems, the theories inherit these properties. This has undoubtedly had an effect on the sort of phenomena that linguistic semanticists choose to study. It separates semantics from all aspects of learning and cognitive representation (Partee 1980, 1981; Jackendoff 1996), and it naturally discourages work on items that are explicitly tied to interactional language use – disfluencies, swears, honorifics, interjections, and other items that Kaplan (1999) characterizes in terms of their USE CONDITIONS. Where such items are studied, it tends to be from the perspective of how they are represented model-theoretically and how they interact with the rest of the compositional system, rather than from the perspective of what they actually mean (in the pre-theoretic sense) when speakers actually use them. Just as it might seem idiosyncratic that Socher et al. (2013) model only the argumentative structure of *but*, these accounts can also seem idiosyncratic in how they attend only to what intensional semantics can easily accommodate.

There is a related, subtler difference worth bringing out. Since any learning-based theory will depend on data sets of utterances, the question arises whether these utterances have a unique intended semantic interpretation. Where there are ambiguities, are they always resolved, or might both speakers and listeners entertain multiple possibilities (Clark 1997)? This issue simply doesn't arise for classical semantic theories, which confine themselves to the question of which representations are possible. A learned theory is likely to rank possible construals, inviting the question of whether the full ranking has communicative or cognitie significance. In other words, the tools of DL are leading us to lose track of the distinction between sentence and utterance, just as intensional theories tend to force it upon us.

4 Compositionality and generalization

Both of the theories sketched in section 2 are compositional in the technical sense originally defined by Montague (1970). However, the weight matrix W in the DL semantics might be seen as compromising the spirit of compositionality because of its global character. Not only is W used in every phrasal combination, but its values are learned from the entire data set. As a result, W can import language-wide information into the local computation of phrasal meaning. Relatedly, it can have the effect of spreading information out across different components of the system. The example in (4) begins to suggest how this can happen: while $[\![\text{not}]\!]^{\text{DL}}$ contributes to the final values, one can make a case that it is W, rather than $[\![\text{not}]\!]^{\text{DL}}$, that encodes the core effect of negation.

It would be easy to dismiss these considerations on the grounds that the compositionality principle is not intended as a statement about how any agent, human or artificial, would learn the semantics of a system. We expect such learning processes to be more holistic. The compositionality principle is meant instead to constrain the final state of the semantic grammar, and here our DL version passes muster.

Still, these considerations should lead us to reflect on the broader rationale for the compositionality principle. The usual story is that compositionality is crucial to our ability to produce and understand creative new combinations of linguistic units, because it offers guarantees about the systematicity and predictability of new units. However, these observations alone do not imply compositionality. The interpretation of a given phrase could be systematic, predictable, and also determined in part by global properties of the utterance, the speaker, the discourse situation, and so forth. And, indeed, it seems to me that our everyday experiences with language are in keeping with this. Listeners greedily use all sorts of information when making sense of others' utterances, and speakers assume they will do this.

In machine learning, by contrast, the goal is not compositionality per se, but rather GENERALIZATION: a system succeeds to the degree that it makes good predictions for entirely new data – to the extent that it displays a human-like ability to creatively produce and consume utterances that are novel in the sense that they are not included in the training data. Compositionality is a highly restrictive strategy for achieving this. If one is designing a system with generalization in mind, one is unlikely to restrict access to potentially useful information a priori. Doing so could weaken the system and, in any case, modern machine learning models learn which pieces of information to pay attention to as part of optimization, so it rarely makes sense to deny them available information during learning.

Finally, there is another dimension to the contrast between compositionality and generalization. As Janssen (1997:461) observes, 'Compositionality is not a formal restriction on what can be achieved, but a methodology on how to proceed' (see also Partee 1984:§7.5). As a result, though it has had a meaningful impact on the accounts semanticists develop, it can only do so much, since we can make pretty much any analysis compositional if we feel pressed to. In contrast, generalization is something we can measure quite precisely using quantitative metrics and specially created data sets used only for assessment. In turn, NLP (like the rest of artificial intelligence) is driven almost entirely by quantitative performance. The effects are both good and bad. On the one hand, even really unusual theories can get a hearing if they post good numbers, whereas unusual linguistic theories can have trouble getting a fair hearing. On the other hand, scientific goals can seem less important than incremental gains on accepted test sets. A truly interdisciplinary DL semantics would, I think, have a chance of finding a balance.

5 Lexical semantics

The area of semantics in which DL would have the largest impact is arguably lexical semantics, and I think it's here that we see most clearly what the trade-offs would be as compared to intensional approaches.

In the DL theory sketched in section 2, lexical items (and, indeed, all meanings) are vectors. The lexicon is in turn a matrix in which the rows are lexical items and the columns capture specific aspects of meaning. One could, in these terms, recapitulate all of intensional semantics: the columns could represent possible worlds, with binary vectors encoding truth in those worlds. This would give us representations for proposition-denoting lexical items. Higher-order tensors could capture the dependencies of other arguments, thereby recreating the typed semantic hierarchy of intransitive verbs, transitive verbs, prepositions, and so forth. However, this is probably not the best use of DL's building blocks. Instead, these representations are more fruitfully used to capture specific dimensions of semantic meaning, in exactly the same way that, for example, phonological segments are modeled as binary vectors in which each dimension corresponds to a feature and 1 means the feature is present/true and 0 means it is absent/false. Such feature representations are common in many areas of linguistics, and were a mainstay of generative semantics, so these ideas are perhaps not so unfamiliar.

What IS unfamiliar about DL modeling of the lexicon is that one rarely has a solid intuitive grasp on what the dimensions in the lexical entries mean. They are typically learned from data via a complicated model, and they are too big and complex to understand analytically. The meaning they encode is largely latent in the relational structure of the full lexical space (Turney & Pantel 2010; Lenci 2018); a further modeling process might reveal that the representations can be used to make accurate predictions about, say, lexical entailment, synonymy, and antonymy, but this is rarely evident from just looking at them. This lack of interpretability is not an inevitable outcome of DL modeling – Pater reviews many cases in which linguists' usual representations are used as the input to DL models to good effect – but it is a likely consequence of truly embracing what DL has to offer the lexical semanticist.

Is this a viable alternative to current theories of the lexicon in semantics? To address this question, I feel we have to confront the fact that the lexicon has largely been neglected by modern semantic theories that find their origins in the work of Lewis (1970) and Montague (1974). The foreword to Carlson (1977) is a send-up of this limitation; then a graduate student, Carlson asked his professors the meaning of *life* and was told that it was, in essence, an atomic and unanalyzed formal symbol **life**, and 'the class then turned to the much stickier problem of pronouns'. In introducing the papers in Montague (1974), Thomason (1974:48–49) is more direct:

> The problems of a semantic theory should be distinguished from those of lexicography [...] A central goal of (semantics) is to explain how different kinds of meanings attach to different syntactic categories; another is to explain how the meanings of phrases depend on those of their components. [...] But we should not expect a semantic theory to furnish an account of how any two expressions belonging to the same syntactic category differ in meaning. 'Walk' and 'run,' for instance, and 'unicorn' and 'zebra' certainly do differ in meaning, and we require a dictionary of English to tell us how. But the making of a dictionary demands considerable knowledge of the world.

I would argue that this doesn't reflect a true division between semantics and lexicography, or an intrinsic limitation of semantics. Rather, it is another case of the tools shaping the theory. The tools developed in Montague 1974 are ideally suited to modeling the functional elements of the vocabulary, but they are cumbersome when applied to open-class lexical items. Of course, the meanings can in principle be REPRESENTED by these logical theories, but when it comes to actually building a lexicon, they are mostly no help at all, and the successful large-scale lexical projects do not use them, opting instead for capturing lexical knowledge in large graphs (Fellbaum 1998; Ruppenhofer et al. 2006).

So the promise of DL semantics is that it will allow us to learn rich representations of the entire lexicon, including open-domain items like *walk* and *zebra*. We can learn such representations from co-occurrence patterns, from visual images, from grounded interactional scenarios, and so forth. The important caveat here is that the representations are unlikely to admit of analytic understanding. We might be able to probe them in various ways and, in doing so, assess whether they have captured a specific set of properties or constraints, but this is unlikely to be evident from high-level inspection or straightforward, exact calculations. It should be said that, given the complexity of natural language lexicons, these analytic limitations might be unavoidable if the goal is a truly comprehensive treatment.

It is worth noting also that, echoing themes of section 3, DL lexical theories discourage firm boundaries between semantics and pragmatics. The information used to create the lexical representations is mostly a record of language use, and the results tend to reflect this. DL theories generally cannot distinguish between literal and non-literal language use, or between denotation and connotation, or between semantic content and pragmatic enrichment. As a result, since these phenomena are pervasive, they take centerstage in a way that they rarely do in intensional theories.

On the other hand, DL theories have been much less successful to date in modeling the functional vocabulary that semanticists have mostly specialized in. This is not a representational challenge, as Clark et al. (2011) show with their compositional distributed models of meaning, but rather one arising from the demands of machine learning. For instance, the DL theory of section 2 is essentially MONOTYPED: every meaning, whether lexical or phrasal, is a vector. Semanticists will immediately see that this is untenable; quantificational determiner meanings are more complex than common noun meanings, so any theory that puts them in the same meaning space is unlikely to do justice to determiners. This can of course have deep consequences for the success of a model, in ways that only careful linguistic argumentation can reliably bring out. For instance, Bowman (2016) shows that even highly effective DL models trained to do the task of NATURAL LANGUAGE INFERENCE (commonsense reasoning about entailment and contradiction) can fail to fully capture the monotonicity properties of quantifiers, which has immediately evident consequences for them as systems that are supposed to reason in language.

6 Looking ahead

In a commentary called 'Computational linguistics and deep learning', Manning (2015:701) quotes the machine learning researcher Neil Lawrence as saying, 'NLP is kind of like a rabbit in the headlights of the Deep Learning machine, waiting to be flattened'. At the time, there was a great deal of optimism about DL throughout the field of artificial intelligence, and prominent DL researchers often described language as the next big area for applying DL tools. In this context, Lawrence's warning invokes an unflattering but common trope: scientists failing to adapt to revolutionary ideas because of an irrational commitment to old questions and techniques. Linguists are especially familiar with this narrative; the story of the Chomskyan revolution, as told in work like Searle (1972) and Harris (1993), is one of intransigent American structural linguists who were flattened by the Cognitive Science machine.

Lawrence's remark is directed at NLP, not theoretical linguistics, but we should still ask whether we are at risk of being flattened by the DL machine as well. As recently as fifteen years ago, NLP semantics was mainly lexical semantics and heuristic shallow semantic representation. Semanticists could rest easy that they were the only ones paying attention to compositional aspects of meaning. This began to change in about 2005, with an outpouring of research on learned semantic parsing systems (see footnote 1). The rise of DL over the past decade has greatly accelerated this change, so that serious semantic interpretation is now the norm in NLP, in the sense that almost all researchers use recursive models that represent meaning compositionally. In my experience, the NLP community is also admirably self-critical, apt to dwell on where its models are failing as a way of moving forward. The shortcomings of DL for linguistic analysis are constantly discussed and debated, and they might eventually lead to DL being supplanted in some sense. Deep semantic analysis isn't going anywhere, though – it's too obviously essential to achieving systems that can produce and interpret language robustly.

What role will formal semanticists play in this new era of deep, learned semantics? It is not a foregone conclusion that the results, values, and methods of formal semantics will survive. Semanticists will have to insert themselves into the DL discourse to make that happen. However, I'd like to avoid this negative framing as much as possible. As our best selves, we can set aside our past theoretical commitments and current methodological attachments and just pose the question of whether DL provides tools, ideas, and insights that could benefit our field. Pater's target article gives us many reasons to say yes, and I think the above discussion does as well. DL appears outwardly to be driven entirely by engineering concerns, but its connectionist roots are still strong and, as Pater documents, connectionism was founded on many of the same core principles as modern linguistics. Both fields can achieve more if they work together.

References

Baroni, Marco; Raffaella Bernardi; and Roberto Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology* 9.241–346.

Baroni, Marco, and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective–noun constructions in semantic space. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1183–1193. Cambridge, MA: Association for Computational Linguistics. Online: http://aclweb.org/anthology/D10-1115.

Bowman, Samuel R. 2016. *Modeling natural language semantics in learned representations*. Stanford, CA: Stanford University dissertation.

Carlson, Gregory. 1977. *Reference to kinds in English*. Amherst, MA: UMass Amherst dissertation.

Clark, Herbert H. 1997. Dogmas of understanding. *Discourse Processes* 23.567–59.

Clark, Stephen; Bob Coecke; and Mehrnoosh Sadrzadeh. 2011. Mathematical foundations for a compositional distributed model of meaning. *Linguistic Analysis* 36.345–384.

Cybenko, George. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* 2.303–314.

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14.179–211.

Fellbaum, Christiane (ed.) 1998. *WordNet: An electronic database*. Cambridge, MA: MIT Press.

Frank, Michael C.; Noah D. Goodman; and Joshua B. Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* 20.578–585.

Goller, Christoph, and Andreas Küchler. 1996. Learning task-dependent distributed representations by backpropagation through structure. *IEEE International Conference on Neural Networks*, 347–352. IEEE.

Harris, Randy Allen. 1993. *The linguistic wars*. Oxford: Oxford University Press.

Hastie, Trevor; Robert Tibshirani; and Jerome Friedman. 2009. *The elements of statistical learning*. 2 edn. Berlin: Springer.

Hornik, Kurt. 1992. Approximation capabilities of multilayer feedforward networks. *Neural Networks* 4.251–257.

Jackendoff, Ray S. 1996. Semantics and cognition. *The handbook of contemporary semantic theory*, ed. by Shalom Lappin, 539–559. Oxford: Blackwell Publishers.

Janssen, Theo M. V. 1997. Compositionality. *Handbook of logic and language*, ed. by Johan van Benthem and Alice ter Meulen, 417–473. Cambridge, MA and Amsterdam: MIT Press and North-Holland.

Kaplan, David. 1999. What is meaning? Explorations in the theory of *Meaning as Use*. Brief version — draft 1. Ms., UCLA.

Klein, Ewan, and Ivan A. Sag. 1985. Type-driven translation. *Linguistics and Philosophy* 8.163–201.

Lakoff, Robin. 1971. If's, and's, and but's about conjunction. *Studies in linguistic semantics*, ed. by Charles J. Fillmore and D. Terence Langendoen, 114–149. New York: Holt, Rinehart, and Winston.

LeCun, Yann; Yoshua Bengio; and Geoffrey E. Hinton. 2015. Deep learning. *Nature* 521.436–444.

Lenci, Allesandro. 2018. Distributional models of word meaning. *Annual Review of Linguistics* 4.151–171.

Lewis, David. 1970. General semantics. *Synthese* 22.18–67.

Liang, Percy, and Christopher Potts. 2015. Bringing machine learning and compositional semantics together. *Annual Review of Linguistics* 1.355–376.

Manning, Christopher D. 2015. Computational linguistics and deep learning. *Computational Linguistics* 41.701–707.

Mitchell, Jeff, and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science* 34.1388–1429.

Montague, Richard. 1970. Universal grammar. *Theoria* 36.373–398, Reprinted in Montague (1974), 222–246.

Montague, Richard. 1974. *Formal philosophy: Selected papers of Richard Montague*. New Haven, CT: Yale University Press.

Partee, Barbara H. 1980. Semantics – mathematics or psychology? *Semantics from different points of view*, ed. by Egli Bäuerle and Arnim von Stechow, 1–14. Berlin: Springer-Verlag.

Partee, Barbara H. 1981. Montague grammar, mental representations, and reality. *Philosophy and grammar*, ed. by Stig Kanger and Sven Öhman, 59–78. Dordrecht: D. Reidel.

Partee, Barbara H. 1984. Compositionality. *Varieties of formal semantics*, ed. by Fred Landman and Frank Veltman, 281–311. Dordrecht: Foris.

Partee, Barbara H. 1995. Lexical semantics and compositionality. *Invitation to cognitive science*, ed. by Lila R. Gleitman and Mark Liberman, vol. 1, 311–360. Cambridge, MA: MIT Press.

Pater, Joe. 2018. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. To appear in *Language*.

Plate, Tony A. 1994. *Distributed representations and nested compositional structure*. Toronto: University of Toronto dissertation.

Pollack, Jordan B. 1990. Recursive distributed representations. *Artificial Intelligence*

46.77–105.

Potts, Christopher, and Roger Levy. 2015. Negotiating lexical uncertainty and speaker expertise with disjunction. *Proceedings of the 41st annual meeting of the Berkeley Linguistics Society*, ed. by Anna E. Jurgensen, Hannah Sande, Spencer Lamoureux, Kenny Baclawski, and Alison Zerbe, 417–445. Berkeley, CA: Berkeley Linguistics Society.

Ruppenhofer, Josef; Michael Ellsworth; Miriam R. L. Petruck; Christopher R. Johnson; and Jan Scheffczyk. 2006. *FrameNet II: Extended theory and practice*. Berkeley, CA: International Computer Science Institute.

Searle, John R. 1972. Chomsky's revolution in linguistics. *The New York Review of Books* 18.12–29.

Smolensky, Paul. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46.159–216.

Socher, Richard; Brody Huval; Christopher D. Manning; and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, 1201–1211. Online: http://www.aclweb.org/anthology/D12-1110.

Socher, Richard; Jeffrey Pennington; Eric H. Huang; Andrew Y. Ng; and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. *Proceedings of the 2011 conference on Empirical Methods in Natural Language Processing*, 151–161. Edinburgh, Scotland, UK: ACL. Online: http://www.aclweb.org/anthology/D11-1014.

Socher, Richard; Alex Perelygin; Jean Wu; Jason Chuang; Christopher D. Manning; Andrew Y. Ng; and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 conference on Empirical Methods in Natural Language Processing*, 1631–1642. Seattle, WA: Association for Computational Linguistics. Online: http://www.aclweb.org/anthology/D13-1170.

Thomason, Richmond H. 1974. Introduction. In Montague, 1–69.

Turney, Peter D., and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37.141–188.

Wilson, Dierdre, and Robyn Carston. 2007. A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. *Pragmatics*, ed. by Noel Burton-Roberts, 230–259. Basingstoke and New York: Palgrave Macmillan.