DOI: 10.26346/1120-2726-248

Doing Linguistics in the GPT era

Roberto Zamparelli

Center for Mind/Brain Sciences, Università di Trento, Italy <roberto.zamparelli@unitn.it>

Chesi's position on the relation between theoretical linguistics and large language models (LLMs) should be taken seriously by anybody who has worked on these two ways of approaching language. This commentary largely shares his general viewpoints, with some qualifications. One is that standardization (and the search for a broad core of compatible analyses) is more important than formalization. The second is that the size and opacity of the materials used for training the largest LLMs makes them essentially useless as models of human linguistic competence (though perhaps not entirely implausible as models of the evolution of a creature capable of using language). On the other hand, smaller models, trained on human-sized amounts of language (as in the the BabyLM challenge) could become useful tools to study human competence and understand the limits of exclusively data-driven approaches.

KEYWORDS: Language models, LLM, theoretical linguistics, computational linguistics.

1. Introduction

Cristiano Chesi's paper is many things: a plea for the formalization of linguistic theories, an invitation to construct shared test-sets to evaluate them (on the model of SyntaxGym[1] or COLA[2]), a defense of the importance of judgment gradability, all too often disregarded as 'performance', and a provocative take on what it means to do linguistic theorizing in the era of ChatGPT (the title itself is a minor provocation, following up the major provocation constituted by Piantadosi's 2023 opinion piece: current Large Language Models, LLMs, disprove generative linguistics). At a higher, strategic level, the paper is an invitation for theoretical linguists to take the methods and goals of computational linguistics quite seriously: comparability of results, attention to gradience, broad within-language coverage. Failure to play this game – Chesi argues – could soon make classic theoretical linguistics marginalized by funding agencies, and ultimately by hiring committees.

Apart from some caveats, discussed in the next sections, I share many of the views expressed here. Even if we firmly believe that theoretical linguistics should be a theory of competence, it is hard to argue that a theory that can predict competence and performance should not

 (received February 2025)

be superior to one that only covers competence, especially if the performance effects are systematic (not due to, say, drunkenness, but to normal language processing constrains). Even better if a theory is capable of disentangling the two sides and talk about competence as the limit case of good performance (language processed with unlimited resources).

Chesi also makes the point that derivational theories of well-formedness should be considered superior to pure representational theories, where the final tree is evaluated regardless of the steps taken to assemble it. The argument here is one of efficiency: representational theories work as filters that throw away ill-formed trees, but derivational ones could prevent those trees from being generated in the first place. But if the superiority of derivations is accepted, it is a small step to argue – as Chesi does – that the best derivation should be the one that processes a sentence such as (1) following the temporal order of the words, i.e. left-to-right and top-to-bottom, not the other way round, as in standard Minimalism.

(1) [Martha [has [come [to [believe [that [Marc [should [have [hired [a [theoretical [linguist.]]]]]]]]]]]]]

These points are familiar for anyone who has followed the work of Chesi, the creator of a formalized, computational, left-to-right grammatical formalism (Chesi 2007; Chesi & Bianchi 2014),[3] but they are given a new urgency by the existence of LLMs, which use machine learning to beat any symbolic theory in 'generative' ability (understood not just as the ability to generate 'correct language', but to generate correct language which is appropriate for a given linguist context). If LLMs are theories – a big *if*, if we assume that 'theory' is not just a relabeling for 'generator', but rather a tool to bring about some increase in understanding – then we have broad coverage and learning from pure data: the empiricist linguist dream.

2. Formalization or standardization?

The issues presented above have at least two facets. One is whether formalization and computational implementation are the right tools to address the crisis that Chesi and others (see especially Pater 2019; Baroni 2021) have identified. The second is how comparable LLMs' use of language is to human use, and what we can learn from it.

There is a trade-off between formalization and coverage. The range of constructions that can be analyzed by Stabler's or Chesi's fully formalized systems is small compared to the variety covered by non-formalized

generative theories of the past (say, Government and Binding; the various incarnations of Minimalism probably sit somewhere in between). At a practical level, there is a tension between formalization and ease of creation: authors need to be convinced that the cost of learning and using formalized notions is worth it, and especially, that they are investing into the 'right' formalism. It is interesting to note, incidentally, that whatever knowledge LLMs have is not formalized (their architectures are).

What we should invest on depends on which aspects of language are important for us. Most papers published in linguistics today are not trying to recast the foundations of the theory, but aim to use existing theories and previous analyses to give a treatment to a specific phenomenon in certain languages. The success of their results can be taken as an indirect confirmation that the bases they rest upon are viable, but very few papers strive to prove that with a different toolbox the analysis of a certain phenomenon would be 'impossible'. Papers can borrow parts of previous theories (say, the probe-goal agreement mechanism) without realizing that taking the whole could actually damage their analysis. Or they can adapt a theory to their needs without thinking that the changes introduced break other analyses based on it.

Nowhere is this more visible than in the domain of functional projections and their properties. The determiner phrase (DP, which others still/again call 'extended NP') contains an NP proper (or maybe an uncategorized root, see Borer 2005), under a projection that could host classifiers in Chinese (Cheng & Sybesma 1999), a projection for adjective-like numerals (*the next three people*) and at least one for determiners (so minimally DP - NumP - ClP - NP, though many authors conflate the first three). However, the names and functions of these projections vary across authors and publications, with N said to denote sets of atomic individuals (Link 1983), or masses (Borer 2005: Ch. 4), or individual kinds (Chierchia 1998), or set of kinds (McNally & Boleda 2004); Cl, also referred to as Num, creating 'discreteness' (Borer 2005), or definiteness (Espinal & Cyrino 2022); Num denoting a cardinality filter (Landman 2003), or a set of degrees (Kennedy 2015), or the creator of plural properties (Ionin & Matushansky 2006), or a function that 'undoes' definiteness, (Cheng & Sybesma 1999; Espinal & Cyrino 2022). D, in turn, could be obligatory (Longobardi 1994) or missing in some languages (Bošković 2005); if present, it could be split in multiple positions to host determiners with different properties, e.g. those that go with predicate nominals (*a/the/two*) and those that do not (Cheng *et al.* 2017); or articles *vs* other determiners (the former seen just as fragments of nominal inflection, Giusti 2015). These heads and functions have a family resemblance that does not translate into identity. What seems to be needed, more than formalization, is 'standardization': adopting a similar vocabulary of items with shared, agreed-upon properties, bringing

each new analysis to bear on a broader set of facts than the cases it was designed for, and doing this by means of public testsets (sets of constructions across languages), as advocated by Chesi.

Indeed, the greatest advantage of a computational implementation may be not cognitive plausibility, but 'shared responsibility'. We see this in open-source software projects: each participants works on a portion of code within a larger project, making changes and documenting the process; when done, he or she proposes to merge the new subpart into the whole; the community tests the addition to see if it breaks anything. If not, it is accepted, until it gets replaced by faster code or superseded by broader changes. There are fights, to be sure, but they have to be resolved, because every project member knows that the best subroutine is worthless if the whole does not work.[4]

This shared construction model of building a global theory may seem illusory for linguistics (a 'young science', we are constantly reminded), but the open-source software movement truly started with the internet, so it is younger, yet it succeeded. The problem is not just having a shared set of problems like COLA, but a shared set of primitives and operations, along with their semantic and syntactic justifications. If these primitives are modified ('forked', in the terminology of open software), the authors of the modifications should feel the obligation to merge them with the mainstream assumptions, having checked their effect on the theory as a whole. Indeed, one positive effect of LLMs, which are generalist by nature, has been to reorient the attention to linguistic 'coverage', after decades spent worrying about the most 'perfect' design for language (Chomsky 1995).[5]

LLMs do not have the problem of conflicting representations: the growth of the deep learning paradigm occurred when its creators shifted from systems that took as input predefined features (decided by researches on grounds of plausibility and often hand-extracted) to so-called 'end-to-end systems', which take in input raw text and produce raw text. These systems have no input bottlenecks, and are free to develop whatever internal representations they find useful to produce an output – a successful strategy in terms of results, but a major contribution to the so-called 'black-box problem': the opacity of their inner workings, and the fact that ten different training rounds of the same architecture could result in ten very different solutions.

3. Theoretical linguistics in the land of GPTs

Are LLMs relevant for understanding humans language? What are they really modeling? A notorious problem for linguistics is that, unlike

for other cognitive domains, we have no animal models: only humans speak. When LLMs came about, it was tempting to see them as the closest thing to an 'animal model' for language. The added bonus was that no animal function has developed by 'imitating' a human function, so in principle LLMs could be even closer to our mental processes, insofar they are reflected in language. Soon enough, cognitively-minded scholars started to complain that trained LLMs were black boxes – but human brains are black boxes, too, and at least we can lesion and alter LLMs in ways that would be highly unethical in children. In addition, we rarely appreciate the fact that a purely symbolic theory of language that could specify the syntax and semantics of every words in every construction a LLMs can cover, down to the level of probability of occurrence in complex and diverse contexts, would most certainly produce analyses as intricate and incomprehensible as the innards of LMMs. Then there is the fact that, unlike static theories, LLMs are keen to help.

This state of affairs has generated a subfield of computational linguistics (LODNA, 'linguistically-oriented deep net analysis', in the terminology of Baroni 2021) whose goal is investigating the differences between what LLMs know about language and what humans know. Can LLMs give metalinguistic judgments? Can we reach an understanding of their inner functioning sufficient to make predictions about their behavior WITHOUT ACTUALLY RUNNING THEM? How important is the size of their training set, and the style of training?

At the level of network analysis, LODNA has a long way to go. The current attempts at breaking the LLM black boxes rarely go beyond identifying 'where' in the network information about specific lexical features can be identified or what part of the text the individual attention heads attend to. The flow of information processing that runs through the models remains poorly understood, except in individual cases (Lakretz *et al.* 2021), in part because each model can develop its solutions in different ways: there is no shared Broca's area for deep learning networks.

On the other hand, when we analyze the 'output' of LLMs with carefully crafted testsets, many of the questions above start to receive an answer. We now know that the largest LLMs show remarkable metalinguistic skills, in terms of binary or graded human-like linguistic judgments (Wilcox *et al.* 2018; Warstadt *et al.* 2019; Wilcox *et al.* 2024; Haider 2023), mixed with some very non-human performances (Ettinger 2020; Katzir 2023; Haider 2023: section 7); sometimes, performances that approximated humans' are only found with styles of training expressly designed to foster systematicity (Lake & Baroni 2023), or when the results have high probability (McCoy *et al.* 2023).[6] It is also clear that the size of the training data matters: state-of-the-art LLMs are exposed to terabytes of

text (many thousands of lifetimes worth of language). To some degree this could be counterbalanced by the fact that LLMs do not actively interact and do not have visual input, but the fact that blind and paralyzed individuals can learn to speak and that young children do not typically ask questions related to linguistic forms makes it unlikely that this deprivation could affect LLMs' performance on core SYNTACTIC phenomena.[7] At the very least, the fact that massive LLMs learn hardly makes the case for the Poverty-of-the-Stimulus (PoS) situation that the human baby might experience. Models trained on more child-sized amounts of language (say, 10 million words, corresponding to 2 to 5 years of age, as in the 'strict-small' category of the BabyLM challenge, Warstadt *et al.* 2023), retain good judgments for testsets like BLIMP[8] or SuperGLUE (Sarlin *et al.* 2020), but fall short in a task the tests the ability to use structural rather than linear order generalizations (MSGS, Warstadt *et al.* 2020). In addition, even the best BabyLM challenge models could not be tested by simply asking it to perform some linguistic task ('zero-shot'), but required additional task-specific fine-tuning; as Chesi notes, they also required hundreds of 'epochs' (i.e. exposures to the same data: 2000 for the best model in the 10M word BabyLM category). Children undoubtedly hear many repetitions, but these are already present in the child-directed-speech data used for the BabyLM training, and in any case the repetitions are likely to follow a Zipfian distribution, not 'every sentence' 2000 times. Last but not least, the judgments in COLA and BLIMP are all about syntax or morphology (modulo NPI violations, and some Verb-Object subcategorization violations: *Kim persuaded it to rain*): conspicuously absent are all semantic data of the form: *Is meaning X available from string Y?* While one could treat the AI as a human subject and try to extract such judgments, to the best of my knowledge all explorations of the metalinguistic capabilities of LLMs have used the acceptability of STRINGS (though SuperGLUE at least requires broader understanding of PASSAGES).

Chesi's position on whether LLM are or will be able to capture human linguistic competence falls in the strong PoS stance: certain phenomena cannot be learned at all without prior biases. I disagree. While the models might never encounter the information that, e.g., *The girl that talking to makes Bill blush is here.* is ungrammatical, we simply do not know which types of 'indirect' evidence current or future models could bring to bear on these cases (for instance, they might note that dangling prepositions to the left of tensed verbs are exceedingly rare, or recognize John-Mary sentences as linguistic example, and judge them according to similar sentences they read in LI). I believe that a combination of more powerful architectures trained with human-sized amounts of data and epochs could still give important contributions to linguistics. When they

will fail on rare constructions that humans get right it, they will need to be supplemented with some kind of structural (i.e. 'innate') knowledge. I suspect that insights on which knowledge we should add will come from theoretical linguistics, but also from the other extreme, from ultralarge models. In my opinion, huge LLMs are useful if they are seen not as models of the learning path of a baby, but as models of the evolution of organisms from undifferentiated proto-cells to college students: 99% of the process is not about ontogeny, but phylogeny. The structures they develop toward the end of their training could be just the ones that we should add to their smaller cousins as 'innate baggage'.

In short, I think theoretical linguists should pay close attention to what happens in the CL community, including their large, cumbersome, unrealistic models. But they should not play the part of ornithologists who think Concorde is superior to DC-9 as a theory of bird flight because it flies faster.

Notes

[1] <syntaxgym.org>.

[2] <paperswithcode.com/dataset/cola>.

[3] The approach in Phillips (2003) is similarly left-to-right, but not formalized; work by Kempson & Cann in the UK (Kempson *et al.* 2000; Cann *et al.* 2005) is similar in spirit, left-to-right, formalized in a more logic-based framework, but as far as I know not computationally implemented.

[4] A dated but still effective description of the process is Raymond & O'Reilly (1999).

[5] Ideally, the aspiration toward a perfect design for language should be evaluated in the context of the degree of perfection of other biological subsystems: vision, sleep, digestion, etc. How perfect is our 'competence' as sleepers?

[6] Part of the empirical problem here is that the LLMs under study change all the time, and not in a transparent or consistent way. A question to e.g. GPT4 that has received a certain answer in one paper might receive a different one in another paper (compare the reports in Haider 2023 and Katzir 2023), and it is quite possible that publications reporting LLMs' limitations are fed back into the system to improve the next version.

[7] This is confirmed by the poor results of the 'loose' category in the babyLM challenge, which was allowed to use multimodal data, see Warstadt *et al.* 2023: section 7.1.

[8] <github.com/alexwarstadt/blimp>.

Bibliographical References

See the unified list at the end of this issue.