Content from Expressions

Juan Luis Gastaldi

Content from Expressions

The Place of Textuality in Deep Learning Approaches to Mathematics

Juan Luis Gastaldi

Abstract

Recent years have seen a remarkable development of deep neural network techniques for data analysis, along with their increasing application in scientific research across different disciplines. The field of mathematics has not been exempted from this general trend. The present paper proposes a philosophical assessment of the epistemological claims and conditions of such attempts. After a quick survey of recent applications of neural models to mathematical knowledge, we address the philosophical significance of those results, focusing on the specific problem of mathematical textuality and the somewhat surprising circumstance that semantic aspects of mathematical knowledge can be inferred from pure syntax. We then analyze the renewed role of distributionalism in neural models and propose an alternative understanding of its relation to meaning. Finally, we present an illustration, based on empirical evidence, of how aspects of arithmetical content such as recursive structure and total order could be inferred through explicit and interpretable means from the distributional properties of a natural language corpus.

Keywords Philosophy of Mathematics · Philosophy of Language · Machine Learning · Natural Language Processing · Deep Learning · Artificial Intelligence · Benford's Law

1 Introduction

In the last two decades, the field of machine learning has experienced a critical revolution due to the remarkable development of Deep Neural Network models (DNNs) together with an ever-increasing availability of digital data. Much of these developments are driven by the economic interests of the leading tech companies, who remain the principal actors in a research field where computational infrastructure and data extraction capacity can become prohibitive

This document is a draft. Please contact the author before sharing or citing.
This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 839730.
File version: 0.0.2023.02.17.

requirements. As a consequence, the main orientations of the field appear inseparable from their application to market products and the increase of economic value. However, along with this evolution, the past years have also witnessed the application of the new neural methods to different aspects of scientific research across the disciplinary spectrum, from natural to social sciences.

Somewhat surprisingly, mathematics has not been the exception to this general trend. Indeed, since the mid-2010s, several efforts have been dedicated to treating different aspects of mathematical knowledge using DNNs of various kinds. From a philosophical perspective, these emerging machine learning approaches to mathematics may very well come as a surprise. For, even more than in any other scientific area, the application of neural machine learning models to mathematical knowledge raises critical epistemological questions due to the formal—i.e., non-empirical—nature generally attributed to mathematics, which contrasts with the strong empirical position assumed by connectionist approaches guiding the application of DNNs.

In the current state of the art, it is not possible to determine if current machine learning methods will end up having a significant impact on future mathematical practices. However, or rather because of that, the current situation represents a most exciting moment in the eyes of the historian and philosopher of sciences. A moment where established certitudes can and need to be regarded afresh to assess if they hold still in a context that has suddenly changed.

If such moments are interesting to philosophers, it is also the philosophical inquiry that becomes interesting, if not altogether indispensable, to critically characterize the stakes of the new situation and propose novel orientations. The present paper intends to represent a modest contribution to this critical task. The new context is that of the regained momentum of the articulation between empirical approaches, experimental methods, and computational means and its effects on their joint generalized epistemological claims. As for the certitudes in question, one of the primary purposes of these pages is to argue that this forces us to revise our conception of the relation mathematics maintains with its own textuality.

Our task will be to call attention to what we think is philosophically significant in the ongoing research on machine learning approaches to mathematics, proposing a specific way of constructing the problem and suggesting understanding principles through a concrete illustration. Accordingly, the plan of the paper is as follows. After a brief presentation of neural network models and a quick survey of recent applications to mathematical knowledge (Section 2), we will address the philosophical significance of those results, focusing on the specific question of mathematical textuality and on the somewhat surprising circumstance that semantic aspects of mathematical knowledge can be inferred from pure syntax (Section 3). We will then inquire into the conditions of possibility of that phenomenon, stressing the renewed role of distributionalism in neural models (Section 4) and proposing an alternative understanding of the relation of the latter to meaning (Section 5). Finally, we will present an illustration, based on empirical evidence, of how aspects of arithmetical content could be inferred through explicit and interpretable means from strictly distributional properties of a natural language corpus (Section 6). We will conclude with a summary and some final remarks.

2 Deep Mathematics

2.1 Deep Neural Networks

In the past few years, the popularity of Artificial Intelligence (AI) driven by Deep Learning techniques has become so overwhelming that it is hardly necessary to present the elementary principles of DNNs. This short introduction is, then, intended only as a way to sketch the essential features of DNNs minimally required for the aims of this paper. For a detailed and systematic presentation, we refer the reader to one of the many available textbooks, such as Goodfellow et al. (2016) or Brunton and Kutz (2022).

From a bird's-eye view, artificial neural networks are devices for transforming vectors. Therefore, their application to any field relies on the possibility of encoding some information as a vector (i.e., a list of numbers), feeding it into the device (or the "model"), and retrieving another vector to be decoded as some other kind of information. The nature of the information is assumed to be of little significance as long as it can be encoded as a vector and a fairly extensive set of examples of pairs of input and output vectors is available to train the model. Indeed, neural network architectures have been proved to be a universal function approximator (Hornik et al., 1989). Therefore, given enough examples of input and output vectors, a neural net will approach the function transforming one into the other, if any.[1]

In the most classical cases, the internal structure of the model is composed of a series of transformations, each having a similar configuration: a linear transformation (i.e., a matrix), a bias (an added vector), and an activation (non-linear) function. Each one of these complex transformations is called a "layer". Hence, when a vector is fed into a layer, it is multiplied by the matrix, the bias vector is added, and the non-linear function is applied element-wise. The "deep" character of neural nets is determined by the composition of many of these layers, so that the output of one layer becomes the input of the next one until a final output vector is produced.

Finally, the model is trained by iteratively updating the components of the matrices (or "weights") and bias vectors of the different layers. This is achieved by performing a gradient descent on a loss function, that is, on a function that measures the error between the output vector and the desired vector. Through an algorithm known as backpropagation, the weights and biases of different layers are progressively adjusted so that the average error over the set of examples hopefully converges to some minimum value.

This scheme corresponds to the most elementary neural models. In the past years, DNNs have become increasingly complex and diversified, from convolutional neural nets, through recurrent neural nets to Transformers, to name only the most popular architectures. Moreover, the training strategies may also vary between supervised, unsupervised, self-supervised, or reinforcement learning. However, unless explicitly mentioned, the details of this wide and evolving

[1] This does not mean that any task is potentially solvable by means of neural models, as a significant part of the AI community tends to think. Many problems, maybe most of them, do not accept to be constructed as a predictive task based on a finite number of supposedly similar cases expressible in the form of vectors. Problems in politics, law, art, and many other domains of social life cannot, by their nature, be framed in such terms without transforming, if not completely erasing that very nature.

variety of models will not be relevant here. What is important is that all of those models remain particularly sophisticated cases of the basic scheme just presented.

2.2 Neural Models for Mathematics

Given the exploratory character of the task in a field that moves with a remarkable speed, there are multiple ways to apply neural models to mathematical knowledge. New results are released every month, making it impossible to provide a complete account of the state of the art. For the purposes of this paper, it is enough to present what we believe to be the main trends recognizable among all these works, pointing to a representative example in each case, with no pretension of exhaustivity.[2] We propose to organize such trends broadly into four categories, determined by whether the research is oriented toward *proofs*, *objects*, *skills*, or *heuristic*.

2.2.1 Proof-Oriented

The first group of works follows the line of what can be taken as the pioneering attempts to apply well-established DNNs to the solution of mathematical tasks, namely that of Alemi et al. (2016). In this approach, theorems are assumed to be the key to mathematics. Accordingly, the aim is to train neural models to find *proofs*. Ideally, one would train the model on an extensive corpus of existing proofs, after which one could feed a yet unproved mathematical statement to the model, which would output a proof, if any. Notice that this is not the same as building a formal logical framework for automatic theorem proving because there is no construction of an explicit logical system here. Eventually, one could imagine that logical systems are just a specific kind of function, even if a partial one, mapping the domain of statements with that of proofs, and we are asking the DNN model to approximate such a function.

In practice, however, things are more subtle. Existing research in this direction focuses on specific subtasks identified within the framework of already existing automated theorem provers (ATP), such as E or Lean. The focus falls mainly on those tasks which appear as computational bottlenecks, in particular, premise selection. Other tasks include predicting the importance or the name of a mathematical statement, the original conjecture for intermediate statements, the same tactics, goals, and subgoals used by humans, or generating intermediate statements for given proofs (cf. Kaliszyk et al., 2017; Bansal et al., 2019).

As a representative example of this line of work, we can mention the original paper of Alemi et al. (2016) on premise selection, that is, "the selection of a limited number of most relevant facts for proving a new conjecture" (p. 1). The strategy is to feed both a conjecture and an axiom into the network, process the input vectors independently through different network architectures producing intermediate vector representations of different kinds, concatenate those vectors, pass them to a fully connected DNN and output a probability measure of the usefulness of the axiom for proving the conjecture. The representation

[2] A more comprehensive account of the state of the art is currently under preparation, to appear in the projected *Handbook of the History and Philosophy of Mathematical Practice* (ed. Bharath Sriraman, Springer Verlag).

of axioms and conjectures at the input is then that of the first-order logic formalization extracted from the Mizar Mathematical Library. The authors tested two different ways of feeding these texts into the model: character by character and word by word (token by token). In the latter case, tokens recognized as identifiers were mapped onto their already processed definition. Other than that, the expressions are fed sequentially with no parsing information. For each conjecture, the model produces a probability distribution over the set of possible premises. To evaluate the model, the authors count how many conjectures can be proved from the top-k of their corresponding probability distributions (for different values of k) using the E automatic prover. They also evaluate the ranking quality, measuring if relevant premises appear at the top of the rank. Such measures are then used to compare the different DNNs architectures.

This work sparked a series of other works in the same direction where, among others, the use of DNNs was shown to increase—rather slightly—the performance of existing ATPs by being integrated into that existing framework. Other corpora were used, and comparisons were made between different neural architectures as to which kinds of theorems each model was good at. The range of tasks was also extended, including the automatic formalization of natural language mathematical texts (cf. Welleck et al., 2021).

As with all cases of mathematical applications, this line of research remains exploratory. However, the results are far from wholly unreasonable or insignificant and continue to improve, suggesting that DNNs are capable of manipulating mathematical proofs to some extent.

2.2.2 Object-Oriented

Another line of research recognizable in the field shares similar strategy, tools, and methods with the first one. However, the focus is placed on manipulating mathematical *objects* rather than statements. Admittedly, proofs can also be considered as mathematical objects, in which case proof-oriented approaches can also fall under this category. At any rate, DNNs are, in principle, indifferent to this distinction. Yet, from a philosophical standpoint, there is a significant difference between, for instance, manipulating numbers to solve arithmetical operations and proving an arithmetical theorem.

In line with the general framework of neural models, the objective, in this case, is to feed the expression of a mathematical object (e.g., equation, terms of a series, etc.) into the model and expect the latter to output some specific result or property about it (e.g., solution, recurrence relation, etc.). Following d'Ascoli et al. (2022), this can be pursued in four different ways, depending on whether either the inputs or the expected outputs are symbolic expressions or numerical data.

The solution of partial differential equations has been a privileged area for the development of this line of research. A notable example and original landmark in this sense is the paper by Lample and Charton (2019), testing the DNNs' capabilities in solving differential equations and symbolic integration. The overall strategy is the same: feeding a first or second-order differential equation as input and asking the model to output its solution. However, unlike the previous case, the authors trained their model on an artificial corpus specially designed for the task. The representation of the input and output was thus well-thought to fit the kind of data on which the chosen DNN architecture

is known to perform well.

Relying on a careful handcrafting of the mathematical object's representations, the model exhibited remarkably high results. Under very precise conditions, the authors showed that their model outperformed Mathematica, Matlab, and Maple for the tasks in question. Like in the previous line of research, many results were developed after this seminal paper, either focusing on the application of DNNs to PDEs (**?**Blechschmidt and Ernst, 2021) or extending the approach to other mathematical objects, such as numerical series (Ryskina and Knight, 2021; d'Ascoli et al., 2022) or linear algebra (Charton, 2021).

2.2.3 Skill-Oriented

A third group of works applying neural methods to mathematical knowledge follows a different strategy compared to the previous two. The idea here is not so much to try to solve specific mathematical tasks but to endow the neural model with general mathematical *skills* capable of spontaneously solving mathematically related problems.

This hope is motivated by a relatively recent revolution in the field of AI, brought about by a new kind of neural architecture called the "Transformer", implementing a specific neural technique known as "attention" (Vaswani et al., 2017). This architecture became popularized through a particular implementation called BERT (Devlin et al., 2018) and achieved public celebrity with the surprising results exhibited by the release of GPT-3, a Transformer language model of 175 billion parameters trained on a massive amount of data (Brown et al., 2020).

This family of models is characterized by the fact of being trained as language models in a highly generic fashion, namely by trying to predict random masked words or tokens in sentences. One of the most surprising features, usually known as "in-context learning", is that, after training it on a vast amount of generic data, one can prompt a few input-output examples of some specific task expressed in natural language, and the model can continue performing that task with surprising accuracy. This behavior was shown to hold for very different kinds of tasks for which the generic model was never explicitly trained. As a consequence, a new training strategy became widespread, consisting of "pre-training" a generic model on a generic corpus of natural language and then continuing the training ("fine-tuning") on a smaller scale on a specialized task or corpus. As pointed out in the comprehensive report of Bommasani et al. (2021), the success of these models (a.k.a. "Large Language Models" (LLMs)) is based on two pillars, namely, the capacity to transfer the content learned between different tasks and the scaling capabilities.

In this framework, it seems natural to prompt the generic model with mathematical tasks. The original GPT-3 paper already showed the arithmetical skills of such a generic model, in particular, by trying to perform up to five-digit elementary arithmetic. The results were rather surprising given that the model was never trained on such a task: the model showed to be almost 100% accurate on 2-digit addition and subtraction and over 80% for three digits (Brown et al., 2020, §3.9.1). These results have been improved in more recent models, and further investigation showed that the generalization capabilities strongly depend on the textual representation of numbers (Nogueira et al., 2020).

Following this new general strategy, significant work has been developed in

the past 2 or 3 years to test the mathematical skills of generic models and fine-tune them for general mathematical knowledge upon what is assumed to be a generic mathematical corpus. In 2021 two models were presented in this sense with the same name, MathBERT (Peng et al., 2021; Shen et al., 2021), which shows the kind of enthusiasm, if not the hype, that this line of research generates.

2.2.4 Heuristic-Oriented

Finally, it is important to mention what looks like a different approach to employing DNN methods in mathematics, namely that of Davies et al. (2021). In this recent paper, the authors rely on different kinds of DNNs to "guide the intuition" of human mathematicians in searching for proofs for existing conjectures. More precisely, machine learning methods are here mobilized to help understand or verify the existence of specific structures or patterns in mathematical objects. The authors show the fruitfulness of this approach in the case of two open problems concerning the structure of knots and the combinatorial invariance of symmetric groups in representation theory.

The relevance of neural methods in this case has been put into question (Davis, 2021), for it is not clear what is the specific value added by DNNs compared to more classical statistical methods already in use by mathematicians in their own practice. However, the idea that DNNs can be used as a *heuristic* tool in mathematical practice seems promising, and it is plausible that more substantial work in this direction will emerge in the near future.

3 The Philosophical Significance of Machine Learning Approaches to Mathematics

3.1 Mathematics As Text

In most of the cases presented in the previous section, the results exhibited by applying current neural models to various aspects of mathematical knowledge concern only elementary mathematical properties. However, the simple fact that those properties can be addressed, if only with relative success, from the radically empirical perspective assumed by current machine learning approaches should be enough to raise a whole series of philosophical questions. In the face of that evidence, one could ask, for instance, what is it, exactly, that these models model or try to model? Is it the manipulation of abstract objects? Of operations? Is it the acquisition of an existing piece of knowledge? Or the human capacity to produce new knowledge? Or, yet, the machine's capability of doing so? More profoundly, is the knowledge attributable to those models *a priori*? And if so, why and how can an empirical learning procedure such as this grasp it? Or is that knowledge empirical and, therefore, not really mathematical? Or maybe mathematical knowledge *is* empirical after all? But then, what distinguishes mathematics from other kinds of empirical knowledge? Under which conditions does a dataset constitute a mathematical corpus? Would corpora from different cultures or different epochs induce different mathematical results? What counts as a mathematical result? And what kind of evidence

can these models provide for the truth, correctness, or even meaningfulness of those results? What is a proof…?

The list of conceptual, epistemological or philosophical questions could be extended at will. And yet, the fruitful encounter between the philosophy of mathematics and current machine learning practices has not yet taken place. On the one hand, one cannot but be surprised about the practically complete absence of epistemological or conceptual reflection—let alone cultural or historical awareness—about the nature of mathematical objects, properties, and knowledge in most, if not all, the literature in question. This absence is all the more surprising that one can imagine multiple ways in which the conceptual clarifications potentially brought about by those reflections could contribute to the tasks under study, especially for an approach that so defies the received analytic view on mathematics, rooted in a purely logical conception. In many cases, the room left by the absence of conceptual reflection is filled, more or less implicitly, by a cognitive perspective where the models in question are assumed to acquire, in their own way, the human cognitive abilities also assumed to be the origin and support of mathematical knowledge. In line with this attitude, the analogy between the computational models and the human brain is frequently present in the field, even if not seriously (i.e., scientifically) addressed. It is hardly surprising, on the other hand, if this lack of conceptual elaboration and the focus on technical details and incremental performance of otherwise non-interpretable models can lead philosophers to more or less silently believe there is nothing of philosophical interest in this approach, which seems reduced to a mere engineering practice.

However, there are many reasons for the philosophy of mathematics to turn its attention to the recent mathematical applications of machine learning, even if, or precisely because those reasons do not coincide with the ones put forward by the main orientations in that field. Starting with the fact that, were this kind of approach to succeed to the point of becoming widely used, the face of mathematics as we know it would certainly change in ways that are not trivial from a philosophical standpoint. Such a modification of mathematical practices would significantly impact the many aspects philosophers have become used to associating with them, such as the role of proofs (Chemla, 2012), the meaning of or even the need for foundations (Wagner, 2019), the nature of explanations (Mancosu, 2008) or the place of experimental procedures (Borwein and Bailey, 2003; Avigad, 2008).

However, the following pages would like to point in a different direction, focusing on *the relation between mathematics and textuality*. Because the success of these methods—be it actual or promised—implies a whole new status for mathematical texts, which goes far beyond the customary understanding of mathematical writing as a simple notation for pre-existing mathematical content or a more or less arbitrary syntax for an independently determined semantics.

The most immediate argument supporting this claim is the basic observation that *texts* (statements, expressions, symbols, characters, etc.) *are all these models can rely on* to perform the mathematical tasks they are asked to perform. Whatever happens within the black box of a neural net happens on the basis of the processing of syntax and *syntax alone*. Admittedly, DNNs are not in themselves purely syntactic entities but the implementation of a reasonably well-specified semantic structure. However, such a structure is highly generic, to the point that it can be successfully applied to very different kinds of data

(images, language, sound, etc.) to perform significantly different kinds of tasks. Given their generic character, if neural models prove to be capable of dealing with mathematical content, such a content is nowhere else to be found than in the corpus of (eventually supervised) textual expressions fed as inputs.

It is essential to understand that *this is not a bug but a feature* of this approach. Indeed, there are multiple ways to enhance neural models with handmade rules and mathematical principles to increase their performance over specific tasks. As soon as accuracy becomes the primary concern, researchers do not hesitate to make heavy use of all sorts of tweaks and tricks, introducing semantic properties by other means. The resulting models progressively become, then, the object of an engineering practice, losing, indeed, much of the philosophical interest concerning the syntax-semantics relation. Nevertheless, the connectionist credo behind DNNs exhorts to develop learning models as generic with respect to tasks as sparing with respect to assumptions about the data. Consequently, whether openly acknowledged or unwittingly, current applications of DNNs to mathematical knowledge grant a critical place to mathematical language and textuality, providing the conditions for an original reflection on its role in the production of mathematical content.

3.2 The Natural Language of Mathematics

Two principal aspects concerning mathematical language stand out for their potential originality in this setting. The first is a consequence of the fact that bare syntactic objects are all these models can actually resort to for processing mathematical knowledge. Thus, without dedicated axiom systems, deductive rules, logical operations, symbolic reasoning, material manipulation, or access to any other kind of context than textual, most of the conceptual focus within this line of research is put on mathematical *representations*. Be it logical statements, formulas, equations, expressions, or symbols,[3] researchers in machine learning have been led to raise the question of textual mathematical representations in ways that can communicate with current research interests in the philosophy of mathematics. For, if something becomes apparent in the vast majority of the works in question, it is how much content is implied in mathematical representations and hence, how much conceptual attention they deserve.

Alemi et al. (2016), for instance, rely on the syntax of proof assistants (Mizar) to represent logical statements while evaluating the efficacy of their model by analyzing it at both character and word levels and elaborating strategies to associate symbols to explicit and implicit definitions. This line of work could thus provide novel perspectives for the design of mathematical languages, as addressed, for instance, by Avigad (2015, 2021). Lample and Charton (2019), in turn, propose encoding the expression of differential equations as binary trees, where the internal nodes represent operators while the leaves feature numbers, constants, or variables. The trees are then re-encoded as sequences following the normal Polish notation to avoid ambiguity and address issues of precedence and associativity while conforming at the same time to the sequential form required for the vector encoding of DNN inputs. Tree structures for mathematical

[3]To the best of our knowledge, there has been no work yet involving the content of mathematical diagrams within this trend. From a different perspective, Sørensen and Johansen (2020) have proposed the use of neural techniques to identify diagrams in mathematical texts.

expressions have also been explored in other neural applications, such as Nangia and Bowman (2018); Zou and Lu (2019); Peng et al. (2021), to name but a few, and the question of representations for mathematical expressions within neural models more generally has been the object of explicit attention (for instance, in Mansouri et al., 2019; Nogueira et al., 2020; Ferreira and Freitas, 2021; Purgał et al., 2021), including thorough studies of mathematical notation (Greiner-Petter et al., 2020a). All of which provides the elements for a fruitful dialogue with current investigations on the role of representations in the philosophy and history of mathematical practices, such as Schlimm (2018); Waszek (2018); Kohlhase et al. (2018). In particular, DNN applications to mathematics have the potential to introduce original perspectives, since their objective is not to find the best representation for a given content (a suitable "notation") but the best representation for given texts so that tasks assumed to rely on their unknown content can be correctly performed by a statistical learning model such as neural nets.

The second remarkable aspect concerning mathematical language is related to the connection between mathematics and *natural language*. Indeed, addressed as an artificial language from a modern logical perspective, mathematical language has been longtime kept apart from natural language by different traditions in the philosophy of mathematics. Departing from this orientation, neural machine learning models are led to operate novel articulations between both, which are far from philosophically trivial. On a superficial level, one can find that, since many of the corpora used are composed of existing published papers (especially in the case of LLMs such as Transformers), the natural language necessarily present in those papers is usually leveraged to contribute to determining the content of mathematical expressions. A conceptual investigation of the attested interaction between both natural and mathematical expressions is likely to provide compelling insight into the mechanisms of such interaction in different corpora, be it from a historical perspective, in line with works such as those of Netz (1999) or Herreman (2000), or from a more cognitive perspective (Giaquinto, 2008; Toffoli and Giardino, 2013; Kohlhase et al., 2018).

More substantially, the vast majority of the learning models used for processing mathematical expressions are none other than those specifically developed for the processing of natural language during the past decade. This means that, within this setting, *mathematical expressions are treated as being themselves a sort of natural language*. Such a position is implicit in all the approaches using word embedding techniques or Transformer models and is even explicitly stated by Lample and Charton (2019). Should the application of natural language models to the processing of mathematical content succeed, even if only in specific cases and on a small scale, we must be careful not to reduce this event to a mere technical feat. Contrary to a view making a drastic distinction between natural and artificial languages, such a success would suggest, otherwise, a close link between both, either because they both share specific properties, because the treatment of natural language constitutes a general framework for studying all kinds of languages, or yet because deep neural models constitute a general framework for the analysis of data, under which both natural language and mathematics fall.

4 The Content of Expressions

4.1 Challenges of a Radical Text-Driven Approach to Mathematics

In this section, we will focus on what constitutes the most radical originality of neural approaches to mathematics, revealing the challenges it represents.

Such an originality lies, as already advanced, in that *mathematical content is derived from expressions alone, semantics from syntax, meaning from pure text.* Yet, in modern mathematics, the access to content is supposed to be provided through explicit and rigorous symbolic mechanisms, such as definitions within a formal language, logical operations, deductive rules, systems of axioms, or formal models. Such symbolic procedures are assumed not only to determine what those expressions mean but also to control how expressions interact and what results from that interaction. Without direct access to those explicit symbolic means—let alone cognitive faculties or social conditions—it is difficult to know how the semantics or the content of expressions can be determined.

The difficulty is even greater if one considers that the content of mathematical expressions DNNs target and to some extent manipulate is far from limited to *referential* aspects. Indeed, the solution of the corresponding mathematical tasks requires much more than associating, for instance, the expression `406` to a particular number or quantity, or to the same entity as the expression `four hundred six`; it goes beyond referring the expression $\mathtt{A} \wedge \mathtt{B}$ to a logical conjunction or $\mathtt{y}'' - \mathtt{y} = 0$ to a differential equation of second order.[4] The tasks in question require, in addition, that enough *operational* content is involved to determine that `406` added to `326` equals `732`, that $\mathtt{A} \wedge \mathtt{B}$ is likely or unlikely to be a premise in the proof of some given logical statement, or that $\mathtt{y}(\mathtt{x}) = \mathtt{c}_1\mathtt{e}^{\mathtt{x}} + \mathtt{c}_2\mathtt{e}^{-\mathtt{x}}$ is the solution to that differential equation.

Resorting to how content is dealt with in natural language, as most models do, might sound promising, because the content of expressions in this case does not depend on explicit rules and other artificial procedures absent in neural models, but tends to rely on more "natural" mechanisms. The latter could then be invoked to think about similar procedures at work in the relation between mathematical expressions and their content, explaining the surprising capabilities of neural nets. However, as natural as it may be, if we restrict ourselves to pure expressions, the access to content in natural language is far from being simple and direct. The meaning of words is usually given through other words, risking and infinite regression, and ostensive definitions are doomed to inescapable ambiguities unless large portions of content are already known, as the radical critiques of Wittgenstein (2009) and Quine (2013), among others, have sufficiently shown.

What is more, the recourse to natural language imposes significant constrains on the treatment of expressions. Indeed, DNNs can only achieve their results by addressing the problem as a predictive task. Thus, given the expressions `406` and `326` as inputs, the model is required to predict the expression `732`, after being trained over a huge quantity of similar cases (but not this specific one). Or more precisely, it is expected to predict the expression `732` from any expression containing `406`, `326` and some expression of the additive operation,

[4] We adopt the convention of writing raw text or pure expressions in a `monospace` font.

as in `406+326=` (or some sequential encoding of it). Now, by approaching the task as if we were dealing with natural language, we require neural models—and sometimes even the same model—to perform those predictions by the same means by which they can predict, say, `you` from `she asked` when generating the expression `she asked you`, however differently structured the two kinds of languages and contents are supposed to be.

If the undertaking seems tortuous, analyzing those means is otherwise challenging. As it is widely acknowledged, DNN models are practically uninterpretable (see for instance Lipton, 2018). In particular, we have no access to formal representations of a model other than its internal state, which in the current state of the art can go from hundreds of billions up to over a trillion parameters. Different methods have been developed in recent years, with the aim of providing tools for the analysis and principles of interpretability of natural language processing (NLP) neural models (see Belinkov and Glass (2019) and Madsen et al. (2021) for a survey). Among them, the method known as *probing* (Conneau et al., 2018) enjoys a certain success. The idea is to use the encoded vector representations of a model to train a classifier over a specific linguistic property considered relevant for performing linguistic tasks (eg., grammatical dependencies). If the classifier achieves a good performance, the properties in questions can be considered to be somewhat encoded in the internal state of the original neural model. However, while several mechanisms have been proposed to show that the information corresponding to the probed property is legitimately attributable to and effectively used by the model (Hewitt and Liang, 2019; Ravfogel et al., 2021), there still remains a fundamental gap between neural models and the symbolic property or structure being probed. In other words, high—and ultimately superhuman—performance on such predictive tasks, which can indeed be conceived to grasp significant aspects of the content of expressions, might not reach that accuracy through the same means through which we are used to do it ourselves, or would expect to do it. Neural models might be able to write sensible natural language or solve natural language tasks without ever following grammatical rules for any acceptable sense of grammatical rule-following. Likewise, those models might be able to solve mathematical tasks with reasonable and even surprising accuracy without ever manipulating any of the symbolic procedures by which we have historically controlled our mathematical practices. Converging results do not imply identical theories, representations, knowledge or capabilities, which in turn does not imply that content is not being grasped.

4.2 How Is It Even Possible? The Feats of Distributionalism

At the face of all these obstacles, by no means unknown in the field, why to place any hope in contemporary machine learning approaches to mathematical knowledge at all? Is it not the whole endeavor simply ill-conceived?

If the proof of the pudding is in the eating, it seems fair to acknowledge that, while the results exhibited so far are certainly modest from a mathematical point of view, they are far from insignificant from a philosophical one. Given the difficulties just mentioned, the slightest evidence that a seemingly impossible task can be achieved deserves philosophical consideration. Moreover, the models in question have shown surprising capacities in the treatment of natural

language, where the richness of content makes the task otherwise challenging. Indeed, despite the many limitations duly identified for these models, the results exhibited by the application of DNNs to NLP in the last decade have been enough to catalyze, if not altogether generate, a revolution in the field (cf., for instance, Manning, 2015).

Due to the close relation between mathematical texts and natural language within this framework, it seems legitimate, then, to assume that the hope placed in neural methods for the treatment of mathematical language is grounded on the mechanisms underlying the success of neural language models, whatever they may be.[5] Accordingly, a philosophical analysis of this phenomenon should focus on the elementary mechanisms responsible for the success of fairly generic learning algorithms in manipulating natural language content when applied to a practically raw corpus of texts, and on the reasons and conditions for those same mechanisms to apply to the treatment of mathematical content.

To avoid resorting to unverifiable analogies between artificial neural models and human learning capacities, let alone structural properties of the human brain, it is imperative to focus on what those models *actually do*. As we have seen, DNNs are statistical models resulting from a learning procedure based on predictive tasks. If such models are capable of grasping linguistic content when operating with raw text, it is then natural to ask what can the source of that content be.

Since the expressions fed into the model are nothing but raw text, i.e. a sequence of empty identifiers, and models are highly generic before training, the source of all content cannot be other than the training corpus. Yet, from the model's viewpoint the corpus is also nothing but a sequence of identifiers, which are entirely arbitrary by definition. Therefore, all the contents neural models can handle, all the manipulations inconceivable without taking content into account that they can make, have no other source than the relations purely arbitrary units maintain with each other in a given corpus. Whatever those units may be or stand for, be it characters, symbols, words, phrases or even bytes or pixels, all neural models can do is leverage the information of their mutual relations in specific corpora, eventually replacing their arbitrary identity with a relational one.

The idea that the content of linguistic units is determined by their relation to other units is not new, and can be traced back at least to the structuralist linguistics of Saussure (1959) at the turn of the 20th century. In the context of neural NLP models that idea has been embraced under a specific form known as the "distributional hypothesis", namely the assumption that words that appear

[5] The symbolic character of mathematical practices cannot, in this sense, be an objection *a priori* to this approach. On the one hand, it is not excluded that neural models can implicitly manipulate some sort of higher order representation akin to symbolic properties or structures (cf., for instance, Manning et al., 2020). On the other, symbolic practices in mathematics as we know them are a relatively recent phenomena. Although mathematical practices across different historical periods and cultures have constantly used expressive means irreducible to those of natural language, the recourse to a foundational role of symbolic practices such as axiomatic methods, formal systems of inference or model-theoretical semantics have only been established in the course of the past century. Mathematical knowledge has very well been produced and evolved for millennia across different cultures without such foundations (cf., for instance, Wagner, 2019) and would arguably continue to do so if alternative principles came to better fulfil the tasks expected from it in existing cultures, even if that requires to re-inspect our existing notion of semantics, content or even understanding, both in natural language and in mathematics.

in the same context tend to have the similar meaning (Sahlgren, 2008; Lenci, 2008; Gastaldi, 2020). This specific form of structuralism also has relatively old roots in the distributionalism of authors like Harris (1960) and Firth (1957). However, despite this longstanding tradition, neural models have introduced a new computational interpretation of distributionalism through the use of word vector representations, or *word embeddings*.

The story goes back to the 2010s, when researchers in the field realized that the first layer of neural models for NLP had a special significance. Indeed, they discovered that if one extracts the first layer of a DNN trained for a specific linguistic task and inserts it as the first layer of another DNN aimed at a different linguistic task, the performance of the latter is shown to improve. Accordingly, the idea emerged of training that first transferable layer as a separate shallow (i.e., not deep) neural network, independently of any specific downstream task. In this way, this new independent model was able to associate a low dimensional dense vector to each word in the vocabulary which could then be used as a generic representation of that word to be fed as inputs into other models.

Significantly, that separate model was trained on the generic task of predicting a word from its context words within a corpus (or vice versa). So the resulting word vector representations, or "embeddings", can be understood as a way of identifying and representing a linguistic unit exclusively through the information of its linguistic (textual) context, encoded as a low dimensional dense vector. As a consequence, through the use of embeddings, the purely arbitrary identity of linguistic units in a neural framework is replaced with an implicit encoding of all the textual contexts in which those units appear in a given training corpus.

As representations of words, embeddings turned out to be endowed with surprising properties. Not only they contributed to increasing the performance of most NLP tasks, but they exhibited, by themselves, significant semantic capacities of different kinds. Starting with the fact that the (typically cosine) distance between the vectors in the embedding space appeared to encode semantic similarity between the corresponding linguistic units. Even more surprisingly, researchers showed that one can identify analogical relationships between pairs of words (such as `prince` : `princess` :: `landlord` : `landlady`), suggesting that the embedding space is structured into subspaces following relevant semantic dimensions.

Since the development of the first embedding models around 2013, such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), word vector representations have undergone several evolutions significantly enhancing their capacity to extract semantic properties from raw text. The evolution was guided by the need to produce contextualized vector representations (i.e., different embeddings for each occurrence of a word in different contexts) over corpora of ever-increasing sizes. This is precisely what Transformer models like BERT achieve, determining the current state of the art in NLP.

The technical details of neural embedding models fall outside the scope of this paper. One can consult (Pilehvar and Camacho-Collados, 2020) for a comprehensive presentation. What is important for us is that word vector representations constitute the key by which the multiple relations of arbitrary units in a corpus are leveraged to capture all neural models can capture of the content of those units.

4.3 Mathematical Embeddings

We can now return to our previous question, and affirm that if any hope is laid on the successful application of neural models to mathematical language, that hope is and cannot but be explicitly or implicitly based on those models' capacity to recover semantic features out of corpora of relevant expressions, relying on the novel incarnation of distributional semantics that are word vector representations. In other words, *it is essentially through vector representations that, within a neural framework, expressions are invested with content and syntax is turned into semantics*. It is, then, on vector representations for mathematical expressions and on the mechanisms connecting them (or not) to mathematical content that a philosophical inquiry about the epistemological basis of neural machine learning approaches to mathematics should focus.

In recent years, several works have been interested in different ways of computing embedding representations for mathematical expressions. Gao et al. (2017), for instance, proposed an early method to produce embeddings for both mathematical symbols and formulas, and showed their relevance for mathematical information retrieval, while acknowledging the limitations to account for mathematical contexts. Krstovski and Blei (2018) showed that one can improve usual mathematical embeddings by considering equations as a single token and computing a vector representation based on the words in natural language surrounding them in texts, thus grasping semantic similarity. Mansouri et al. (2019) propose to compute embeddings based on hierarchical representations, such as Symbol Layout Trees and Operator Trees, exhibiting improvements in information retrieval tasks, resorting to hand-engineered pre-processing tools. A different approach to embeddings can be identified in more recent works. Purgał et al. (2021), for instance, propose different ways to encode logical structure in embeddings, to be leveraged in tasks related to theorem proving, while Ryskina and Knight (2021) compute numerical embeddings over Online Encyclopedia of Integer Sequences and then probe them with a series of mathematical tasks, such as divisibility, sequence completion or analogies (i.e. rule of three). While not insignificant, these recent results remain modest.

From a more critical perspective, Naik et al. (2019) show that vector representations for numbers produced through existing embedding techniques have difficulties accounting for magnitude (e.g. $3 < 4$) and numeration (e.g. $3 = \textit{three}$). More generally, Greiner-Petter et al. (2019) assess embedding techniques for mathematics in a more systematic way and exposes their semantic limitations due, among others, their incapacity to capture components of complex structure. A critical survey of different approaches and techniques of mathematical embeddings can be found in Greiner-Petter et al. (2020b); Thawani et al. (2021).

This entire line of work is of particular interest from the viewpoint of a text-driven philosophy of mathematics. It has also the merit of calling attention to the elementary mechanism upon which DNNs for mathematics ultimately rely, naturally bringing the discussion closer to conceptual stakes. However, there at least two reasons why this recent work seems insufficient from an epistemological standpoint. On the one hand, with very few exceptions like Purgał et al. (2021); Ryskina and Knight (2021), most of those works do not seem interested in the operational content of expressions. The semantics addressed here is mostly required for what is usually known as "mathematical knowledge

management", including tasks as information retrieval, labelling, classification of texts, document ranking, query recommendation or entity recognition. Admittedly, none those tasks could be performed in an unsupervised way without recovering some aspect of the mathematical expressions' semantics. However, it is not clear how this content could be leveraged to characterize the properties of mathematical objects and operate with them. On the other hand, in all these works, embedding techniques are being adopted somewhat uncritically, without inquiring into the reasons that could explain their success in extracting semantic properties from expressions in natural language. Not only the mechanisms underling the distributional hypothesis might not be entirely valid, but it might very well be that their validity for natural language, if any, does not carry over to mathematical textuality. Invoking the distributional hypothesis in this case does not really help unless we have a better understanding of the mechanisms underlying distributional semantics, which is astonishingly absent in the current state of NLP. This problem is all the more pressing as, in their current form, embedding properties are shown to be very brittle. Their efficacy is indirect, by increasing the performance of downstream tasks. On their own, the semantic similarity established through embeddings (not to speak about semantic analogy) does not look like a stable framework where to sit the firm construction of a systematic semantics, let alone mathematical semantics.

Accordingly, if a stronger treatment of mathematics is expected from current machine learning techniques, both a better conceptual basis for the connection between distribution and content and a deeper understanding of distributional properties of mathematical expressions are needed. In the last two sections of this paper, we will propose a fresh start to address these two issues from a philosophical perspective.

5 Toward a Structural Semantics

5.1 Making Sense

The first step toward a better grounded distributional account of mathematical knowledge is to clarify, from a conceptual standpoint, the relation distributionally defined expressive units, such as embeddings, maintain with the content they are supposed to convey.

Since the beginning of the deep learning revival of distributionalism, this question is assumed to be answered by invoking the distributional hypothesis, conveniently associated with a use or usage-based theory of meaning (Lenci, 2008). Generally attributed to Wittgenstein, the latter holds that the meaning of a linguistic unit is determined by how that unit is used. However, both philosophical principles came to be adopted and referred to rather uncritically in the field,[6] paving the way to a situation where the meaning import of distributional models—and DNNs in particular—is considered sufficiently accounted for by exhibiting high accuracy in the accomplishment of predictive downstream tasks, as Andreas (2018) correctly points out.

It is easy to see how this implicit assessment of the semantic grounds of distributionalism falls short of any conceptual or philosophical requirements:

[6] For a critical assessment of the association between those two principles, see Gastaldi (2020).

the problem is considered solved by not dealing with it. The attempt to move beyond such a situation received some attention by the computational linguistics and NLP communities, starting with the so-called "semantics mega-thread" in Twitter (Wolf, 2018). Within this context, a position begun to stand out claiming that, despite all appearances, models ungrounded "in the real world" dealing exclusively with linguistic forms (i.e., raw text) do not and cannot have any relation to meaning. In its most articulated version (Bender and Koller, 2020), the argument features as a restatment of Searle's famous Chinese Room argument (**?**), relying, at its core, on a definition of meaning as "the relation between a linguistic form and communicative intent" (Bender and Koller, 2020, p. 5185).

Although far from enjoying a large consensus in the field, this skeptical position has revealed fruitful to raise important critical debates (Bender et al., 2021) and motivate interesting theoretical results (Merrill et al., 2021). Moreover, alternative perspectives have found it difficult to articulate a coherent image for the semantics of current distributional language models.[7] However, from a strictly philosophical viewpoint, such a perspective has the weakness of excluding any connection between pure expressions and meaning *by definition*. Incidentally, it also rules out any possibility of conceiving meaning without the presence of a subjective (and possibly only human and individual) intentionality. Such a principled position is, therefore, of little use to address local epistemological issues as in the case of mathematics, where meaning is not indifferent to form, and communicative intent, while not entirely absent, is far from exhausting mathematical content. Incidentally, resorting to a radical distinction between natural and artificial languages to argue that a notion of meaning as communicative intent only applies to the former is not an option here, since the relevance of that distinction has been undermined by the very nature of our problem, as we have shown in the previous sections.

Coming up with a different general definition of meaning at this stage risks encountering similar obstacles, utlimately turning the discussion into a pure definitional one. A possible way out, however, is to leave aside the delicate notion of meaning for a moment, and focus on the fact that, in a significant number of cases (but not necessarily in all of them) the outputs of DNN models applied to either natural or mathematical language *make sense*.

The notion of sense-making has been the object of some conceptual elaboration in the last two decades of the past century, around the program proposed par Dervin (1983); Dervin and Nilan (1986); Dervin (1986). However, in all this literature, sense-making is conceived as the cognitive act of subjects who make sense *of* the world in which they live, bringing as back to an intentional problem. But when we say that the results of neural models make sense, by no means we mean to say that they make sense "of" something, let alone the world, or less still that those models have any intentionality. The way in which these models make sense appeals to a much more superficial, although not less substantial notion of sense. When a neural model yields `732` when prompted `406+326=`, or $\mathtt{y(x)} = \mathtt{c_1 e^x} + \mathtt{c_2 e^{-x}}$ to $\mathtt{y''} - \mathtt{y} = \mathtt{0}$, or any of the surprising texts to which LLMs have already got us used to, those results make sense, even be-

[7] The pervasive topics of intelligence, consciousness and sentience within the more general field of AI cannot but add more confusion to this debate, conceiling the fact that, at both its logical and linguistic roots, the contemporary notion of meaning emerged from an antipsychological standpoint. See note 10 below.

fore we know anything about the principles motivating the production of those expressions. In other words, those expressions make sense *as a pure effect*.

One could object that this is a degenerate understanding of the notion of sense and meaning. But the truth is that this notion of sense as an effect has a long history in Western thought outside the Aristotelian and analytical tradition, from the Stoics up to Deleuze (1990). However, it is not necessary to enter into doctrinal details here. For our purpose within the limits of this paper it is enough to recognize that the very fact that the behavior of DNNs urges us to reflect on the true conditions of meaning is a proof that it is indeed sense that it is being produced, for pure senseless results would not have that effect. Acknowledging that neural models make sense in this sense of "sense" is already a lot. For it allows us to suspend the difficult question of the very substance of meaning and turn it into this other one, namely: *what must sense be for DNN models to be able to make sense as they do*.[8]

The first thing to notice is that the notion of sense as it is used here is independent of those of consciousness or understanding. Anyone who has ever dreamt at night should know that it is possible to make sense in unconscious states[9] and everyone who has mechanically copied the habits of a radically foreign culture should be ready to accept that one can make sense without understanding. Notice also that this notion of sense is independent of the notion of truth: the expression `Paris` makes sense as the answer to the question `What is the capital of France?`, but also does `London` or `Babel`, for instance in a lie, a fiction, a joke, or a metaphor. Irreducible to truth, understanding or consciousness, the notion of sense implied in our most ordinary experiences of sense-making is, moreover, not an absolute one. What makes sense in some cases or contexts does not necessarily do in others. But we would be wrong to think that this relative character condemns this notion of sense to randomness or individual arbitrariness. On the contrary, the fact that what makes sense is relative to a context protects it from the arbitrary modification by single individuals: contexts are not something individuals can simply change on their own. Sense is thus to be placed at the level of a *culture*, conceived as a complex system of contexts, which is at once supra-individual and non-universal. And it does not seem unreasonable to hold that making sense in a culture requires little more (but no less) than identifying the regularities that culture is made of and extending them, even if in a deviant or creative way.

From this standpoint, neural models can be seen as particularly convoluted yet performant devices for identifying regularities in a corpus which is nothing but a partial—i.e., biased—expression of a culture. It is, then, in this sense that, regardless of "learning meaning", "understanding", "being intelligent" or "conscious", "knowing the truth", and all the other metaphors by which we might be more or less impressed, we can say that, to a surprising extent, and strictly relative to the inescapably biased corpus they are trained on and the cultural environment in which they output their results, neural models make sense.

[8] "As they do", that is, to a large extent and in relatively specific contexts, but neither completely, nor universally or absolutely.

[9] Dreams can be fragmentary or paradoxical, but never completely senseless.

5.2 Formal Content

The notion of sense advanced so far does little more than to rebaptize the problem of meaning, proposing what can be seen as a weaker version of it. However, accepting this weaker version as a legitimate part of our experience of meaning can not only prevent us from running into false problems, but can also open us to novel perspectives on our objects. This is especially true if we turn our eyes away from models and focus on the expressions themselves.

Leaving aside the question of their substantial meaning, expressions that make sense can be said to have a *content*. Take for instance the embedding representation corresponding to a unit or token in our corpus. While it is generally accepted that the meaning of such embedding is in most cases unclear or uninterpretable, it would be hard to deny that it associates to that unit a specific content, if only because it encodes some information characterizing that unit as such. And it is indeed based on that content that sense is produced as an effect. Content can thus be understood as an objective counterpart or correlate of the sense made by an expression.[10] More significantly, unlike meaning or sense, which are usually treated as unanalyzed wholes, the notion of content allows for a pluralist treatment. We can certainly conceive of an intentional content of expressions, corresponding to the communicative intent of speakers, as we can conceive a referential content informing a truth-conditional semantics, or even a psychological content. But none of them has reasons to exhaust or monopolize the content involved in the sense expressions can make. The kinds of content will always remain as varied as the ways expressions have of making sense.

Relying on this notion of content, in Gastaldi and Pellissier (2021), we have proposed to consider the existence of a *formal content* of expressions.[11] With

[10] "Objective" here is meant in relative way, as "non-subjective": contents are not to be attributed to any subject or agent, but to expressions themselves. A notion of objective content like this one is not new, and can already be found in the philosophies of Frege and Husserl (cf. Benoist, 2002), and their antipsychologist sources, such as Bolzano's notion of "proposition in itself" (cf. Proust, 1999). However, in the sense intended here, it finds its root in Louis Hjelmslev's introduction of the notion of content in the field of linguistics, as a generalization of Saussure's notion of "signified" (Hjelmslev, 1953, §13). The main purpose of that introduction is to characterize all the semantic aspects of signs which can be studied from the internal perspective of signs themselves, without borrowing to other sciences or disciplines outside linguistics. The notion of content has then a primarily epistemological role, namely, contributing to establish linguistics, and more generally, semiology, as an autonomous discipline.

[11] The notion of formal content might appear eccentric if not altogether paradoxical to anyone assuming the traditional exclusive alternative between form and content. However, here again, the idea of a formal content is well rooted in a philosophical tradition that goes back at least to German Idealism with its problematization of the opposition between form and content, either in its transcendental (Kantian) or dialectical (Hegelian) versions. A more recent version of this notion can be found in Frege, for whom the notions of individual, concept, judgement and truth value constitute elementary forms of contents supporting the formal foundations of arithmetical contents (Gastaldi, 2014). The idea can also be found in the artistic and theoretical work of the Russian formalists, where artistic content is seen as the effect of formal procedures (Lemon and Reis, 2012). But once again, our notion of formal content recognizes its most direct source in Hjelmslev's linguistic elaboration of the notion of content, which draws from Saussure's famous principle stating that "language is a form, not a substance" (Saussure, 1959, p. 113). Thus, for Hjelmslev, both expressions and contents can be analyzed into substance and form, linguistic or semiological analysis being defined by the study of the relation between *forms of expressions* and *forms of content* (Hjelmslev, 1953, §13). However, our use of this notion here does not exactly correspond to Hjelmslev's.

it, we intend to characterize that dimension of content which finds its source in the internal relations holding between the expressions of a language. This dimension is usually neglected from the referential or truth-conditional—but also pragmatic or intentional—standpoints which have dominated contemporary semantic theories. However, its prominent role is well-known across a myriad of sense regimes: poetry, music and art in general, games and other rule-following practices, including law, and, of course, formal sciences, and mathematics in particular. Indeed, within a standard model of arithmetic, it is part of the mathematical *content* of the expression `406` that, when associated with the expression `+326`, it can be replaced with the expression `732`. The set of all such relations the expression `406` can maintain with all the arithmetical expressions can then be said to characterize its arithmetical content to a large extent, whatever all those expressions might stand for.[12] Certainly, one could still claim that all these sense regimes relate to natural language only in a derived and deviated fashion, even in the cases in which natural language plays indeed a central and inescapable role, like poetry or law. Yet, as we have seen in the previous sections, if there is something recent neural models have taught us about natural language despite their weak epistemological import, it is that the distinction between natural language and other kinds of language cannot be taken for granted. But more significantly, that the formal content of language is responsible for much more sense effect than we would initially be ready to attribute to it from the perspective of a natural attitude.[13]

In an attempt to further specify the notion of formal content bringing it closer to its operational conditions, in that same work, we identified three dimensions that seem to characterize it. In other words, three different systems of relations between linguistic or semiological units contributing to their sense effects and derivable, at least in part, from the distributional properties of a corpus. Thus, we proposed to distinguish between the *syntactic content*, namely the content a unit receives as a result of the multiple dependencies it can maintain with respect to other units in its context; the *characteristic content*, resulting from the inclusion of a unit in a class of other units by which it accepts to be substituted in given contexts; and finally an *informational content*, related to the non-uniform distribution of units within those substitutability classes. Take, for instance, the unit `gavagai`, whose meaning is famously assumed to be completely undetermined (Quine, 2013). However, as soon as that unit is put in relation to others in an expression like `the gavagai is on the mat`, part of its indeterminacy is lifted and a content (syntactic, in this case) can be attributed to it as a result of the sense that expression makes. That this is a content at all can be easily felt by contrasting that expression with another one which does not make sense (in usual English language) like `mat the on is gavagai the` (borrowing from Chomsky's famous example). That such a content is specific can be shown by comparing the different ways of making sense of that expression and this other one: `the people gavagai on the mat`. Likewise, much can be said about `gavagai` if we can establish that it can be substituted by `cat`, `dog`, `bird`, `rabbit`, `spider`, etc. (thus receiving a characteristic content) and even

[12] This idea is not foreign to the axiomatic approach to mathematics, and more generally the formalist approach to its foundations, ideologically dominant throughout the 20th century up to our days. However, the latter typically attributes a primary role to logic in the control of the interaction between expressions that is voluntarily absent in our example.

[13] For this notion of natural attitude, see Husserl (1983, ch 1, §1).

more if we know that its probability is the lowest compared to all the other units of that class (informational content) (see Gastaldi and Pellissier, 2021, §2.2).

Those three might very well not be the only dimensions of formal content, and it is important to recall that formal content is not the only content expressions can rely on to make sense. Yet, those three dimensions give us already an idea of how to account for the sense effects of a strictly formal content in a more detailed and principled way than mechanically invoking the elusive action of the distributional hypothesis. For all these dimensions implicitly defined by the distributional properties of a corpus can be, in principle, made apparent by drawing explicitly the latent units over which they act. Those units are no other than the ones determined by the relation of substitutability between the explicit or observable units of the corpus. This is trivially so for the characteristic content, but also holds for the syntactic and informational contents. Indeed, in the former, when looked at closely, the dependencies in question hold between classes of substitutable terms rather than between terms themselves,[14] while in the latter, unlike the typical approach in the field, probability distributions and information measures are computed only with respect to units with equivalent characteristic content.

The idea of latent or implicit units determined by the relation of substitutability between expressions recognizes both a logical and a linguistic source in contemporary thought. On the logical side, it was Frege who gave substitutability its most celebrated formulation, as a defining principle of his generalized notion of function (eventually understood as a propositional function). Indeed, in his *Begriffsschrift*, Frege introduces functions over expressions (*Ausdrücke*) by regarding parts of those expressions as replaceable, implicitly considering the class of all the expressions that can be replaced at that point, and eventually the subclass of expressions yielding a given (truth) value of the function (Frege, 1972, §9-10). On the linguistic side, the idea of substitutability of expressions in given linguistic contexts informed the concept of associative or paradigmatic relations which, together with that of syntagmatic relations, constitute the key principle governing the mechanism of language in the structuralist view which revolutionized linguistics at the turn of the 20th century (Saussure, 1959, §II-V). The explicit derivation of implicit paradigmatic units, or *paradigms*, became thus a major stake of modern linguistic analysis, until the structuralist approach was relegated by the generativist program. The revival of distributionalism in the analysis of language through DNNs provides the conditions for a renewed approach to paradigmatic units. What is more, contemporary proof theoretical approaches to computational logic allow an interpretation of paradigms as types which would permit to bring together the logical and the linguistic traditions in an original fashion (cf. Gastaldi and Pellissier, 2021).

The explicit derivation of paradigmatic units would represent an evolution from current distributional perspectives to a renewed structuralist framework. Such evolution is marked by the fact that, while individual tokens are considered to be determined by the local context in which they appear, higher order units imply a global account of the corpus which is only implicit in current models.

[14] For instance, the syntactic content `gavagai` receives from `the` is the same that `cat`, `dog`, or `rabbit` would receive by being at the same position, which is the same it would receive from `a` or `one`, as long as these terms are seen as substitutable. As soon as the dependency relies on the specific terms themselves, we can conceive it as a lexical dependency, rather than a syntactical one, as in the proper name `The Great Gatsby`.

Accordingly, meaning can be conceived not so much as the use of words in similar local contexts, as advocated by the distributional hypothesis, but as the effect of subsuming expressions under a structure underlying the corpus to which they are assumed to belong. In this sense, it comes as no surprise from a structuralist standpoint that the most fundamental objective neural language models optimize, including Transformers, is that of predicting words throughout all the contexts that constitute the corpus, thus resulting in a language model that attributes high probability to all the words in the vocabulary that would be acceptable substitutions of the original word in given contexts. As is hardly surprising for a structuralist that so much sense can be made by levering only this kind of information.

Following this orientation, it seems possible to advance from the bare observation that neural models make sense and distributional units bare similarity relations to a finer account of the content implied therein. Drawing from those sources of inspiration in the tradition of modern theories of language, it might be possible to leverage distributional properties to derive explicit representations of paradigmatic units possibly at work in current neural language models, and reconstruct structural and logical aspects underlying the use of linguistic expressions.

6 Arithmetic from an Expressionist Perspective

The structural perspective on distributionalism just given does not rely on any specific characteristic of the expressions in question.[15] As such, it intends to provide insight on how content can result from pure expressions through the same mechanisms across multiple language modalities.[16] In particular, the study of the three dimensions of formal content considered cannot only contribute to the intelligibility of sense-making for expressions in natural language, but also in mathematics. In the remaining pages of this paper we will propose an illustration of this fact which, although elementary and highly speculative, can suggest intelligibility principles for either existing or alternative machine learning approaches to mathematical knowledge.

Our illustration concerns elementary arithmetic, trying to tackle the main problem advanced in previous sections, namely how arithmetical content can result from the distributional properties of pure expressions. More precisely, we will focus on arithmetical objects, that is *numbers*, interested in how to reach "numbers through numerals", following the suggestive formula of Schlimm (2018). However, our approach will be at once more radical and more humble than what this formula suggests. More radical, because interested as we are in the kind of raw expressions neural language models are capable of processing, we cannot assume numerals exist as an independent sign regime, without already assuming significant aspects of arithmetical content. More humble, because we

[15] It might seem that we have restricted ourselves to sequential data. However, the notion of context involved does not assume such a limitation, and notions of context can be proposed for the analysis of non-sequential data.

[16] Owing to this genericity, the perspective provided here should be viewed as contributing to the program of a general semiology, as set up by Saussure (1959, §Introduction, III.3).

will be interested in only two aspects of arithmetical content: the *recursive structure* and the *total order* of natural numbers

Although those two aspects are far from exhausting the complex content of arithmetical expressions, they are, however, essential to the concept of number if one believes the modern reconstruction rooted in the logical perspective dominating the philosophy of mathematics since the end of the 19th century. From a classical axiomatic standpoint on natural numbers, such as the Dedekind-Peano axioms (cf. Sieg and Schlimm, 2005), both the recursive structure and the total order are guaranteed by the single-valued successor function. Yet, such an account is strictly semantic in the sense advanced in previous sections: pure expressions do not play any role in the content attributed to numbers, and can only be thought to represent that content *a posteriori*.[17] If we adopt a perspective where no semantics is assumed, both aspects of the content of natural numbers should then be regarded as independent. Yet, none of them is a property of characters themselves. If it is nonetheless possible to derive them from pure text, then, as we saw, it can only be done through the latter's distributional properties. The notions of characteristic, syntactic and informational content advanced in previous sections will guide us through this exercise.[18]

6.1 The Characteristic Content of Digits

The first issue to be addressed is the identification of numerical expressions (or numerals) as a autonomous class of expressions. Because, from the viewpoint of neural model representations, numerals are indistinguishable from any other kind of token. As we said, considering the class of numerals as given (say, as the set of sequences of digits, assumed, in turn, to be a specific class of characters) implies already assuming a global unit accounting for much of the content supposed to be derived. A naive distributional approach, by which all numerals could be identified as having similar distributions, would not be of much help here since expressions like `two`, `million`, `many` or `several` exhibit distributions similar to them, without being numerals nonetheless. More sophisticated treatments of distributions, such as DNNs, are certainly able to distinguish between numerals and other numeric or quantitative expressions, but the reasons for that distinction remain opaque, and thus unreliable for a characterization of the content of numbers.

However, numerals can be distributionally characterized in a more principled way if instead of considering lexical or even morphological tokens, we go all the way down to the character level. Take, for instance, the English Wikipedia corpus (Wikimedia Foundation, 20220301.en dump). From the character viewpoint, the corpus appears as a long sequence of characters, without other explicit structure. To accentuate this lack of structure, we can even consider removing all punctuation marks, including white spaces, as well as lowercasing all char-

[17] In this sense, a more interesting approach to the recursive content of natural numbers can be found in Frege's pre-logical work, namely his *Habilitationsschrift* (Frege, 1874), where he attempts to characterize numbers through a recursive structure identifiable over domains of functions (cf. Gastaldi, 2016). This unorthodox approach, which is still somewhat present in Frege's attempt to provide a functional language for arithmetic in his *Begriffsschrift*, certainly explains the generality of the notion of sequence and succession proposed in that work, which revealed to be insufficient to guarantee total order.

[18] A computational implementation giving access to the experimental data presented in this section can be found on `https://github.com/Gianni-G/MathML-2_code`.

acters, which also has the effect simplifying our example.[19] A typical sample from this corpus looks like this:

`asof2007reviewsestimateaprevalenceof1-2per1000forautism`

In such a corpus, every character is assumed to be equally different from all the others. In particular, digits do not have any special internal characteristic separating them from the rest of the characters, and unlike usual formal systems, we cannot rely on given rules to produce "well-formed formulas" which would separate numerals, as strings of digits, from other possible combinations of elementary characters. In other words, numerals cannot be assumed to exist as such.

Yet, the simplest of distributional properties, namely the left or the right 1-character length context of each character, can easily make that circumstance appear. Indeed, consider the matrix A where each element $A_{i,j}$ corresponds to a measure of the association between the two characters c_i and c_j. A natural choice for that measure would be the frequency in the corpus of the 2-character sequence c_ic_j, or a normalized version of it representing its empirical probability $p(c_i, c_j)$. A better and widespread measure, however, is provided by the *pointwise mutual information* (pmi), namely (an information-theoretical transformation of) the ratio between the actual probability $p(c_i, c_j)$ of the pair of characters c_ic_j and the probability one would expect if both characters were independent, given by the multiplication between the probabilities of each one of the characters: $p(c_i)p(c_j)$. Concretely:

$$\mathrm{pmi}(c_i; c_j) = \log \frac{p(c_i, c_j)}{p(c_i)p(c_j)}$$

A positive or a negative pmi indicates that the two characters are positively or negatively associated, i.e. they appear more or less frequently than if they were independent, whereas a pmi close to 0 indicates that both characters are practically independent. A normalized version of pmi (npmi) can be achieved by dividing the pmi by the joint self-information, i.e. $-\log p(c_i, c_j)$. All values of npmi will then lie between a minimum of -1 (in the limit) and a maximum of 1, with 0 indicating independence.

With the npmi as association measure, we can then build the matrix A where $A_{i,j} = \mathrm{npmi}(c_i; c_j)$. Figure 1 shows such a matrix for the first 40 most frequent characters of the Wikipedia corpus, which account for more than 99.7% of the occurrences in that corpus. Interpreted distributionally, this matrix features all the information about characters with respect to their 1-character length right or left contexts. Indeed, the ith row $A_{i,:}$ collects all the association measures of the character c_i in all the right contexts. Likewise, the jth column $A_{:,j}$ presents the the association measures of the character c_j in all the left contexts.

The importance of this matrix is that *rows and columns can be used as explicit distributional vector representations or embeddings* (cf. Levy and Goldberg, 2014). These explicit vector representations (EVR) can help to establish two important facts. First, (row or column) EVRs of digits are very similar to each other and significantly different from the rest of the characters. This is hardly surprising if we already know what digits are and how they work. But if

[19] None of this is strictly necessary. A character-based model would still be able to grasp the same properties analyzed here without any pre-processing of the corpus.

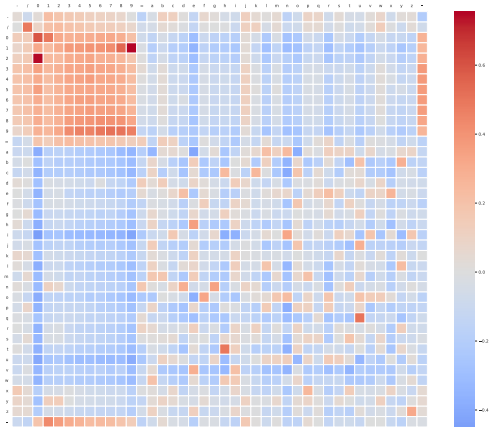


Figure 1: Normalized PMI matrix for the 40 most frequent characters in the English Wikipedia corpus.

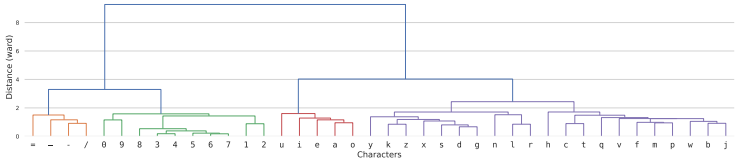


Figure 2: Hierarchical Clustering of the concatenation of left and right EVRs for each character in the Wikipedia corpus. The distance is computed using Ward variance minimization algorithm.

we stick to a purely distributional approach, this means that, whether we consider right or left contexts, distribution is enough to constitute digits as a group of characters clearly separate from the rest. By performing a classic hierarchical clustering on either row or column vectors, or even better, on a concatenation of both, thus considering the complex right-left context for each term, we can easily confirm this fundamental fact (Figure 2).[20]

The fact that the EVRs corresponding to the characters `1, 2, 3, 4, 5, 6, 7, 8, 9, 0` constitute a cluster indicates that, in the current corpus, any of those elements could, in principle, appear at the place of any other, for distributional similarity of terms entails similar predicted probability for different contexts in distributional models. This is other way of saying that those characters share the same characteristic content. Such a content determines the existence of digits as a class, if only implicitly. While a capital letter D can function as a symbolic representation of that class, the properties on which it rests actually provide a way to conceive and manipulate that representation in

[20]The clustering is computed using Ward variance minimization algorithm. The clear distinction between the 4 main clusters is, however, consistent across different methods.

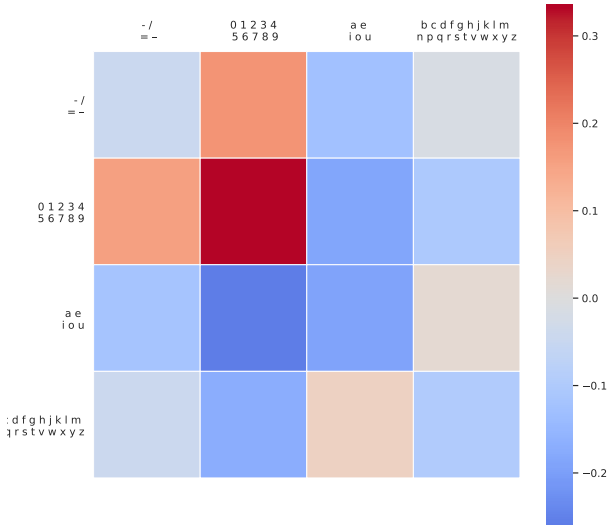


Figure 3: PMI matrix of clusters of characters, represented by the average vector of the members of each cluster.

distributional terms, for instance, as the average vector of all the EVRs of its members or the centroid of their cluster.

6.2 Syntactic Content and Recursive Structure

The characteristic content of digits is not enough yet to define the content of numerals or even characterize them as an identifiable class of expressions. After all, the clustering procedure resulted in the identification of other classes as well. Without further determinations, there are no reasons to believe that what we know to be digits and sequences of digits are endowed with a specific content, let alone an arithmetical one.

Yet, if we use the clustering information to draw a compressed version of our original PMI matrix A by replacing all the row and column EVRs by the distributional representation (e.g. average vector) of their corresponding clusters, we can reveal a singular property of the class of digits (Figure 3). Indeed, unlike all the other classes, when considered as a whole, *the class of digits is positively associated with itself*. The analysis of the EVRs of each digit in the original matrix shows, moreover, that every member of the class exhibits a high pmi with all the other members, including itself. This means that, as a general (statistical) rule, digits only appear in the context of digits. This is a remarkable fact, which puts us on the track of a distributional characterization of the recursive content of numbers.

Following Tomalin (2011), from the standpoint of linguistic structures, recursion appears to refer to two distinct principles at least: *iterative constructional*

devices and *self-similar syntactic embedding*. If we recall the Dedekind-Peano axioms, both principles concur to define the recursive content of numbers. On the one hand, the successor function plays the role of a device governing the iterative application of a rule (usually interpreted as "adding 1"). On the other, the stipulation that the successor of a number is also a number guarantees the self-similarity of objects through the application of the rule. Here again, the association between both principles is imposed as a semantic requirement, forced, as it were, upon any syntax intending to represent numbers. If a recursive structure is to be found at the level of numerals alone, a more subtle articulation is to be expected.

Returning to the remarkable property of the class of digits, by which digits overwhelmingly appear in the contexts of other digits, we can see here the rudiments of an iterative device. This can become even clearer if we recall that, in its most elementary form, a generative language model can be seen as a function f of a sequence of linguistic units yielding yet another unit for that sequence. When f yields a digit based on the presence of a digit in its argument, it mobilizes the distributional properties we have explored so far. However, the strict iterative structure is difficult to see if we only focus on individual tokens, like current machine learning approaches are typically supposed to do. Indeed, no recursion is explicitly at work other than trivially if the most elementary of models is conceived to produce, for instance, the sequence `406` as the result of applying f to `4` and then iteratively to the resulting values: $f(\texttt{4}) = \texttt{0}$, $f(\texttt{0}) = \texttt{6}$. However, we have seen that the embedding representation of all those characters is highly similar. What f receives at each step is then a very similar vector. It is then possible to imagine how a stronger notion of recursion can be here at work. If an explicit representation of the class D of digits were available based on the latter's characteristic content, then an iterative device would become manifest since, in this case, D turns out to be a fixed point of f, i.e., $f(D) = D$.

Such a fixed point can account for the iterative behavior of digits in writing. As such, it determines a significant aspect of their *syntactic content*, contributing in this way to characterize a specific class of character sequences. Indeed, numerals can be seen as the sequences of characters generated by this fixed point, which thus informs part of their recursive structure. Under this light, numerals appear as the Kleene closure of D, typically denoted D^*.

Assuredly, we cannot assume all these symbolic representations are explicitly encoded in current distributional models. What is more, the functioning of current language models is far from reducible to predicting the next token based on the previous one alone. However, nothing prevents them from having such a fixed point for a component f of a more complex function f' representing the actual language model. In which case, the latter would count numerous mechanisms preventing them from getting stuck in the infinite loop entailed by any fixed point. It goes without saying that determining whether an elementary iterative device as the one provided by a fixed point $f(D) = D$ is present and identifiable in current neural models is an empirical issue. What is important for us is that such a device is derivable in principle from distributional properties without relying on any other means than the usual mechanisms to infer features of their latent structure, provided paradigmatic units are granted an explicit representation in one way or another.

Notice that the iteration device here is not to be mistaken for the successor function. Its role is strictly syntactic, generating a digit given another digit

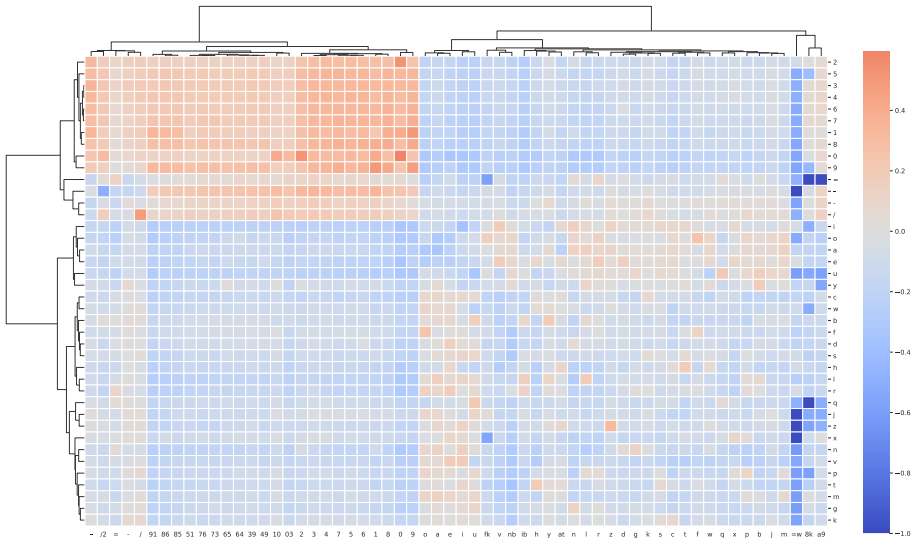


Figure 4: Normalized pointwise mutual information (npmi) matrix with hierarchical clustering of rows and columns including 10 random 2-digit terms and 10 random 2-character terms for control.

in the context. As such the iterative device provided by $f(D) = D$ seems to suffice to characterize the *syntax* of numerals as sequence of digits, but not necessarily their *content*. More precisely, if, following Gastaldi and Pellissier (2021), we consider clusters of EVRs, such as D, as a formal types describing the interaction of terms in a corpus (here, the characters), numerals could be understood as terms of the types D, $D \otimes D$, $D \otimes D \otimes D$, etc., where $\otimes$ denotes the compositional relation of concatenation lifted from terms to types. However, it is not clear what distributional property characterizes all those sequences of characters as having a common content. In other words, what makes the different types D, $D \otimes D$, $D \otimes D \otimes D$, etc. be "of the same type"?

The answer can come from another remarkable fact of the class of digits, motivating the second aspect of recursive structures, namely self-similar syntactic embedding. For if we consider all the terms of the type $D \otimes D$, i.e., all 2-digits sequences, and compute their EVR as we did for 1-character terms[21] it appears that, in the corpus in question, *they are all similar to those of the terms of type* D, i.e. the digits. We can easily confirm this circumstance again by adding the corresponding rows to our previous PMI matrix A and performing a clustering of rows and columns (independently this time). We can clearly see, then, that 1 and 2-digit terms are clustered together (Figure 4).

As it has been said, terms that have similar embedding representations are considered by the model as possible substitutions of each other in given contexts. In our words, we have that *terms of the types* D *and* $D \otimes D$ *have the same characteristic content*. If explicit representations were available for such types, this

[21] That is, computing the normalized version of the $\mathrm{pmi}(d_1 d_2; c_j) = \log \frac{p(d_1 d_2, c_j)}{p(d_1 d_2)p(c_j)}$ for a 2-digit sequence $d_1 d_2$ with respect to all the characters c_j to their right, and adding the resulting vector to the one built from the same measure but with respect to characters on the left.

would mean that both types are equivalent modulo their characteristic content, or that the characteristic content of digits is preserved through concatenation. Certainly, the finite character of the corpus will make the EVRs of sequences of digits progressively approach the unigram distribution of characters in the corpus after a few number of concatenations, moving away from the EVR of D. Yet, it would only take to encode in some explicit fashion the substitutability between $D \otimes D$ and D unambiguously resulting from the data for a distributional model to capture an elementary principle of self-similar syntactic embedding.

That substitutability of types, based on the similarity of their characteristic content—which we can informally denote by $D \simeq D \otimes D$—, together with $f(D) = D$, provide a distributional characterization of the recursive structure of numerals. Together, those two principles guarantee that a digit can always be concatenated with a digit yielding a meaningful expression, and that a sequence of digits has the same characteristic content of a digit and hence concatenation between digits preserves the type of content. In this way, a particular combination of the characteristic and the syntactic content of digits together with the characteristic content of numerals can account for the latent recurrent structure of the latter, which can be considered a fundamental aspect of the semantic of numbers.

It is easy to see that the recursive structure at play here does not coincide with that of the successor function. While every application of the iteration rule does indeed yield an entity of the same kind as the previous one, nothing guarantees that the resulting one is the successor of the original one. As announced, from a distributional perspective, the recursive behavior of numerals is independent, in principle, from their order. In this sense, the recursive structure derivable from distributional properties in the way suggested here is significantly more general than the one represented by the Dedekind-Peano axioms. One can even imagine that the same principles could account for the recursive structure of positional numerical writing systems much more complex than the decimal one, including the subtleties of the latter, such as the fact that the class of the first digit in numerals excludes `0`, which we decided to leave outside the scope of this paper in the sake of simplicity. Notice however that it would suffice to identify a similar structure over a class containing a single member to retrieve a structure like Dedekind-Peano's. Such a possibility is not entirely foreign to natural language corpora such as the one used here. Indeed, characters such as `*`, `_`, `~` appear to meet those conditions. Interestingly, according to the perspective to be presented in the following section, order would follow directly from the recursive structure in that case.

6.3 Informational Content and Total Order

As fundamental as it may be, the recursive structure is only one of the many aspects of the content of numbers, no more than the characteristic content is only one dimension of the formal content of numerals. If the characteristic content were the only content of expressions, all numerals would have the same content. Their syntactic content does not seem to help in this case. Although the analysis of the syntactic content of numerals falls outside the scope of the present paper, we can see that the clustering in Figure 3 suggests that, as a class, numerals relate to the class of characters such as `-`, `/` or `=`, i.e., operators and relations, which in turn relate to numerals. However, from a strictly syntactic

perspective, any numeral accepts to be replaced by any other in the context of operator or relation characters and other numerals, unless the content of the latter is already differentiated in some other way.

Being acquainted with the semantics of number, we know that such a differentiation finds its source in their total order, i.e., the property that, for any two numerals, it is always possible to determine which one is greater or smaller than the other. The successor function ensures this property trivially. However, for the expressions obeying the recursive structure just derived, order should result from distributional properties alone. According to the different dimensions of content, if that differenciation is supposed to be formal, its source cannot be other than the *informational content*, that is, the specific probability distribution of terms sharing the same characteristic content.

The idea that total order, as a defining property of natural numbers, can be derived from probability distributions of numerals is certainly unintuitive. However, it is known that numerals occurring in many kinds of real-life datasets obey a particular distribution known as Benford's law. The law was originally identified as an empirical observation by Newcomb (1881) receiving further elaboration later by Benford (1938). It states that, in such datasets, the distribution of the leading or first significant digit of numerals in positional notion (e.g., `4` in `406` or `7` in `0.00732`) is not uniform as one would intuitively expect, but follows a logarithmic law, namely:

$$p\left(\mathtt{d}_1\right) = \log_b\left(1 + \frac{1}{d_1}\right) \tag{1}$$

where $\mathtt{d}_1$ denotes the 1st digit, d_1 its numerical value (correlated to its order), and b the base of the notation system. Thus, for a positional system in base 10 as the one we are analyzing, we have that the probability of the digit `1` to appear as the leading digit in numerals is ∼0.301 (i.e., $\log_{10} 2$), while that of 9 is ∼0.046 (i.e., $\log_{10}\left(1 + \frac{1}{9}\right)$).

It is not here the place to enter into the details of this law and its possible explanations.[22] What is important for us, is that Benford's law provides a principled way of connecting the statistics of digits to the order of the numbers they represent. For if it is possible to predict the probability of the first digit $\mathtt{d}_1$ of numerals in specific corpora based on its numerical value d_1, the reciprocal should also be true. It should be enough to consider the inverse of Benford's formula, namely:

$$d_1 = \frac{1}{b^{p(\mathtt{d}_1)} - 1} \tag{2}$$

Notice that (1) is strictly monotonic, and hence so is (2), which guarantees that total order is preserved.

Significantly, Benford's law is not limited to the first digit. In his original paper, Benford had already shown how to compute the probability of the mth

[22] Benford's law received many explanations, including the spread of data over several orders of magnitude, the constant growth rate of processes represented by numbers, the scale or base invariance of those processes, or the action of the Central Limit Theorem (cf. Miller, 2015, for an overview). However, in recent years, Whyman et al. (2016) showed that the law is "a consequence of the structure of positional digit systems" (p. 5), and provided a corresponding alternative formula for the statistical estimation of first digit's probabilities, convergent with the traditional formula.

digit in a numeral. What is more, it is possible to provide a generalization of Benford's formula for the sequence of the first m digits (cf. for instance Miller, 2015, ch. 2):

$$p(\mathtt{d}_1 \otimes \mathtt{d}_2 \otimes \ldots \otimes \mathtt{d}_\mathtt{m}) = \log_b \left(1 + \frac{1}{\sum_{j=1}^{m} d_j b^{m-j}}\right) \tag{3}$$

where, $\otimes$ denotes the concatenation of digits, as before.

It is easy to see how the sum in the right-hand side of this formula encodes the polynomial structure underlying the positional notation system. But the important thing for us is that, like the original formula, this generalized version can also be inverted.

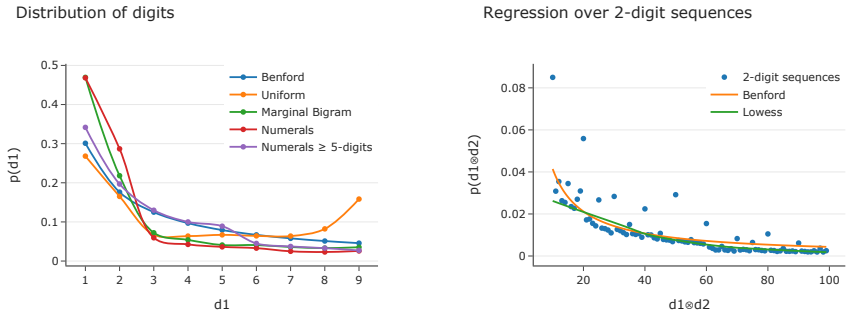
In sum, Benford's law ensures that the numerical values of sequences of digits, and *a fortiori* their total order, can be retrieved from their probability. It guarantees that a function turning the latter into the former, as complex as it may be, exists. As such, the probability distribution it describes constitutes the theoretical distribution of our numerals and determines, in this way, their informational content. Relying on this analytic characterization, such probability function could, in principle, be approximated from the empirical statistics of an appropriate corpus.

But only *in principle*. In practice things are not so simple. The first thing to verify is whether the corpora we are interested in, i.e. natural language corpora, is one of those "appropriate" corpora, that is, whether they actually verify Benford's Law. For concreteness, we will stick to the English Wikipedia corpus.[23] In this corpus, the unigram distribution of digits is nowhere near Benford's.[24] This is normal since the law should hold only for leading digits, while the unigram distribution is an aggregate of all digits. Yet, this already tells us that a naive approach will not do, and that the position of digits will need to be somewhat taken into account. If we consider all the pairs of a non-digit character and a digit, the marginal probability of digits presents a more optimistic landscape. Here, the empirical probability of digits does indeed induce the expected order, except for `9`, whose probability is slightly above that of `7` and `8`. However, the corresponding probabilities are not quite those predicted by Benford's Law. A more accurate approach, extracting all digit sequences from the corpus and estimating the empirical probability of their leading digit confirms this circumstance, while exhibiting a slight overestimation of `2`, without however affecting its order. This fact gives us a clue of what might be the source of the divergence with respect to Benford's law. Featuring encyclopedic texts, dates, and years in particular, are certainly overrepresented in Wikipedia, especially those corresponding to the past 40 years. A circumstance that can be confirmed by computing the probability distribution for 4-digits sequences.[25] If we perform a rough normalization by computing the leading digit of numerals of 5 digits or more, then we can see that the distribution approaches Benford's, and more significantly, induces the expected order over digits (Figure 5a).

[23] Although we are confident that the weak conclusions we will draw here would still hold in other commonly used natural language corpora, their validity needs to be empirically established case by case.

[24] For simplicity, we leave here aside the thorny consideration of `0`.

[25] The 10 most probable sequences are: `2010`: 0.021, `2011`: 0.018, `2008`: 0.018, `2012`: 0.018, `2009`: 0.018, `2007`: 0.017, `2006`: 0.017, `2014`: 0.017, `2013`: 0.017, `2015`: 0.017. All of them are in the top 30 most probable numerals of all lengths.



(a) First significant digits

(b) First 2 significant digits

Figure 5: Distribution of leading digits in the Wikipedia corpus.

One could then infer without risking much that the significant-digit law is indeed present in natural language corpora of relatively big size, if the bias effect of the aggregation of different numerical systems (like dates) is taken into account and somewhat filtered out. This is not a strong assumption since current neural models rely on tokenizing mechanisms, such as BPE (Sennrich et al., 2016), that lexicalize frequent sequences and have an implicit normalizing effect on the distribution of units, including numerals.[26] One could also resort to other known techniques, such as different forms of implicit or explicit matrix factorization, which could retrieve Benford's distribution of first digits from empirical distributions of corpora in a natural way. But what is important for us is not so much to retrieve Benford's exact distribution, but a distribution of digits that induces an order corresponding to that of their numerical content in a robust and reliable way. Despite the fluctuations on the empirical distributions, and modulo the normalizations mentioned, there are indications that this might actually be the case.

However, even if this is assumed as a working hypothesis, it represents little progress towards the ultimate goal, which is to ensure the total order, not only of first significant digits, but of all digit sequences. The empirical distribution of numerals in real-life corpora as the ones we are interested in is way too brittle to rely on. In addition, the very nature of the problem implies that the probability of finding any particular numeral in a given corpus becomes practically zero after some relative low value (say, a couple of thousands). The strong generalization capacity expected from a principle of total order for numerical expressions of arbitrary length and the practically negligible data available in empirical corpora makes a direct and naive derivation of such a principle—for instance through simple regression over sequences of digits—effectively out of reach.

Yet, not everything is lost if we recall that our numerals are not reducible to a simple class of sequences. Although not analytically defined in this framework, the distributional characterization we proposed in the previous section presented the numerals as a recursive construction from digits. In other words, the class of numerals (which we can informally denote N), as long as it can

[26] Our current work in progress suggests that this is indeed the case for Benford's Law.

be distributionally characterized as such, is *structured*. Significantly, both here and there, the challenge is to identify a distributional principle which would allow a purely syntactic analysis to reach the content of N from that of D. As for its characteristic content, the answer came from the recursive effect of the concatenation operation $\otimes$ when considered at the level of explicit representations of a class of characters, namely D. The key to this construction was the similarity of the characteristic content of D and $D \otimes D$, directly derivable from the distributional properties of the corpus. A similar strategy might be followed for the generalization of the informational content of D to N.

All we need is that the principle of the informational content of D, i.e. the law behind the probability distribution of its members ensuring the order of their numerical content, is preserved through the operation $\otimes$. We already know that the law for the leading digit (1) can be generalized for sequences of digits of arbitrary length (3), and that both formulas are invertible, so that a function governing the informational content of both D and N exists. If only could we find a way to describe the latter from the former, a distributional characterisation of total order as the informational content of numerals would remain possible.

If we look back at equation (3) it appears that this is indeed possible. For the generalized version of Benford's law is expressed as a function of the numerical value of the composing digits. But we have a way to compute those values from the probabilities of the digits, namely through the inverse of the simple law, as expressed in (2). Since, additionally, (3) is invertible, we have a way of expressing the value (i.e. the order) of the concatenation of digits out of the probability of its components.

Take, for instance, the simplest case of 2-digit numerals. If we denote $\mathcal{B}$ Benford's formula (1) for leading digits and $\mathcal{G}$ its generalized form (3), and $\mathcal{B}^{-1}$ and $\mathcal{G}^{-1}$ their respective inverses, then we have that the probability of the concatenation of digits is a function of the probability of those digits, namely:

$$p\left(\mathtt{d}_1 \otimes \mathtt{d}_2\right) = \mathcal{C}\left(p(\mathtt{d}_1), p(\mathtt{d}_2)\right) = \mathcal{G}\left(\mathcal{B}^{-1}\left(p\left(\mathtt{d}_1\right)\right), \mathcal{B}^{-1}\left(p\left(\mathtt{d}_2\right)\right)\right)$$

It thus appears that, as a certain combination of $\mathcal{B}^{-1}$ and $\mathcal{G}$, the function $\mathcal{C}$ invests the operator $\otimes$ with a probabilistic content (thus extending the informational content from classes or types to operators or connectives). Such a content is not purely abstract since, thanks to the explicit forms of $\mathcal{B}^{-1}$ and $\mathcal{G}$, $\mathcal{C}$ can also be explicitly given. Indeed, for $\mathtt{d}_1$ and $\mathtt{d}_2$ the first two digits of a numeral, we have that:

$$\mathcal{C}\left(p\left(\mathtt{d}_1\right), p\left(\mathtt{d}_2\right)\right) = \log_b\left(1 + \frac{1}{\frac{1}{b^{p(\mathtt{d}_1)}-1}b + \frac{1}{b^{p(\mathtt{d}_2)}-1}}\right)$$

It is possible to generalize $\mathcal{C}$ so that $\otimes$ can take arguments which are themselves composed, thus becoming associative. The generalization is somewhat subtle because, as announced, the length of the sequences in the arguments needs to be taken into account. That length is a syntactic property, and as such, available for a purely syntactic analysis, so such a condition does not fall outside the distributional framework.[27] In any case, what is important for us is

[27] Indeed, Ryskina and Knight (2021) show that numerical embeddings can reliably predict the length of the numerals.

that a compositional principle inducing the order of numerals out of the order of its component digits seems to be available for a statistical analysis that takes into account our distributionally defined classes and operators (or our types and connectives).

Certainly, it doesn't seem reasonable to expect that an acceptable approximation of Benford's law, mapping precise probabilities into numerical values and vice-versa, can be achieved in this way. But the truth is that this is not strictly necessary. Because we are not here concerned with numerical values as such, which can very well be considered a high level emergent property of numerical systems. We are only concerned with total order. Benford's law provides a principled way to show that the order of numerals correlates with their probability, due to the principles of the notation system. But in its concrete form, it is only one of the many imaginable functions realizing such a mapping. Looking at the problem from a wider perspective, it is possible to see that Benford's law is only a particular case of a function ensuring the *lexicographical order* of compositions from the order of the components, any of which would perfectly fulfill the requirements for the construction of a total order suggested here.[28] Deriving the informational content of $\otimes$, and inducing with it a total order on N appears then much less challenging for current statistical models, including DNNs, than deriving Benford's laws. A simple local regression over the empirical probabilities of 2-digit characters could be enough to approximate such a function (cf. Figure 5b). The recursive structure of N, invested by the informational content of $\otimes$ would then suffice to guarantee the total order of all its members.

7 Conclusion

In this paper, we addressed the capacity of current deep neural machine learning models to deal with mathematical knowledge. Instead of following a widespread attitude in the field consisting in probing different capabilities of those models, we adopted a strictly philosophical approach, proposing a conceptual elaboration of the significance of those results from the perspective of the philosophy of mathematics and raising the question of the conditions of possibility of the epistemological claims implicitly or explicitly guiding the research in the field.

We thus suggested that, from a philosophical standpoint, the most original aspect of the application of current neural models to mathematical knowledge concerns the relation of mathematics to its own textuality, namely the fact that mathematical content is expected—and shown to a still modest but promising extent—to result from the analysis of pure expressions. When assessed at an epistemological level, this radical feature can become the occasion to rethink the nature of mathematical language in ways that are convergent with many current trends in the philosophy of mathematics. In particular, owing to the privileged use of neural language models, current machine learning applications to mathematics provide the opportunity to reassess the multiple connections between mathematical and natural language.

After exposing the challenges and possible resources for deriving mathematical content from a linguistic analysis of pure expressions, we showed that

[28] To the best of our knowledge, the signification of Benford's law with respect to all functions realizing lexicographical order is still to be studied.

the chances of success of such an undertaking are fundamentally tied to the foundations of distributional analysis, the critical assessment of which seems underestimated in the current state of the art. Therefore, in the last two sections of our paper, we engaged in a philosophical and conceptual elaboration of the elementary principles of distributionalism. We first focused on the relation of pure expressions to meaning, proposing a conception of sense as an effect, rooted, among others, in the different dimensions of a formal content (characteristic, syntactic and informational). Finally, armed with those concepts and relying on empirical data of a natural language corpus, we proposed a detailed illustration of how a principled distributional approach would be able to account for significant aspects of the content of numerical expressions in a way compatible with the original challenges identified in this paper.

The conceptual and sometimes highly speculative elaborations presented in this paper are not intended as an exact framework for the construction of yet another formal model but as philosophical orientations and rudiments for a thorough epistemological investigation of the new and often surprising line of research opened by the application of deep neural models to mathematical knowledge. When critically assessed—which in short means: when, instead of dwelling on metaphoric discussions about neural models, we take them for what they are, namely textual approaches to mathematical practices—then, the study and adoption of such models might turn out to be fruitful for the philosophical but also the historical and sociological investigation of mathematics. We hope these pages represent a first step in this direction.

References

Alexander A. Alemi, François Chollet, Niklas Een, Geoffrey Irving, Christian Szegedy, and Josef Urban. Deepmath - deep sequence models for premise selection. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2243–2251, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Jacob Andreas. Meanings and belief states, 2018. URL http://blog.jacobandreas.net/meaning-belief.html.

Jeremy Avigad. *Computers in Mathematical Inquiry*, chapter 11, pages 134–150. Oxford University Press, New York, 2008.

Jeremy Avigad. Mathematics and language, 2015. URL https://arxiv.org/abs/1505.07238.

Jeremy Avigad. The design of mathematical language, August 2021. URL http://philsci-archive.pitt.edu/19508/.

Kshitij Bansal, Sarah M. Loos, Markus N. Rabe, Christian Szegedy, and Stewart Wilcox. Holist: An environment for machine learning of higher-order theorem proving (extended version). *CoRR*, abs/1904.03241, 2019. URL http://arxiv.org/abs/1904.03241.

Yonatan Belinkov and James Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational*

Linguistics, 7:49–72, 04 2019. ISSN 2307-387X. doi: 10.1162/tacl_a_00254. URL https://doi.org/10.1162/tacl_a_00254.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://www.aclweb.org/anthology/2020.acl-main.463.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Frank Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938. ISSN 0003049X. URL http://www.jstor.org/stable/984802.

Jocelyn Benoist. *Husserl et Frege sur le concept*, pages 2003–224. Vrin, Paris, 2002.

Jan Blechschmidt and Oliver G. Ernst. Three ways to solve partial differential equations with neural networks — a review. *GAMM-Mitteilungen*, 44(2):e202100006, 2021. doi: https://doi.org/10.1002/gamm.202100006.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Jonathan M. Borwein and David H. Bailey. *Mathematics by experiment - plausible reasoning in the 21st century*. A K Peters, 2003. ISBN 978-1-56881-211-3.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin,

Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2 edition, 2022. doi: 10.1017/9781009089517.

François Charton. Linear algebra with transformers. *CoRR*, abs/2112.01898, 2021. URL https://arxiv.org/abs/2112.01898.

Karine Chemla. *The history of mathematical proof in ancient traditions*. Cambridge University Press, Cambridge, 2012.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198.

Stéphane d'Ascoli, Pierre-Alexandre Kamienny, Guillaume Lample, and François Charton. Deep symbolic regression for recurrent sequences. *CoRR*, abs/2201.04600, 2022.

Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis, and Pushmeet Kohli. Advancing mathematics by guiding human intuition with AI. *Nature*, 600(7887):70–74, December 2021. doi: 10.1038/s41586-021-04086-x. URL https://doi.org/10.1038/s41586-021-04086-x.

Ernest Davis. Deep learning and mathematical intuition: A review of (davies et al. 2021). *CoRR*, abs/2112.04324, 2021.

Gilles Deleuze. *The logic of sense*. Columbia University Press, New York, 1990.

Brenda Dervin. An overview of sense-making research: concepts, methods and results to date. International Ccmmunications Association Annual Meeting, 1983.

Brenda Dervin. From the mind's eye of the user: The sense-making qualitative-quantitative methodology. *Sense-making methodology reader*, 1986.

Brenda Dervin and Michael Nilan. Information needs and uses. *Annual review of information science and technology*, 21:3–33, 1986.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

Deborah Ferreira and André Freitas. STAR: Cross-modal [STA]tement [R]epresentation for selecting relevant mathematical premises. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3234–3243, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.282. URL https://aclanthology.org/2021.eacl-main.282.

John Rupert Firth. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford, 1957.

Frege. Methods of calculation based on an extension of the concept of quantity. In *Collected Papers on Mathematics, Logic, and Philosophy*, pages 56–92. Basil-Blackwell, 1874. ISBN 0-631-12728-3.

Gottlob Frege. *Conceptual Notation, and Related Articles.* Oxford University Press UK, Oxford, 1972.

Liangcai Gao, Zhuoren Jiang, Yue Yin, Ke Yuan, Zuoyu Yan, and Zhi Tang. Preliminary exploration of formula embedding for mathematical information retrieval: can mathematical formulae be embedded like a natural language? *CoRR*, abs/1707.05154, 2017. URL http://arxiv.org/abs/1707.05154.

Juan Luis Gastaldi. *Une archéologie de la logique du sens : arithmétique et contenu dans le processus de mathématisation de la logique au XIXe siècle.* Theses, Université Michel de Montaigne - Bordeaux III, September 2014. URL https://tel.archives-ouvertes.fr/tel-01174485.

Juan Luis Gastaldi. Frege's *Habilitationsschrift*: Magnitude, Number and the Problems of Computability. In Fabio Gadducci and Mirko Tavosanis, editors, *History and Philosophy of Computing*, pages 168–185, Cham, 2016. Springer International Publishing. ISBN 978-3-319-47286-7.

Juan Luis Gastaldi. Why can computers understand natural language?: The structuralist image of language behind word embeddings. *Philosophy & Technology*, 05 2020.

Juan Luis Gastaldi and Luc Pellissier. The calculus of language: Explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*, 2021. doi: 10.1080/03080188.2021.1890484.

Marcus Giaquinto. *Cognition of Structure*, chapter 2, pages 43–64. Oxford University Press, New York, 2008.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, Cambridge MA, London UK, 2016.

André Greiner-Petter, Terry Ruas, Moritz Schubotz, Akiko Aizawa, William I. Grosky, and Bela Gipp. Why machines cannot learn mathematics, yet. *CoRR*, abs/1905.08359, 2019. URL http://arxiv.org/abs/1905.08359.

André Greiner-Petter, Moritz Schubotz, Fabian Müller, Corinna Breitinger, Howard Cohl, Akiko Aizawa, and Bela Gipp. *Discovering Mathematical Objects of Interest—A Study of Mathematical Notations*, page 1445–1456.

Association for Computing Machinery, New York, NY, USA, 2020a. ISBN 9781450370233. URL https://doi.org/10.1145/3366423.3380218.

André Greiner-Petter, Abdou Youssef, Terry Ruas, Bruce R. Miller, Moritz Schubotz, Akiko Aizawa, and Bela Gipp. Math-word embedding in math search and semantic extraction. 125(3):3017–3046, 2020b. ISSN 1588-2861. doi: 10.1007/s11192-020-03502-9. URL https://doi.org/10.1007/s11192-020-03502-9.

Zellig Harris. *Structural linguistics*. University of Chicago Press, Chicago, 1960. ISBN 0226317714 0226217714.

Alain Herreman. *La topologie et ses signes: éléments pour une histoire sémiotique des mathématiques*. L'Harmattan, Paris, 2000.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks, 2019.

Louis Hjelmslev. *Prolegomena to a Theory of Language*. Wawerly Press, Baltimore, 1953.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi: 10.1016/0893-6080(89)90020-8.

Edmund Husserl. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy: First Book: General Introduction to a Pure Phenomenology (Husserliana: Edmund Husserl – Collected Works, 2)*. Springer, paperback edition, 1983.

Cezary Kaliszyk, François Chollet, and Christian Szegedy. Holstep: A machine learning dataset for higher-order logic theorem proving. *CoRR*, abs/1703.00426, 2017. URL http://arxiv.org/abs/1703.00426.

Andrea Kohlhase, Michael Kohlhase, and Taweechai Ouypornkochagorn. Discourse phenomena in mathematical documents. In Florian Rabe, William M. Farmer, Grant O. Passmore, and Abdou Youssef, editors, *Intelligent Computer Mathematics*, pages 147–163. Springer International Publishing, 2018.

Kriste Krstovski and David M. Blei. Equation embeddings, 2018.

Guillaume Lample and François Charton. Deep learning for symbolic mathematics, 2019.

Lee T. Lemon and Marion J. Reis, editors. *Russian Formalist Criticism: Four Essays, Second Edition (Regents Critics)*. University of Nebraska Press, 2012. ISBN 978-0803239982.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *From context to meaning: distributional models of the lexicon in linguistics and cognitive science, Italian Journal of Linguistics*, 1(20):1–31, 2008.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, pages 171–180, 2014.

Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16 (3):31–57, jun 2018. ISSN 1542-7730. doi: 10.1145/3236386.3241340. URL https://doi.org/10.1145/3236386.3241340.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey, 2021. URL https://arxiv.org/abs/2108.04840.

Paolo Mancosu. *Mathematical Explanation: Why it Matters*, chapter 5, pages 134–150. Oxford University Press, New York, 2008.

Christopher D. Manning. Computational Linguistics and Deep Learning. *Computational Linguistics*, 41(4):701–707, 12 2015. ISSN 0891-2017. doi: 10.1162/COLI_a_00239. URL https://doi.org/10.1162/COLI_a_00239.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907367117.

Behrooz Mansouri, Shaurya Rohatgi, Douglas W. Oard, Jian Wu, C. Lee Giles, and Richard Zanibbi. Tangent-cft: An embedding model for mathematical formulas. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 11–18, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368810. doi: 10.1145/3341981.3344235. URL https://doi.org/10.1145/3341981.3344235.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060, 2021. doi: 10.1162/tacl_a_00412. URL https://aclanthology.org/2021.tacl-1.62.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Steven J. Miller. *Benford's Law: Theory and Applications*. Princeton University Press, 2015. doi: 10.23943/princeton/9780691147611.001.0001.

Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. Exploring numeracy in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3374–3380, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1329. URL https://aclanthology.org/P19-1329.

Nikita Nangia and Samuel Bowman. ListOps: A diagnostic dataset for latent tree learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 92–99, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-4013. URL https://aclanthology.org/N18-4013.

Reviel Netz. *The Shaping of Deduction in Greek Mathematics*. Cambridge University Press, 1999.

Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4(1):39–40, 1881. ISSN 00029327, 10806377.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. *CoRR*, abs/2003.06713, 2020.

Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. Mathbert: A pre-trained model for mathematical formula understanding. *CoRR*, abs/2105.00377, 2021. URL https://arxiv.org/abs/2105.00377.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

Mohammad Taher Pilehvar and Jose Camacho-Collados. Embeddings in natural language processing: Theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, 13(4): 1–175, November 2020. doi: 10.2200/s01057ed1v01y202009hlt047. URL https://doi.org/10.2200/s01057ed1v01y202009hlt047.

Joëlle Proust. Bolzano's theory of representation. *Revue d'histoire des sciences*, 52(3/4):363–383, 1999. ISSN 01514105, 19696582. URL http://www.jstor.org/stable/23633692.

Stanisław Purgał, Julian Parsert, and Cezary Kaliszyk. A study of continuous vector representations for theorem proving. *Journal of Logic and Computation*, 31(8):2057–2083, 02 2021. ISSN 0955-792X. doi: 10.1093/logcom/exab006. URL https://doi.org/10.1093/logcom/exab006.

Willard Van Orman Quine. *Word and Object, new edition (The MIT Press)*. The MIT Press, paperback edition, 2013. ISBN 978-0262518314.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.15. URL https://aclanthology.org/2021.conll-1.15.

Maria Ryskina and Kevin Knight. Learning mathematical properties of integers. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 389–395, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.blackboxnlp-1.30. URL https://aclanthology.org/2021.blackboxnlp-1.30.

Magnus Sahlgren. The distributional hypothesis. *Special issue of the Italian Journal of Linguistics*, 1(20):33–53, 2008.

Ferdinand de Saussure. *Course in General Linguistics*. McGraw-Hill, New York, 1959. Translated by Wade Baskin.

Dirk Schlimm. Numbers through numerals. the constitutive role of external representations. In Sorin Bangu, editor, *Naturalizing Logico-Mathematical Knowledge: Approaches from Psychology and Cognitive Science*, pages 195–217. 2018.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 1715–1725, Berlin, Germany, August 2016. ACL.

Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil T. Heffernan, Xintao Wu, and Dongwon Lee. Mathbert: A pre-trained language model for general NLP tasks in mathematics education. *CoRR*, abs/2106.07340, 2021.

Wilfried Sieg and Dirk Schlimm. Dedekind's analysis of number: Systems and axioms. 147(1):121–170, 2005. ISSN 1573-0964. doi: 10.1007/s11229-004-6300-9. URL https://doi.org/10.1007/s11229-004-6300-9.

Henrik Kragh Sørensen and Mikkel Willum Johansen. Counting mathematical diagrams with machine learning. In Ahti-Veikko Pietarinen, Peter Chapman, Leonie Bosveld-de Smet, Valeria Giardino, James Corter, and Sven Linker, editors, *Diagrammatic Representation and Inference*, pages 26–33. Springer International Publishing, 2020.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. Representing numbers in NLP: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.53. URL https://aclanthology.org/2021.naacl-main.53.

Silvia De Toffoli and Valeria Giardino. Forms and roles of diagrams in knot theory. *Erkenntnis*, 79(4):829–842, November 2013. doi: 10.1007/s10670-013-9568-7. URL https://doi.org/10.1007/s10670-013-9568-7.

Marcus Tomalin. Syntactic structures and recursive devices: A legacy of imprecision. 20(3):297, 2011. ISSN 1572-9583. doi: 10.1007/s10849-011-9141-1. URL https://doi.org/10.1007/s10849-011-9141-1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Roy Wagner. *Does Mathematics Need Foundations?*, pages 381–396. Springer International Publishing, Cham, 2019. ISBN 978-3-030-15655-8. doi: 10.1007/978-3-030-15655-8_17.

David Waszek. *Les représentations en mathématiques*. PhD thesis, 2018. URL http://www.theses.fr/2018PA01H231. Thèse de doctorat dirigée par Panza, Marco Philosophie Paris 1 2018.

Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. Naturalproofs: Mathematical theorem proving in natural language. *CoRR*, abs/2104.01112, 2021. URL https://arxiv.org/abs/2104.01112.

G. Whyman, E. Shulzinger, and Ed. Bormashenko. Intuitive considerations clarifying the origin and applicability of the benford law. *Results in Physics*, 6:3 – 6, 2016. ISSN 2211-3797.

Wikimedia Foundation. Wikimedia downloads. URL https://dumps.wikimedia.org.

Ludwig Wittgenstein. *Philosophical investigations*. Wiley-Blackwell, Chichester, West Sussex, U.K. ;, 4th ed. edition, 2009. ISBN 1-282-55045-4.

Thomas Wolf. Learning Meaning in Natural Language Processing - The Semantics Mega-Thread, 2018. URL https://medium.com/huggingface/learning-meaning-in-natural-language-processing-the-semantics-mega-thread-9c0332dfe28e.

Yanyan Zou and Wei Lu. Text2Math: End-to-end parsing text into math expressions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5327–5337, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1536. URL https://aclanthology.org/D19-1536.