Symbol Grounding in the Age of LLMs

Benjamin GASKIN[1]
[a] *The University of Sydney, Australia*
ORCiD ID: Benjamin Gaskin https://orcid.org/0000-0003-0277-2883

Abstract. This paper considers symbol grounding in its practical and theoretical aspects. Taking up the theoretical perspective, we begin by considering the relative inefficiency of large language models in acquiring language. A framework is introduced based on the concept of morphological computation and formalised with reference to conditional Kolmogorov complexity: that the form of embodied experience scaffolds human language acquisition. This argument is extended to consider the symbol grounding problem, with particular reference to the origin of language in both the individual and historical sense. It is argued that, while humans also make use of statistical learning, the process of symbol grounding via morphological computation is essential at the origins of language and during early development. It provides a minimal ontology in terms of objects, containers, processes, etc.—basic features which language models must instead brute force by statistical means. The paper closes by reconsidering the symbol grounding problem in light of recent advances, particularly the promise of multi-modal models and robotics, and ultimately concludes that the status of the symbol grounding problem depends upon our aims in the pursuit of artificial intelligence.

Keywords. Symbol grounding, large language models, multi-modal language learning, multi-modal language models, robotics, morphological computation.

1. Introduction

Symbol grounding can be seen to have two aspects which are commonly conflated, which we will here frame in terms of the movement of meaning between world and model. These correspond, moreover, to separate aims in terms of what it would mean to solve the symbol grounding problem. The *practical* aim of symbol grounding is to ensure predictable behaviour—that internal symbols correspond regularly and comprehensibly to concrete actions; it is thus concerned with the movement from model to world. This aim is simply a matter of alignment, of aligning the 'conceptual' structures of artificially intelligent systems with the world of activity. Such an alignment is of increasing importance as we develop and put in place advanced systems such as transformer-based multi-modal language models, and particularly with the further integration of these systems with robotics as in the case of Google's RT-2 [1]. The *theoretical* aim, on the other hand, is more nebulous, with the basic question being how symbols attain their meaning; it is thus concerned with the movement from world to model.[2] While related, these two movements cannot be treated as identical. An understanding of the latter could

[1] Corresponding Author: Benjamin Gaskin, bgas0204@uni.sydney.edu.au

[2] The notion that meaning is ultimately derived from the world is not uncontroversial by any means, but it is highly plausible if we accept an evolutionary perspective on the origin of animal life and human cognition. We might further take this paper to be presenting a formalised and testable hypothesis in support of this view.

well be done without, for instance, as the aim of achieving safe and effective social robotics might well be achieved as a purely practical question.

The present paper, however, will focus primarily on this theoretical aspect of symbol grounding. To this end we will provide a framework for the theoretical aim of symbol grounding, for the movement of meaning from world to model, with reference to the concept of morphological computation. This term is typically used with reference to systems wherein physical bodies reduce the computational load of a system performing a given task, as where the skeletal structure of a bipedal robot is designed to 'carry' the computational demands necessary for controlling locomotion on uneven surfaces [2]. We will here by analogy consider linguistic meaning in terms of morphological computation, with a focus that crucially includes—but also essentially exceeds—the physical body. The idea, in short, is that the morphology of experience provides a structure which humans use in the acquisition of language. Symbol grounding, in other words, is achieved through embodied experience within a social and physical environment.

We here intend to further provide a practical and testable formalisation, suggesting that the design of multi-modal robotics might leverage these aspects in the training process rather than solely emphasising the movement from model to world. The practical difficulty here will be in designing architectures capable of successfully integrating the full spectrum of multi-modal information, although we will here introduce early findings with existing architectures that support our thesis. While the current generation of models have made impressive breakthroughs, including in multi-modality and robotics, and further seem primed for more, even RT-2 is not truly embodied in any integrative or developmental sense. It is rather based upon a pre-trained vision-language model which is subsequently trained to produce sequences of token-based action instructions [1]. This is significant in its own right, but such downstream integrations of robotics will not allow for the sort of world-to-model symbol grounding with which we are here concerned.

These considerations of symbol grounding in its theoretical aspect, and particularly the framing of this in terms of morphological computation, initially stemmed from the puzzle presented by the relative inefficiency of language learning in large language models as compared to humans. To achieve a linguistic facility comparable to an adult the language model must be exposed to a far greater quantity of linguistic material than a child, not mention the further disparity of costs in terms of energy and other materials. The question, then, is why this would be the case—what is it that allows the child to learn so much more efficiently? We will here suggest an answer in terms of morphological computation, here formalised with reference to conditional Kolmogorov complexity, before turning to further consider the matter of symbol grounding in light of this view.

Of course, there are those who would argue that the inefficiency here is a matter of neurological architectures, that the human brain is simply a much more efficient system compared to even state-of-the-art transformer models. Proponents of such an view would expect then that this relative disparity in terms of efficiency could be ameliorated and ultimately brought into line with architectural advances, perhaps in a form akin to the structure state space model developed by Gu and Dao [3]. Taking this particular proposal as an example, we may note that this only promises a system where the computational cost of processing a given unit of text scales linearly with the length of that text—as against the quadratic scaling of self-attention in the standard transformer model. There is, however, little reason to believe that such benefits for inference efficiency will reduce the quantity of training data necessary for linguistic performance; it is this measure with which we are foremost concerned, and in terms of which we here define efficiency.

2. The Kolmogorov complexity of language

We thus propose the following argument from morphological computation: that the relative efficiency of human language learning, as against that of large language models, results from leveraging the totality of experiential data to which the child is exposed during cognitive development. This can be formalised with reference to the concept of Kolmogorov complexity [4]. As a basic concept, Kolmogorov complexity is a measure of the shortest program that will reproduce in full a given mathematical object. The extent to which such an object can be compressed is thus taken as a measure of its complexity. We will here extend this argument with reference to what is known as conditional Kolmogorov complexity, which is a measure much the same but conditional upon access to some auxiliary system. This has been used in transfer learning, for instance, to measure how much constructive information one task provides about another [5]. The conceptual argument underlying our basic thesis here is that as parsimonious designer as evolution would certainly make full use of all available information in the construction of language.

Our case can be made by way of a classic example of Kolmogorov complexity. Take the following sequence: 14159265358979323846264338327950288419 71…. We might on the face of it consider this a complex sequence—that is, one which is not readily compressible, which has a high Kolmogorov complexity. Supposing, however, one knew the structure that generated this series—in other words, that these are digits of Pi—then its complexity is significantly reduced. This example, while imperfect, illustrates this basic notion of conditional Kolmogorov complexity: our argument is that having access to the structure of experience is analogous to the insight provided here by recognising Pi.

Of course, we must make a further modification of Kolmogorov complexity for our purposes here. Instead of the length of a program we will consider the length of training data considered as a series of tokens—i.e., the quantity of linguistic information—necessary for an individual to acquire language. By 'acquire language,' we here refer to a given level of aptitude as defined by some standardised evaluation or benchmark such as GPQA or MMLU. We must also consider this sequence to correspond equally to written and spoken language, as children will learn primarily through spoken language.

Taken together, therefore, we will compare two sequences: that which the human requires in the course of their 'training,' and that required by a particular language model. We will then align these in terms of an adult human, or representative population thereof, and a language model with equivalent performance on some standard benchmark. The formalisation resulting from this perspective would then be as follows:

$$K(L|E) < K(L)$$[3]

The intuition captured by this formalisation, in short, is that integrative access to the structure of embodied perceptual experience will reduce the complexity of language. While the number of tokens used to train the current state-of-the-art in large language models—such as Claude, Gemini, and GPT-4o—remains a secret, one estimate of the disparity is that language models require three orders of magnitude more data than even a highly literate adult human [6]. Our argument is that much of this difference may be explained by the role of experience as morphological computation in language learning.

[3] Where K refers to Kolmogorov complexity measured by the length of training data as a string of language tokens, L refers to a given level of linguistic aptitude, and E refers to embodied perceptual experience.

3. Experience as morphological computation

We will here provide two examples for a further categorisation of this sort of morphological computation, broadly considered in terms of external and internal symbols—corresponding to structures with perceptual and proprioceptive roots. External symbols, which tend most obviously to depend on sight, include, for instance, a *tree*. The tree is a common enough object in the human environment, albeit perhaps rarer than it once was, but it is something we can expect a developing infant to encounter. They are thus made familiar with the structure of this symbol through their experiential encounters with trees, most significantly through vision. This experience grounds the symbol by providing the shape which is its meaning—that is, the nature of the thing whereby the symbol circumscribes its proper usage. There are also further elaborations from this basic symbol, likewise bound, as with the wide variety of linguistic forms metaphorically derived from that of the tree: family trees, tree searches, roots, branches, and so on.[4]

Internal symbols, on the other hand, depend upon such modalities as the proprioceptive and enactive. Such symbols include, for instance, to *kick*. Here we can compare the learning process which occurs in a language model to that which occurs in a human child. The word 'kick' for a language model is first tokenised, then converted to an n-dimensional embedding which represents the meaning of the word as derived from its co-occurrence with a textual corpus. This representation then features in the set of training data for the model, whereby—the process of loss minimisation during prediction—the model derives the meaning of the word (i.e., its proper, or probable, usage) with reference to the statistical structure of the training data. The language model learns to use the word, as it were, by brute forcing its meaning, by distilling its usage from a mass of text. The child, in contrast, knows first what it is to kick and only later learns the word. Our argument, in short, is that this experiential structure, the embodied meaning of 'kick,' provides a scaffolding for the efficient learning of meaning as usage.

While the separation of these cases, of external and internal symbols, may be somewhat contrived, the basic idea remains the same: in either case we are concerned with the notion that the morphology of experience reduces the complexity of language acquisition by circumscribing the usage of a term—that is, experiences operates here akin to morphological computation. There are two simple studies which support this view. Wang and colleagues used a vision encoder to initialise the hidden state of a language model—albeit using the earlier long short-term memory (LSTM) architecture rather than a transformer model [7]. This yielded statistically significant benefits for model performance, particularly for nouns and verbs. Similarly, Zhuang and colleagues found that integrating visual data during a word-based language modelling task provided a particular benefit in situations where training data was limited [8]. These studies both provide initial support for our thesis, albeit in cross-modal rather than truly multi-modal cases. While early, they are nevertheless clear cases in line with our basic perspective.

[4] This is not to say that someone who had not encountered a tree could not understand these, and in fact one may well acquire an understanding of trees from familiarity with the pattern underlying this coherence in alternative forms; it is not inconceivable that one might use tree search algorithms to provide structure to trees.

4. Learning and the edges of language

We must yet note, of course, that there are certainly cases where human language learning takes place by what might be called a statistical method. This may well be especially common in the contemporary world, where many words only distantly refer to any sort of experiential structure. The notion of comprehension, for instance, hardly triggers any immediate thought of hands—of prehensile grips—in the population at large. This is only a simple example, but we can generally see much in the development of language as an increasing abstraction of meaning by transfer and recombination. The argument hidden here, which we ought to state explicitly, is that symbol grounding in the sense with which we are concerned is most essential at the earliest origins of language.

There are cases in which children have been deprived of language entirely, where they are understood to have missed the 'critical period' and hence seem to never develop language in the full sense of this term [9]. These seem to be instances where statistical learning, imitative and social learning, has been barred—and to that extent, they are significant for indicating the necessity of this. More important for our purposes, however, are cases where groups of deaf children have, in the absence of any extant sign language, been found to develop their own language from the ground up, as it were [10]. Our argument is that this ground can be understood in terms of communication structured by morphological computation based on embodied activity and perceptual experience.

We might compare this to the possibility of language models, left to themselves, learning language. Such a thing is patently absurd, even for multi-modal models or those endowed with robotics, at least within the current paradigm. It may be possible to create the necessary conditions and capabilities, but that would seem to require something else entirely. Of course, there are undoubtedly confounding features in this example beyond symbol grounding—perhaps most notably, the lack of any intrinsic motivation or activity. This argument obviously cannot be taken further than it will go, but it does seem to take us somewhere: what exactly is it that allows for the original development of language?

Taking up the case of sign languages again, we can note that the early stages of signed languages are highly imagistic: signs at first closely resemble that to which they refer [11]. This gives way to a process of 'conventionalisation,' of increasing abstraction, but this abstraction is itself achieved through the creative combination of experientially grounded structures rather than the invention of meaning ab nihilo. This is evident in English, at least, where there seem to be no words which—if taken all the way to their etymological roots—do not ultimately refer back to some more or less concrete situation.

Indeed, this is true only in ordinary language but even at the very edges of our most abstract sciences. Niels Bohr described the difficulty with quantum mechanics, for instance, in terms of our reaching the edges of the imaginable [12]. Taking the example of the wave-particle duality of light, we see that both sides of this paradox refer to simple and concrete phenomena: waves, particles. The problem here, however, is that we have no experience of anything, apart perhaps from light itself, which is at once wave and particle.[5] These symbols thus exist in a state of tension akin to binocular rivalry; that the

[5] The fact that this may be the sole instance of our concrete encounter with such a thing may explain the prominence of this metaphor for the expression of paradoxes more broadly.

whole seems to oscillate unstably between two alternatives. If we are to grasp light as a phenomenon without reference to mathematics then we must hold it by handles derived from experience, by symbols which have their ultimate grounding in this concreteness.

We have nevertheless acknowledged both the possibility and place, even the importance, of statistical learning in human language. Here we may suppose, however, that some sort of minimal ontology is acquired by way of morphological computation during development. Leaving aside the strict task of taxonomy, we can suggest some seemingly necessary structures as candidates for this. Most significant is that class which Lakoff and Johnson refer to as 'ontological metaphors,' of which the most prominent is the object [13]. Quine [14] noted the centrality of this concept—which is itself the structuring principle for our concept of a concept—in stating that "We are prone to talk and think of objects … for how else is there to talk?" Other examples include processes and containers, as well as more nebulous concepts such as causality and intention.

The idea, in short, is that morphological computation provides a minimal ontology which serves as the basis of human language; and furthermore, that experience is particularly suited to serving this role in that it is derived from the world to which we refer. This does not mean that it is perfect, as is evident from Bohr's comment, but only that it is natural—and perhaps most importantly for evolution's demands, highly efficient.

5. Conclusion

We began by considering the relative inefficiency of language learning in large language models compared to human children. This has been taken as a route to reconsidering the problem of symbol grounding in terms of the movement of meaning from world to model, which we have here argued takes place as a form of morphological computation whereby embodied activity and perceptual experience provide qualitative structures which scaffold language. We have further noted the place of statistical-type learning in human language use but have argued that morphological computation is yet the source of a minimal ontology which structures the whole of human language. Experience, in other words, is the ground from which it grows, however this material might be altered or repurposed afterwards. From this summary, then, we will turn now to considering the implications of this view—and particularly for the question of multi-modal robotics.

At the outset we noted that symbol grounding could be considered in terms of two aspects: the practical, from model to world, and the theoretical, from world to model. We have spoken here in terms of the theoretical, but really our measure has throughout retained a practical bent—that we have framed this discussion with reference to the relative efficiency of language learning in neural networks making use of the transformer architecture and human beings. This brings us to the position where we might readily imagine a form of artificially intelligent robotics which makes use of these features, wherein the concept of morphological computation so informs its design as to allow for equivalent efficiency to the human case. Such an instantiation may well provide proof in terms of our formalisation by way of conditional Kolmogorov complexity, and we have here noted some initial efforts in this direction with findings that conform to our thesis.

More broadly, however, this aim is akin to the stated goal of the field variously known as developmental or cognitive robotics, with its major proponents being Minoru Asada, Angelo Cangelosi, and Giorgio Metta [15]. The overarching motive in this work is that efforts in artificial intelligence ought to move in concert with an understanding of the developmental processes at work in humans. That humans should serve as an inspiration for efforts in artificial intelligence is, of course, hardly new—with the neural network itself plainly based, more or less loosely depending on the variant, upon an abstract model of the neuron, upon action potentials, and so on. The mode of learning, however, has remained distinct; although evidently this has not been a strict impediment.

The recent success of large language models, from this perspective, suggests that artificial intelligence may not require embodiment, and the question at hand is whether even the apparent difficulty of symbol grounding in the practical sense requires anything more than a mass of training data, compute, and the bolt-on addition of multi-modality and robotics. The state-of-the-art in present multi-modal robots, after all, is not naturally multi-modal or even robotic; rather these use linguistic intelligence as an interface to coordinate these modalities.[6] The question, then, is whether this is sufficient—with which we return again to the theoretical versus practical, albeit now at a different angle.

The ultimate question here must be the significance of symbol grounding: what is the purpose of work in this area? If it is a matter of alignment then we may soon count ourselves as having solved this problem on the whole. This is particularly true when we consider that, in many cases, perhaps most, the sort of symbol grounding through morphological computation which we have outlined here is likely little more than a matter of efficiency even for human language users. The need for this form of symbol grounding, moreover, may be significantly reduced in a society that has been abstracted and commodified, particularly with the rise of digital technologies and virtual realities.

There are yet cases at the very edges of language where something further may be necessary, where thought must dip once more into the well which was its origin, as for poetry and perhaps also in the case of paradigm shifts as described by Thomas Kuhn [16]—and yet this is hardly necessary for the general population. The requirement for direct symbol grounding seems most prominent in the earliest stages of language, both at a population and an individual level; beyond which we might rest satisfied rather with the practical alignment of meanings. This is supposing, of course, that the peculiar limitations of large language models—such as the ‘reversal curse’ [17] or the ‘blindness’ of vision language models [18]—are not associated, and here this may be a precarious assumption, with the absence of a prior integrative access to experiential grounding.

If, however, our interest is not purely practical, if we are interested in this also as a theoretical question, then we must consider the whole from a different perspective. Here these matters of minor importance must rather be our aim, with which we will surely encounter a whole raft of further difficulties. The notion of experience, for instance, has here been treated as common—for in some sense it is, at least for those in the audience of a biological origin—but it is hardly so when considered as a question of engineering. It is not clear whether multi-modality is at all equivalent to the unity of experience,

[6] The latest fundamental model from OpenAI, GPT-4o, is end-to-end multi-modal—but the specifics of this remain proprietary; it is likely akin to ImageBind (Girdhard et al., 2023) and Unified-IO 2 (Lu et al., 2023).

whether vision through pixel matrices does not differ fundamentally from its transduction in the retinal complex. If we are to evaluate our present prospects for the symbol grounding problem therefore, we must ask what this means to us and what our interests are, whether we want intelligence or something more.

References

[1] Brohan A, Brown N, Carbajal J, Chebotar Y, Chen X, Choromanski K, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818. 2023 Jul 28.

[2] Müller VC, Hoffmann M. What is morphological computation? On how the body contributes to cognition and control. Artificial life. 2017 Feb 1;23(1):1-24.

[3] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752. 2023 Dec 1.

[4] Ming LI, Vitányi PM. Kolmogorov complexity and its applications. In: Algorithms and complexity. Elsevier; 1990. p. 187-254.

[5] Mahmud M, Ray S. Transfer learning using Kolmogorov complexity: Basic theory and empirical evaluations. In: Advances in Neural Information Processing Systems 20 (NIPS 2007). 2007.

[6] Frank MC. Bridging the data gap between children and large language models. Trends in Cognitive Sciences. 2023 Nov;27(11):990-992.

[7] Wang W, Vong WK, Kim N, Lake BM. Finding Structure in One Child's Linguistic Experience. Cognitive science. 2023 Jun;47(6):e13305.

[8] Zhuang C, Fedorenko E, Andreas J. Visual grounding helps learn word meanings in low-data regimes. arXiv preprint arXiv:2310.13257. 2023 Oct 20.

[9] Vyshedskiy A, Shreyas M, Dunn R. Linguistically deprived children: Meta-analysis of published research underlines the importance of early syntactic language use for normal brain development. bioRxiv. 2017 Jul 21:166538.

[10] Kegl J, Senghas A, Coppola M. Creation through contact: Sign language emergence and sign language change in Nicaragua. In: DeGraff M, editor. Language creation and language change: Creolization, diachrony, and development. Cambridge (MA): MIT Press; 1999.

[11] Corballis MC. The recursive mind: The origins of human language, thought, and civilization. Princeton (NJ): Princeton University Press; 2011 Jun.

[12] Bohr N. On the notions of causality and complementarity. Science. 1950 Jan 20;111(2873):51-4.

[13] Lakoff G, Johnson M. Metaphors we live by. Chicago: University of Chicago Press; 1980.

[14] Quine WV. Ontological Relativity and Other Essays. New York: Columbia University Press; 1969.

[15] Cangelosi A, Asada M, editors. Cognitive Robotics. Cambridge (MA): The MIT Press; 2022.

[16] Kuhn TS. The structure of scientific revolutions. Chicago: University of Chicago press; 1997.

[17] Berglund L, Tong M, Kaufmann M, Balesni M, Stickland AC, Korbak T, et al. The reversal curse: Llms trained on "a is b" fail to learn "b is a". arXiv preprint arXiv:2309.12288. 2023 Sep 21.

[18] Rahmanzadehgervi P, Bolton L, Taesiri MR, Nguyen AT. Vision language models are blind. arXiv preprint arXiv:2407.06581. 2024 Jul 9.