Minds and Machines
https://doi.org/10.1007/s11023-021-09571-w

ORIGINAL ARTICLE

(What) Can Deep Learning Contribute to Theoretical Linguistics?

Gabe Dupre[1]

Received: 17 February 2021 / Accepted: 24 August 2021

Abstract
Deep learning (DL) techniques have revolutionised artificial systems' performance on myriad tasks, from playing Go to medical diagnosis. Recent developments have extended such successes to natural language processing, an area once deemed beyond such systems' reach. Despite their different goals (technological development vs. theoretical insight), these successes have suggested that such systems may be pertinent to theoretical linguistics. The competence/performance distinction presents a fundamental barrier to such inferences. While DL systems are trained on linguistic performance, linguistic theories are aimed at competence. Such a barrier has traditionally been sidestepped by assuming a fairly close correspondence: performance as competence plus noise. I argue this assumption is unmotivated. Competence and performance can differ arbitrarily. Thus, we should not expect DL models to illuminate linguistic theory.

Keywords Philosophy of linguistics · Philosophy of artificial intelligence · Machine learning · Competence and performance

1 Introduction

Over the last 15 years or so, deep learning approaches have become dominant in computer science. For a wide variety of tasks, if we want a computer to do something for us—identify objects in photos, diagnose diseases, drive cars, etc.—we are better off providing the computer with a large set of exemplary data, and allowing it to learn how to complete the task for itself, rather than programming a solution ourselves. This is as true in the case of linguistic tasks, such as translation, inscribing speech, filtering 'spam', and others, as it is in any other domain, if not more so. Reasonably enough, these successes have led various theorists to ask whether these applied, computational models might provide insight to theoretical linguistics, the

✉ Gabe Dupre
G.g.dupre@keele.ac.uk

[1] Keele University, Newcastle, UK

Published online: 10 September 2021

discipline aimed at describing and explaining the properties of human natural language. In this paper, I shall argue that, for the core tasks of theoretical linguistics, there are principled reasons why this is unlikely to be the case. Specifically, deep learning systems are unlikely to prove illuminating to theoretical linguistics, on the grounds that linguistic theories are guided not solely by normal linguistic behaviour, but by the constraints of adjacent disciplines, such as neuroscience, evolutionary theory, developmental psychology, etc. and thus the rules and structures they posit are, from the perspective of observable speech, *sui generis* and quirky, and thus that deep learning systems trained on such observable speech behaviour are unlikely to capture them.

2 Different Targets, Shared Paths?

I shall use the term 'Natural Language Processing' or 'NLP' to refer to the applied scientific project of designing a computational system capable of performing linguistic tasks. Paradigmatic such tasks include translation from one language to another, transcription of speech, conversation (by so-called 'chatbots'), classifying written texts as hate speech or spam, etc. Clearly, automated systems with such capacities can be both useful and valuable, which motivates massive investment in the development and improvement of such systems. Nowadays, the most common, and most successful, approach to the creation of such systems is Deep Learning ('DL'), wherein the computational system 'learns' for itself how to perform on the basis of, typically very, large sets of data.[1] Assessment of such tasks is pragmatic, typically involving broadly behavioural measures, such as accuracy of translation (as compared to that of a skilled human translator). The central question for NLP is: do these systems do what we want them to?

I shall use the term 'Theoretical Linguistics' or 'TL' to refer to the basic scientific project of describing and explaining the properties of human language. Theoretical linguistics is itself further broken up into various sub-disciplines, focusing on different aspects of human language: centrally, phonology, morphology, syntax, and semantics. For the purposes of this paper, I shall assume that TL is a branch of cognitive psychology, and thus that a true theory of human language will thereby provide an account of the distinctive features of human psychology that enable us to learn and use language.[2] For example, a true phonological theory will provide an

[1] Of course, the degree to which these systems are constrained by the goals, structures, etc. provided by the programmer varies, and should not be undersold.

[2] While this 'mentalist' position is very widespread within linguistics, there are alternatives. Famously, Katz (1984, 1980) argued that linguistics should be construed as a branch of mathematics ('Platonism'), and Devitt (2006) has argued that linguistics should be understood as concerning properties of concrete public symbol types ('nominalism'). Within linguistics, Gazdar et al. (1985) are explicit that they do not view the success of their theory as dependent on whether it accurately describes human psychology (p. 5), although whether they endorse a Platonist or nominalist approach is left unclear. I will simply be assuming mentalism for the purposes of this paper, and thus my arguments will be relevant only to the question of what we can learn from DL about natural language *understood as a psychological capacity*.

abstract description of the ways that speakers represent the sounds of their language, identifying the ways they group and differentiate speech sounds (as similar/different, legitimate/illegitimate, etc.). Such descriptions will be needed both in the cross-linguistic case in the description of putative linguistic 'universals' (e.g. that all languages that allow sonority decreases, such as '*lb*' in their syllabic onsets, also allow sonority increases, such as '*bl*', but not vice versa), and in the description of specific languages (e.g. that English prohibits sonority decreases in onsets). Likewise for other branches of linguistics. Importantly, while the evidence for or against such claims will typically come from observations about language use, paradigmatically including native speaker judgements of the (un-)acceptability of certain expressions, it may also come from myriad other domains, such as psychological or neurobiological theory, evolutionary and human history, cross-linguistic comparisons, learnability considerations, etc. Assessment of such work consists in the application of standard 'theoretical virtues': simplicity, empirical coverage, consilience, etc. The question for TL is: does a given theory accurately describe human psychology?

Despite my insistence that TL can be informed by such neighboring disciplines as neurobiology and developmental psychology, it is important to distinguish the goals of these different disciplines. Firstly, as I will use the term, TL is a *synchronic* discipline. That is, the goal of TL is to describe the acquired states of linguistic competence in normal humans. This differentiates TL from developmental linguistics, which aims to describe the psychological mechanisms involved in acquiring such a steady state, as well as the trajectories involved in such acquisition. Secondly, TL is a *computational level* discipline, in the sense of Marr (1982). That is, it aims to describe such steady states of linguistic competence at a relatively abstract level, in terms of the cognitive tasks such states are deployed to solve and the information appealed to in such cognitive activities. This differentiates TL from psycholinguistics and neurolinguistics, which are aimed at lower-level descriptions of the algorithmic processes used in executing such computations, and the neurobiological mechanisms which realize or instantiate such algorithmic processing.

These distinctions are important as they provide a restriction on the scope of my argument: investigation of DL models is unlikely to provide insight into TL, viewed as a synchronic and computational level discipline. My argument is thus consistent with recent work, e.g. Pater (2019) and Linzen (2019), that has argued that DL may provide insight into the mechanisms by which linguistic competence is acquired. Linzen argues that while we are unlikely to be able to read off human competence from language-trained machines (as "there are significant differences between the syntactic representations that RNNs acquire and those of humans." (p. 105)), once the structure of human competence is identified as a target, it is possible that DL systems can be used to model language acquisition on the basis of linguistic data. Likewise, it is possible that DL-style systems could provide insight into how linguistic competence is *realized*, without such systems being suitable tools for discovering the computational-level properties of human language.[3]

[3] As an epistemological pluralist, I am also open to the idea that DL models may *indirectly* constrain TL theorizing, by showing that some proposed linguistic competences are more plausibly learned or neurobiologically realized than others. The argument in the paper is aimed at a potential direct inference from

Given the quite different goals of DL and TL, there is no requirement that the results of either project will be relevant to the other. This mutual independence is a common-enough feature of applied and basic science. Aeronautical engineering need not inform, nor be informed by, the biological study of animal flight. However, in practice, especially in cognitive domains, applied and basic scientific disciplines often display high degrees of productive 'cross-fertilisation', with insights flowing in both directions.

In the case of flight, the reason for the broad independence of applied and basic research stems from the fact that there are multiple different ways of 'solving the problem' of flight. Rigid wings with a source of propulsion will do it, but so will flappable wings of the right sort. As one of these is biologically more plausible, while the other is easier to produce mechanically, the two traditions have taken off on distinct paths, leaving little room, or motivation, for collaboration. However, it has not typically seemed that this type of situation is present in the case of cognitive capacities. In many cases, both applied and basic researchers are in the position of asking 'how-possibly' questions: how can we identify a 3-D scene on the basis of a 2-D retinal projection? how could a mechanical system display *rationality* in its transitions between states? etc. In such a case, the applied and pure theorists are both attempting to envision systems capable of performing the task (albeit, often with different constraints), and so the potential for mutual illumination is high. And indeed, this is what we see when we look to the shared history of Artificial Intelligence and Cognitive Science. On the one hand, cognitive theories provide accounts of how various tasks are solved by humans, which themselves can serve as models when designing machines capable of solving the same problem. On the other, if a machine can be developed to solve a problem, this provides an existence-proof that the algorithms it implements have the capacities we are seeking to explain in humans, and thus suggests the hypothesis that this is indeed the way our own minds work.

The domain of language has exemplified this cross-fertilisation as much as any other. Early work in generative linguistics, such as Chomsky (1957/2002, 1965) (partially developed, relevantly to my current point, while Chomsky was employed at MIT to work on a machine translation project) was hugely influential in the early development of language processing systems, which until around the 1980s largely worked by incorporating the rule systems characterised by such theoretical work. In the other direction, the very existence of physical computers has provided the dominant framework for theorising about the mind since the latter half of the 20th Century: the Computational Theory of Mind.

Further, NLP may seem an even better source for theoretical insight than other branches of AI, given that NLP systems are designed to engage with human language users. This may provide a constraint on these systems, ensuring that their behaviour doesn't deviate too far from that of humans, which may be absent in other

Footnote 3 (continued)
the fact that some DL model has solved a linguistic task by forming certain sorts of structures or inferential patterns to the claim that human language works in a similar or analogous way.

domains, such as computer vision, or object classification, wherein the computational systems interact not with other agents, but with the environment.

The view that the most successful engineering projects in the linguistic domain could provide insight for our pure science theories of human language thus suggests itself very naturally, given the significant successes such a program has generated recently. To see how such inferences might go, in the next section I shall describe in some more detail how DL models of natural language work, and how we might leverage from them theoretical insight.

3 How Do Machines "Learn Languages"?

While there is lots of variation in DL models, depending on the specific task they are designed to solve, the resources available, and other factors, the basic functionality of such systems is relatively straightforward. I will focus on *supervised*, *end-to-end* neural network models.

These models consist of a collection of inter-connected nodes. Each node functions to transform an input signal, from each node causally upstream of it in the network, into an output signal, which is transferred to those nodes causally downstream of it. These nodes are organised into layers, with each node in a given layer largely causally responsive to nodes in the previous layer. The 'deep' in 'deep learning' reflects the fact that these neural networks have many such layers. A major axis of variation for such networks is their topology: traditional 'connectionist' networks were highly connected, with each node in a given layer connected to every node in the layers on either side, while modern deep networks are typically much more internally heterogeneous. Input nodes comprise the first layer of such a network, and receive their inputs exogenously. Such inputs must therefore be 'encoded' so as to be readable by the system. For example, a machine translation program will take an encoded sentence in a given language as input, which will send a wave of activation through the network. The properties of this wave will be determined by the individual functions of each node. The wave will culminate with the output of the final layer, which can then be decoded, for example as a sentence in the target language.

The most important feature of such networks is that the signal transformation functions of each node need not be, and usually are not, determined by the programmer. Instead, they are 'learned', i.e. dynamically modified in response to training. Supervised training consists in providing the network with a range of input-output pairs, a way of comparing its own outputs, given these inputs, with the 'correct' pairings provided, and a method for modifying the functions of internal nodes so as to better fit these correct pairings. For example, in a machine translation system, we could provide a wide range of paired English and French sentences (e.g. from the proceedings of Canadian Parliament). If we initially randomise the functions of the nodes in this system, its initial outputs will be very unlike the intended translations of the input sentences. But by using a learning algorithm, such as backpropagation, we can compare the system's 'guess' of the translation with its genuine translation, and modify the internal weightings in such a way that they are more likely to produce the target output. Repeated application of this process enables the system to

'hill-climb', strengthening the functions which are on the right track, and dampening those that lead to mistranslation. Given enough nodes, and enough training data, the system will converge on a function capable of generating these data, and if things go well, generalising beyond to novel cases. Much engineering work in machine learning is aimed at ensuring this latter.

Paradigmatic work in NLP uses end-to-end modeling, where the input and output are taken as given, and no constraints are placed on how the system calculates the function mapping the former onto the latter. For example, Wu et al. (2016) present the Google Neural Machine Translation (GNMT) system, which was able to achieve results comparable to those "achieved by average bilingual human translators on some [...] test sets" (p. 20). This system takes entire sentences in a given language as input, mapping these onto output sentences as a whole, rather than identifying and translating individual words and phrases and using these to sequentially construct a translated sentence.

Now, of course, producing a system capable of performing machine translation (or transcribing speech, answering questions, or any other NLP task) does not, on its own, provide theoretical insight into natural language. The input-output performance of such systems takes us no further than what human translators are capable of doing, without detailed knowledge of contemporary linguistics. However, it could be hoped that 'opening up' such systems, and looking at how they perform these tasks might indeed be illuminating in this way. This is the proposal that seems most *prima facie* plausible as a contribution from DL to theoretical linguistics, and is the one I shall be investigating.

Before elaborating on my argument that this proposal is mistaken, i.e. that DL systems are unlikely to complete linguistic tasks in the way that humans do, and thus that we are unlikely to learn about human language through investigation of such systems, it will be essential to get clear on just how DL systems could, in principle, provide theoretical illumination. Specifically, investigation of the classificatory schemes adopted by such systems could provide a source of evidence pertinent to selecting theories of how human beings classify linguistic expressions. I turn now to spelling out such a proposal.

Theoretical linguists often debate about whether two surface linguistic phenomena should be viewed as resulting from distinct underlying operations, or from the same operation, as applied in different ways, or to different expressions. For example, traditional work in generative linguistics distinguished between two linguistic operations responsible for sentences apparently consisting of a single nominal argument serving as subject of multiple verbs. 'Control' involves a sentential subject serving as semantic argument of both a main and an embedded verb, and is underlain by a structure in which the subject nominal is grammatically the subject of the main verb, with an unpronounced anaphoric expression (denoted 'PRO') serving as grammatical argument of the embedded verb, as in "Mikhail_1 wants PRO_1 to read regularly". Mikhail here is both the (hopeful) reader and the wanter, and this result is achieved by enforcing co-reference between the sentential subject and its anaphoric PRO. Raising, on the other hand, involves the sentential subject only genuinely being the semantic argument of the embedded verb, but appearing in the surface subject position of the main verb, as in "Mikhail_1 seems t_1 to read

regularly", wherein Mikhail is the reader, but not the 'seemer'. Raising structures result from moving the embedded subject to sentential subject position, leaving behind a trace (or, in most contemporary theories, an unpronounced but identical copy). While phonologically, we hear the nominal expression at the beginning of the sentence, semantically it is interpreted in its original, embedded location. Arguments that, despite the surface similarity between "Mikhail wants to read regularly" and "Mikhail seems to read regularly", these result from different underlying processes appeal to observations like: raising verbs can take semantically null expletive subjects, while control verbs can't ("It seems Mikhail reads regularly" vs. *"It wants Mikhail reads regularly"). However, in line with the explanatory scheme of contemporary Minimalist grammar, which problematises the introduction of formal tools like indices into grammatical structures, Hornstein (1999) argues that we should view both kinds of surface structure as stemming from the same sorts of process, specifically reducing control to raising.[4]

Cases like this, where linguists are unsure as to whether some surface phenomena result from distinct or similar underlying operations, provide possible targets for DL-based investigation. By investigating the internal workings of an NLP system, we could see whether these phenomena are treated in the same, or different, ways by such systems. And this could provide a *prima facie* argument that they are treated likewise by our linguistic faculties. Indeed, it could be argued, at a certain level of description what the linguist and the DL system are doing is the same: trying to abstract the underlying regularities in linguistic data. Given this, it might be surprising if the two traditions *didn't* converge on their linguistic analyses.

There are various methods that have been proposed for the task of looking inside the black box of DL systems. Early work by Sanger (1989) and Elman (1990, 1991) showed how to use statistical tools like cluster analysis and principle components analysis, applied to connectionist systems, to identify the classifications made by these systems. Very roughly, the idea is that we can get some understanding of the 'reasoning' behind the system's overall input-output behaviour by looking at correlations between responses to different stimuli and between antecedents of different outputs. If the same or similar patterns of activation at given nodes or locations are reliably activated in response to different inputs, this suggests that the system treats these inputs as identical or similar. Likewise, if different outputs are produced by similar hidden nodes, this provides insight into the similarities the system ascribes to these outputs. Further, correlations between nodes in adjacent layers can identify common internal trajectories, indicating the patterns of inference that the system makes. On the basis of these styles of analysis, Sanger and Elman identified analogues of phonological and grammatical classification occurring internal to connectionist systems.[5]

[4] See Boeckx et al. (2010) for a monograph-length elaboration and defense of this proposal.

[5] Interestingly, Elman is very explicit in his assumption that such investigation into the internal workings of these connectionist systems should contribute to cognitive theories of human linguistic capacities. See e.g. Elman (1991) p. 197.

Of course, the systems used in contemporary commercial NLP software are *massively* more complex than such simple connectionist systems. They typically have many more nodes, organised into many more layers, are trained on massive corpora, and feature elaborately articulated topological structures. This leads to severe difficulties in interpreting the internal working of such systems. Indeed, the 'opacity' of such systems has recently been emphasised in philosophical discussions of the ethical implications of the increasing role of algorithms in our social lives. Despite these difficulties in 'scaling up', there is reason to think that suitable analytic tools will be able to shed light on the internal workings of even large DL systems. Work such as Paudyal and Wong (2018) is already showing promise in this area.[6]

So, on the one hand, NLP systems are displaying ever-improving performance on standard linguistic tasks, to the point where they are at human or near-human levels. On the other, there is sustained effort dedicated towards identifying the internal workings of such systems. Together, these suggest the potential for contribution to theoretical cognitive science. If machines can perform the tasks that we can, and we can discover how they do so (i.e. which classifications they draw, which transitions they make, etc.), this may provide grounds for thinking that we perform these tasks in the same way. These machines would provide both an existence-proof that such solutions to these tasks are possible, and provide a model we can examine and even experiment on without the usual practical and ethical concerns that arise in trying to understand the psychology of human subjects. Returning to our earlier example, an NLP system trained on a naturalistic corpus containing both control and raising verbs could be inspected to see how it classifies such expressions, and the similarity or disparity between such classifications could then be leveraged as evidence in theoretical debates concerning the similarity of the processes used in generating such expressions in human language.

Something along these lines is proposed by several theorists. Norvig (2017) for example claims that "examination of the properties" of a "model containing billions of parameters" can provide "insight" into the workings of human language (p. 63). A recent survey of the relevance of DL models to cognitive science has claimed that "When deep neural models achieve state of the art performance in tasks highly related to human cognition, it is natural to ask what these models can suggest to cognitive science. This question is even more compelling when the task is the highest cognitive function of humans: language." (Perconti and Plebe (2020) p. 8). This line is then expanded on later in the paper with the claim that DL models "[already] process aspects of full-fledged human language" (p. 9). And a popular textbook in NLP states that "While practical utility is something different from the validity of a theory, the usefulness of statistical [i.e. computational, corpus-based] models of language tends to confirm that there is something right about the basic

[6] See Creel (2020) for an overview of this topic, and discussion of some strategies for avoiding 'opacity' in DL systems, and Prince and Schwarcz (2019) and Johnson (Unpublished Manuscript) for discussion of both the ethical need for such analysis, and the difficulties that arise in developing it.

approach." (Manning and Schutze (1999) p.4).[7] I believe, however, that a core methodological assumption of linguistic theory, the distinction between competence and performance, provides principled reason to be sceptical that such insight will be forthcoming.

4 Competence and Performance: A Functional Approach

Language, it is often claimed, is sound with meaning. We can update this intuitive and traditional proposal so as to better include gestural (and, perhaps, written) languages: language is perceptible sign with meaning. While issues surrounding meaning are notoriously murky, at least on this proposal the vehicles of these meanings, soundwaves, manual gestures, ink marks, etc., seem ontologically unproblematic and empirically investigable. A human linguist, or an artificial system, is able to examine linguistic tokens, investigate their properties and relations, and generalise about them.

The problem is, this proposed definition is incorrect. In line with the assumption that linguistics is a branch of cognitive science, mainstream generative linguistic theory assumes that linguistic expressions are not, strictly speaking, sounds, or publicly observable signs, at all. Rather, they are internal, psychological structures. Publicly observable signs may result from the generation of such internal structures, but only by involving a variety of other psychological processes to 'translate' the latter into the former. Some of these processes will be linguistically specific, but some will not, the latter being either specific to something other than language or psychologically general. Relatedly, some of these processes may be quite systematic and predictable, but some will not. This is all to say that perceptible signs are complex interaction effects, influenced by linguistic structures, but not determined by them. Given this, there is no guarantee, indeed it is empirically unlikely and presupposed by most linguists to be false, that robust scientific theories or empirically confirmed generalisations will apply to such public, perceptible signs.[8]

Consider a standard psychological model of linguistic production which maps a thought onto an utterance.[9] On standard assumptions, such a process requires: identification of relevant lexical items, morphological combination and transformation of these items to form words[10], grammatical composition of these words into a

[7] All of these proposals are more modest, and much more plausible, than the claim made in some popular, public-facing, discussions (e.g. Anderson (2008)) that DL models can not merely supplement, but *replace* theoretical science.

[8] This is in fact one of the major objections to Devitt's nominalist approach to linguistics: many of the central theoretical posits of linguistic theory seem to be essentially properties of psychological representations, not of public symbols. See e.g. Collins (2008).

[9] For simplicity, I am assuming that the thought is generated prior to linguistic processing, but this is not a requirement. See Dupre (2020) for discussion of the idea that the generation of a thought just is the generation of a linguistic structure.

[10] It is controversial whether this is a component of grammatical structure-building or a distinct process. See e.g. Embick and Noyer (2007) and other papers in part II of Ramchand and Reiss (2007) for discussion.

hierarchical syntactic structure, and phonological processing, sensitive to all three prior processes, generating a linearised structure suitable for use by the articulatory systems. However, even this structure is itself not a public, observable sign, but a psychological representation. Specific systems of speech production, as well as general mechanisms of motor control, modulated by various aspects of the speaker's mind including their communicative intentions, knowledge about their audience, etc., must be recruited to produce anything perceptible. On the basis of this public sign, similar, but (modulo peripheral differences) reversed, processes in the hearer can, if all goes well, recreate the internal (syntactic, morphological, phonological, semantic) representations that in the speaker caused such a production.

Some of these processes are the targets of linguistic theorising. Others, however, are not. For example, in my dialect both /kʰæt/ , with a final alveolar stop, and /kʰæʔ/ ending on a glottal stop, are legitimate ways of pronouncing 'cat'. There are systematic rules of phonology which ensure this. However, which of these licensed options I opt for in a given conversational context will depend on a wide range of non-linguistic, and more-or-less unsystematic factors, including who I am speaking to, how formal the situation is, how fast I am speaking etc. Even taking these into account, however, there is probably some residual variability. Further, if I am imitating a friend's accent, or eating, or drunk, my utterances of 'cat' may not resemble either of these. This variation in mapping stable, internal representations onto perceptible utterances is simply outside of the scope of linguistic theory.

All of this is simply to say that linguistics is a theory of *competence*, the underlying rule-system(s) partially responsible for the acquisition and use of language, whereas sounds, and any other perceptible signs, are instances of performance, actual linguistic behaviour.[11] Components of competence are causal influences on performance, but do not determine it. Performance thus provides evidence for a theory of competence, but not its subject matter.

Consider, in this light, a language model. A language model can take as input a linguistic string, and produce its best guess of a plausible continuation. Upon its recent unveiling, GPT-3, the Generative Pre-trained Transformer 3, a deep neural network language model of this sort, received international media coverage, due to its often uncanny ability to engage in conversation with humans and complete various kinds of linguistic tasks in human-like ways. Such a system is, obviously, aimed at reproducing human *performance*. One of its central goals, then, is to learn, on the basis of massive amounts of corpus-data, which utterances (in this case, inscriptions, not sounds) are acceptable and which are not. That is, it must learn a function capable of generating the (presumably infinite) set of acceptable English strings, on the basis of its lexical inventory.[12] Call this set '*P*', and call a function generating it

[11] 'Competence' here is being used in the technical sense introduced to linguistic theory by Chomsky (1965). This differs from other uses, such as Miracchi (2019)'s, for which a 'competence' is a reliable, agential behavioural capacity, not an underlying representational system which may or may not explain such a capacity.

[12] Even this is something of an idealisation. Being able to convincingly engage with human speakers need not involve being able to generate all acceptable English utterances. Many acceptable utterances would rarely if ever come up in conversation, e.g. the semantically bizarre or trivial, and so a failure to treat such utterances as legitimate may not undermine the performance of a language model.

'f_{Perf}'. Let's even assume, idealising a bit, that f_{Perf} is indeed acquired by GPT-3, so that it never produces unacceptable utterances, and, given the right sort of stimulus, could in principle produce any acceptable utterance.

Identifying f_{Perf} is no small feat. It will require not merely correctly classifying the data it is trained on, but accurately extrapolating from this to any arbitrary member of '*P*', in the manner of 'zero-shot learning' exemplified by GPT-3. Systems capable of this are able to correctly classify novel stimuli as members of categories not found in their training data. However, even learning capable of supporting such zero-shot classification subserves performance, and as we shall see there is no reason to think that the best system for reproducing performance is likely to closely match the systems actually used by human speakers.

As the discussion of this section has aimed to show, the goal of theoretical linguistics is not to identify f_{Perf}. The closest thing to *P* in theoretical linguistics is the set of legitimate linguistic structures. Call this set '*C*'. TL, then, is interested in identifying the function capable of generating C, call this 'f_{Comp}'.[13,14] f_{Comp} will then be a function from the basic constituents of language (i.e. morphemes) to the complex structures generated to serve at the interface between language and other cognitive systems, centrally semantic interpretation and motor control.

As we've seen, capturing f_{Perf} is likely the best case scenario for NLP. On the one hand, the engineering goals of this program are behaviourally defined. It doesn't matter to Google whether automated question-answerers, translators, text summarisers, etc. follow the same morpho-syntactic rules as humans, what matters is that their outputs are sufficiently similar to ours. On the other, DL systems require training data, and lots of it. But data, by its very nature, is observable. f_{Comp} maps unobservable representations onto unobservable representations, and so doesn't produce anything a system could be trained on, and thus there is no way for an DL system to learn it directly.

5 The Distance Between Competence and Performance

The question, then, is: is there reason to think that learning f_{Perf} will shed light on the function f_{Comp} that theoretical linguists are interested in.

Ultimately, this question boils down to a question of the distance between competence *C* and performance *P*. A fairly common assumption in both theoretical linguistics and NLP is that this distance is fairly minimal. Performance, on this view is competence plus some noise. Another way to put this is to say that one can largely abstract away from whatever extra-linguistic mechanisms are used in language

[13] Note that, whereas NLP is aimed at identifying *any* function capable of generating *P*, TL is aimed at identifying the specific function that actually generates *C* for human speakers. Another way to put this is to say that NLP aims at identifying f_{Perf}, understood as a function-in-extension, whereas TL is aimed at identifying f_{Comp} as a function-in-intension.

[14] Note also that f_{Comp} is itself an abstraction from the various more specific functions that will be the target of linguistic sub-disciplines: morphologists aim to uncover the function from morphemes to words, syntacticians the function from words to phrases, etc.

production/perception without missing out on much. If performance is a fairly direct reflection of competence, plus or minus a bit, identifying f_{Perf} might be highly informative for the theoretical linguist. And indeed, many paradigmatic cases of performance deviating from competence do seem to be 'noisy': people *umm* and *ahh*, repeat words, and revise sentences midway through, they mumble their words due to inattention, inebriation, a mouth full of peanut butter, etc. Even more systematic cases, such as memory constraints precluding very long or complex sentences, don't seem like they would pose a deep worry for NLP systems, capable as they are of extrapolating from the simple cases they encounter. And many NLP systems, especially convolutional neural nets, are often praised precisely for their ability to abstract away from noise in the stimulus.[15]

However, it is an empirical assumption that performance relates to competence in this fairly tidy way. There are multiple ways that it could turn out to be false. For one, 'performance effects' need not simply consist in adding noise to the signal of competence. It is perfectly possible that performance systems could add signal of their own. That is, non-language-specific psychological systems could contribute information to utterances that doesn't correspond to any aspect of linguistic structure. Some kinds of utterance, such as interjections or greetings, are plausibly of this sort. 'Hello', 'ouch', 'hey', etc. do not combine with other linguistic expressions in grammatical ways, and so may be genuinely para-linguistic. And there is no reason in principle why we couldn't learn whole phrases like this.[16] Our utterances of such phrases would then be more akin to the production of a drawing or a diagram, a product of general intelligence rather than language-specific systems, but this fact would not make itself apparent in the public sign itself.

Another possibility is that the process of uttering ('externalising') an expression produces something with very different properties, altering, destroying, or adding features in ways that preclude any 'backwards path'. If the externalisation of a linguistic structure involves these complex transformations, even if they are highly systematic, there is no reason to expect that a DL system trained on these outputs will be able to recreate the original structures. The point is not merely that multiple distinct processes can produce the same outputs (although this is of course true and important). Rather, it is that the underlying structures may be far from the simplest ways of generating the observed output, and thus the observed behaviour may be uninformative of the processes and rules used in its generation. Of course, although the set P does not logically determine that a specific function generated it, if there

[15] Additionally, it may work in NLP's favour that they are often trained on *written* languages, which typically conform more closely to stable conventions than spoken language. On the other hand, much work in NLP involves training systems on language use in online contexts which may bring with it various internet-specific deviations from linguistic rules.

[16] Note that my view that we can learn complex linguistic expressions as memorized tokens is a far cry from the proposal made by construction grammarians (e.g. Goldberg (2006) and Tomasello (2003)) that this is *all* that learning a language amounts to. While I believe that we should retains the 'rules and representations' approach of mainstream generative grammar, I believe that our linguistic behaviour need not exclusively stem from such rule-governed processes and is likely instead 'supplemented' by something more akin to the acquisition and deployment of constructions.

were a simple relationship between C and P, then a system aimed at identifying f_{Perf} might identify f_{Comp} as a likely underlying process. However, if C and P are related only in complex and unpredictable ways, discovering f_{Perf} may tell us very little about the target of interest, f_{Comp}.[17]

It is only the assumption that the contribution of the processes involved in externalising, i.e. in mapping C onto P, are either unsystematic (noisy) or uninteresting (simple), that allows us to infer from performance to competence. I believe that a strong empirical case can be made that these externalising processes cannot be idealised away. That is, in both of the manners just described, performance deviates from competence in substantial ways.

Firstly, it is plausible that a wide range of perfectly ordinary utterances are best viewed as conventional productive idioms, incorporated into speakers' linguistic repertoires wholesale, rather than generated by one's language faculty. The cases will vary from language to language, but examples of this in English plausibly include: echo-questions [1], pied-piping [2], non-constituent conjunction [3], and subject-dropping [4]:

1. Alejandro said what?
2. To whom shall I address the letter?
3. Holmes knows, while Watson merely suspects, that the butler did it.
4. Saw Marcel at the shops earlier. Seems well.

Each of these, in one way or another, poses problems for grammatical theory. Of course, this is not to say that grammatical theories which allow for 1-4 are *impossible*. But they are more complex, in ways that put significant pressure on developmental and evolutionary theories of language.[18] For this reason, it is often better to treat utterances of these types as outside of the domain of grammatical theory, as 'learned exceptions' to the rules governing our grammars.[19] To the extent that this is correct, DL systems trained on naturalistic data are liable to identify linguistic patterns which don't correspond to the underlying grammatical rules.

The second kind of deviation is plausibly widespread as well. Debates rage on about the extent to which the hierarchical structure of grammatical representations can be extracted from linearised performance data (see Linzen et al. (2016) for relatively recent empirical results), but much of this literature undersells how much information is lost in the process of externalising. Mainstream generative theory

[17] This point is thus the inverse of Firestone (2020)'s point that the *failure* of a machine to display human performance may tell us little about whether such a machine shares human competence.

[18] See Dupre (2020), Guasti and Cardinaletti (2003), for cases 1 and 2 respectively.

[19] The 'learned exceptions' I discuss here are unlike what are typically called 'idioms' in the generative literature (see e.g. Gehrke and McNally (2019)). The cases above are syntactically unexpected, on standard grammatical assumptions. But semantically, they are quite predictable. What Gehrke and McNally call idioms are the opposite kind of case. "Andreas kicked the bucket" is syntactically unexceptional, but its typical intended meaning is unpredictable without particular knowledge of the idiom in question. I leave open whether these 'semantic' idioms will pose similar problems for inferring facts about human language from DL systems trained on linguistic corpora.

adopts the 'copy theory' of movement, according to which argument expressions are typically (perhaps invariably) found at multiple locations in a sentential structure. Most familiarly, when a sentential subject is also the object of a relative clause, as in "The team that the millionaire bought lost", standard theories claim that the underlying syntactic structure contains two tokens of the noun phrase 'the team', one in each place it receives semantic interpretation (i.e. the structure is closer to: "[The team$_1$ [that the millionaire bought the team$_1$] lost]"). Contemporary generative theory suggests such displacement is much more widely found than has previously been assumed. For example, the 'VP-Internal Subject Hypothesis' (Koopman & Sportiche, 1991) has it that every sentential subject undergoes movement from its original position as the specifier of a VP towards the 'left-periphery' of the sentence where it is pronounced. Facts of this nature are obviously not made evident by the produced utterance, and so are unlikely to be uncovered by any DL system trained on corpus data. Similar kinds of worries can be raised on the basis of programs like distributed morphology (Embick, 2015), which argue that the words that seem, at the surface, to be the units of language are in fact complex products of a variety of underlying processes. Again, such complexity will be, from the perspective of any DL system trained on performance, invisible.

Note that the import of these phenomena is not that it is *impossible* that DL systems could, on the basis of such data, extract a complex of linguistic rules that correspond to the rules governing human language. It is just that there is no particular reason to expect this. The relation between internal structure, which theoretical linguists care about, and utterances is highly complex, likely more complex than some systems which simply generate these surface forms directly. As the theoretical goals of NLP turn not on identifying this complex underlying system, but just on generating these surface forms, there is no expectation, or requirement, that NLP systems will point towards the targets of theoretical linguistics.

A common thought is that, given that human beings learn their languages from their linguistic environments, it must be possible for a machine to do so as well. And indeed, given that a human is capable of acquiring a given language on the basis of exposure to a finite set of utterances, there must be some way a machine could be designed such that if we were to provide this set of utterances as input, it would develop in ways that would allow us to 'read off' human linguistic competence from its internal organization. The question is: how likely is it that such a machine would be constructed given the engineering goals of NLP? What I hope to have shown is that this is highly unlikely. On the one hand, if the generative tradition I am appealing to is on the right track, human abilities to acquire a language depend in large part on linguistically-specific innate structures. The aim of this discipline is to identify these structures, and whatever modifications the process of language acquisition produce in them. This is made difficult by the fact that much of the empirical data for linguistic theories is unreflective or only partially reflective of such structures, being explained instead by a wide variety of extra-linguistic psychological systems. On the other hand, the engineering project of NLP centrally relies on the use of general (i.e. not language specific) learning mechanisms, aimed at reproducing observable data. It is thus far from obvious why we should expect the end products of these quite different projects, TL and DL, to converge. Of course, an NLP designer *could*

wait for the completion of a TL theory of natural language, and incorporate such results into a DL system, but this would obviously not provide any insight for TL, and would be unlikely to further the practical goals of DL.

These empirical issues point to a deeper, methodological, reason why we shouldn't expect NLP and theoretical linguistics to converge on linguistic systems. Contrary to empiricist accounts of science, science is not aimed at predicting observations, but about using observations as cues to underlying reality.[20] A linguist, then, in developing her theories, is not aiming to predict performance, but to describe underlying competence. In performing this latter task, she may draw on a wide range of evidence. If a description of a linguistic phenomenon is found in one language, but is unlike that found in any other, this can be reason to reject this description. If neurobiological evidence suggests speakers treat two linguistic phenomena differently, this can be used to argue against proposals that they are the same. If a purported linguistic rule cannot be learned from environmental data, it must be innate, and if it is innate, it must be explained developmentally and/or evolutionarily, which brings with it substantive constraints. If children learning their native language, or adults learning a second language, reliably make mistakes of a particular sort, this may indicate that there are internal biases towards languages with a feature which explains this. And so on. None of these sources of data are available, or pertinent, to the task of NLP systems trying to extract regularities from linguistic corpora. Thus, NLP systems will be relevant to theoretical linguistics to the extent that these sources of evidence are irrelevant to the latter. Most work in theoretical linguistics indicates that this will not get us very far.

This last point shows us that the worry is not merely one of underdetermination of theory by evidence. That is, it is not merely that, given that competence falls short of determining performance, extrapolating from performance (as DL systems do) will not necessarily tell us about one determinant of performance, namely competence. This is true, but weak. The stronger point is that there are substantial reasons, empirical and methodological, to think NLP will, and indeed should, identify patterns and rules quite unlike those proposed by theoretical linguists. To the extent that NLP remains an engineering discipline, then, aimed at reproducing performance, it is unlikely to be of use to theorists of competence. Of course, one could look for ways to incorporate these kinds of constraints (neuro-biological and evolutionary plausibility, compatibility with results from first and second language acquisition, etc.) into a DL system. In doing so, however, one would deliberately be moving away from the engineering goals of NLP.

While I find the empirical arguments for a significant distance between competence and performance compelling, it is worth noting that they are highly controversial. In particular, this feature of linguistic theorizing is particularly prominent within mainstream generative grammar, as exemplified by the Minimalist Program (Chomsky, 1995). A variety of other linguistic approaches are, in the terms of Sag and Wasow (2011), much more "surface-oriented". That is, these approaches, such as Head-Driven Phrase Structure Grammar (Pollard & Sag, 1994), Lexical-Functional

[20] See Dupre (Forthcoming) for an argument along these lines.

Grammar (Kaplan et al., 1981), Simpler Syntax (Culicover & Jackendoff, 2005), and Construction Grammar (Goldberg (2006) and Tomasello (2003)), insist on much stricter correspondence between the structure licensed by a given linguistic theory and observable properties of linguistic utterances. The arguments for this are largely methodological, rather than empirical: linguistic theory is on better evidential footing when it aims to capture all or most observable aspects of language, rather than just that, perhaps small, subset of observable linguistic behaviour reflective of a purported causally distant competence. While I hope that some of the discussion above has pointed to why such arguments are mistaken, the goal here is of course not to argue for one linguistic theory over another.

It is worth noting, however, that even those theorists who defend such 'surface-oriented' approaches to language tend to endorse some version of the competence-performance distinction. Christiansen and Chater (2016) say that "Language, like any other empirical phenomenon, is the product of many factors, and explaining language requires separating the influence of these different factors." (p.234). Jackendoff (2002) (p. 34) argues for a 'soft', 'methodological' interpretation of this distinction. And even Michael Tomasello (2000), who elsewhere (Ibbotson & Tomasello, 2016) claims that such a distinction makes a grammatical theory immune to falsification by observation, recognizes the empirical need to disentangle the effects of performance constraints from the contributions of competence in explaining linguistic behaviour. This is for good reason. It is a truism that people's speech behaviour need not correspond to their knowledge of the rules of language, given the ubiquity of disfluencies, false starts, etc. in conversation. The question then is just how wide this gap is. While I am willing to bet, with mainstream generative linguistics, that the gap is quite substantial, for the purposes of this paper, I am happy with the conditional conclusion: *to the extent that* mainstream generative proposals are on the right track, and competence and performance may bear only faint resemblance to one another, DL models are unlikely to provide insight concerning the nature of linguistic competence.

To reiterate a point from earlier, the gap I have argued for between competence and performance does not show us that NLP will be totally irrelevant to linguistics generally. In particular, it is plausible that, once linguistic competence is described, computational models could be used to test how much of this could be learned from the child's linguistic environment.[21] This would be very significant in apportioning linguistic knowledge between the innate and the learned. What I hope to have shown is that the prior task of identifying the structure of competence itself is unlikely to be advanced by examining the computational models utilised in NLP.

[21] Although there are difficulties here, too, as machines typically require significantly more data than child learners, thus making inferences from their abilities to ours uncertain.

6 Conclusion

Quine (1960) (p. 8) compared language users to topiary elephants. While the external shapes may be identical, they are made so by quite different internal structures. Decades of work in generative linguistics seems to show that this analogy is empirically mistaken in the case of human language, for which the universal, internal constraints seem to vastly outstrip the role of experience in ensuring similarity between speakers. But it seems entirely appropriate when comparing human and machine learners. Powerful DL systems may be capable of reproducing the observable shapes of human language. And, of course, this should be viewed as a significant success from an engineering perspective. But, if the generative tradition is on the right track, they are likely to do so only with radically different internal properties, and so are unlikely to be theoretically illuminating.

Acknowledgements A draft of this paper was presented as a Royal Institute of Philosophy lecture at Keele University. Thanks to the audience at that talk for helpful feedback. Thanks also to Gabbrielle Johnson, for commentary and discussion on several versions of this paper.

Funding Funding was provided by Leverhulme Trust.

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine, 16*(7), 16–07.

Boeckx, C., Hornstein, N., & Nunes, J. (2010). *Control as movement*. Cambridge University Press.

Chomsky, N. (1957/2002). *Syntactic structures*. Walter de Gruyter.

Chomsky, N. (1965). *Aspects of the theory of syntax*. The MIT Press.

Chomsky, N. (1995). *The minimalist program*. The MIT Press.

Christiansen, M. H., & Chater, N. (2016). *Creating language: Integrating evolution, acquisition, and processing*. MIT Press.

Collins, J. (2008). A note on conventions and unvoiced syntax. *Croatian Journal of Philosophy, 8*(23), 241–247.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science, 87*(4), 568–589.

Culicover, P. W., & Jackendoff, R. (2005). *Simpler syntax*. Oxford University Press.

Devitt, M. (2006). *Ignorance of language*. Oxford University Press on Demand.

Dupre, G. (2020). What would it mean for natural language to be the language of thought? In *Linguistics and philosophy* (pp. 1–40).

Dupre, G. (Forthcoming). Realism and observation: The view from generative grammar. *Philosophy of Science*.

Elman, J. L. (1990). Finding structure in time. *Cognitive science, 14*(2), 179–211.
Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*(2), 195–225.
Embick, D. (2015). *The morpheme: A theoretical introduction* (Vol. 31). Walter de Gruyter GmbH & Co KG.
Embick, D., & Noyer, R. (2007). Distributed morphology and the syntax/morphology interface. In G. Ramchand & C. Reiss (Eds.), *The Oxford Handbook of linguistic interfaces*. Oxford University Press.
Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences, 117*(43), 26562–26571.
Gazdar, G., Klein, E., Pullum, G. K., & Sag, I. A. (1985). *Generalized phrase structure grammar*. Harvard University Press.
Gehrke, B., & McNally, L. (2019). Idioms and the syntax/semantics interface of descriptive content vs. reference. *Linguistics, 57*(4), 769–814.
Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.
Guasti, M. T., & Cardinaletti, A. (2003). Relative clause formation in romance childs production. *Probus, 15*(1), 47–89.
Hornstein, N. (1999). Movement and control. *Linguistic Inquiry, 30*(1), 69–96.
Ibbotson, P., & Tomasello, M. (2016). Evidence rebuts Chomskys theory of language learning. *Scientific American, 315*(5)
Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.
Johnson, G. (Unpublished Manuscript). Proxies aren't intentional, they're intentional.
Kaplan, R. M., Bresnan, J., et al. (1981). *Lexical-functional grammar: A formal system for grammatical representation*. Citeseer.
Katz, J. J. (1980). *Language and other abstract objects*. Rowman and Littlefield Publishers.
Katz, J. J. (1984). An outline of platonist grammar. In T. G. Bever, J. M. Carroll, & L. A. Miller (Eds.), *Talking minds: The study of language in cognitive science* (pp. 17–48). MIT Press.
Koopman, H., & Sportiche, D. (1991). The position of subjects. *Lingua, 85*(2–3), 211–258.
Linzen, T. (2019). What can linguistics and deep learning contribute to each other? response to pater. *Language, 95*(1), e99–e108.
Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMS to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics, 4,* 521–535.
Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.
Miracchi, L. (2019). A competence framework for artificial intelligence research. *Philosophical Psychology, 32*(5), 588–633.
Norvig, P. (2017). On chomsky and the two cultures of statistical learning. In *Berechenbarkeit der Welt?*, (pp. 61–83). Springer.
Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language, 95*(1), e41–e74.
Paudyal, P., & Wong, B. W. (2018). Algorithmic opacity: making algorithmic processes transparent through abstraction hierarchy. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, (pp. 192–196). SAGE Publications
Perconti, P., & Plebe, A. (2020). Deep learning and cognitive science. *Cognition, 203,* 104365.
Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.
Prince, A. E., & Schwarcz, D. (2019). Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev., 105,* 1257.
Quine, W. V. (1960). *Word and object*. MIT Press.
Ramchand, G., & Reiss, C. (2007). *The Oxford Handbook of linguistic interfaces*. Oxford University Press.
Sag, I., & Wasow, T. (2011). Performance-compatible competence grammar. In R. Borsley & K. Börjars (Eds.), *Non-transformational syntax: Formal and explicit models of grammar*. Wiley.
Sanger, D. (1989). Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. *Connection Science, 1*(2), 115–138.
Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74*(3), 209–253.

Tomasello, M. (2003). *Constructing a language*. Harvard University Press.
Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.