doi:10.20944/preprints202109.0062.v1

Article

Discover AI knowledge to preserve Cultural Heritage

Leonardo Ranaldi [1]*, Fabio Massimo Zanzotto [2]

[1] Department of Innovation and Information Engineering Guglielmo Marconi University; l.ranaldi@unimarconi.com
[2] Department of Enterprise Engineering University of Rome Tor Vergata; fabio.massimo.zanzotto@uniroma2.it
* Correspondence: l.ranaldi@unimarconi.it;

Abstract: Documenting cultural heritage by using artificial intelligence (AI) is crucial for preserving the memory of the past and a key point for future knowledge. However, modern AI technologies make use of statistical learners that lead to self-empiricist logic, which, unlike human minds, use learned non-symbolic representations. Nevertheless, it seems that it is not the right way to progress in AI. If we want to rely on AI for these tasks, it is essential to understand what lies behind these models. Among the ways to discover AI there are the senses and the intellect. We could consider AI as an intelligence. Intelligence has an essence, but we do not know whether it can be considered *"something"* or *"someone"*. Important issues in the analysis of AI concern the structure of symbols -operations with which the intellectual solution is carried out- and the search for strategic reference points, aspiring to create models with human-like intelligence. For many years, humans, seeing language as innate, have carried out symbolic theories. Everything seems to have skipped with the advent of Machine Learning. In this paper, after a long analysis of history, the rule-based and the learning-based vision, we propose KERMIT[1] as a unit of investigation for a possible meeting point between the different learning theories. Finally, we propose a new vision of knowledge in AI models based on a combination of rules, learning and human knowledge.

Keywords: Machine Learning; Natural Language Processing; Deep Learning

1. Introduction

For centuries, humanist scholars have contributed to the preservation of memory and knowledge of the past [2,3] by analyzing symbolic documents with symbolic minds and, today, they are seeking the help of artificial intelligence (AI) to speed up their analyses. Currently, AI "minds" are dominated by non-symbolic, obscure statistical learners, which have cancelled symbols in their controlling strategies. Humanist scholars can hardly control these statistical models by using their knowledge expressed with symbols.

There are then two theories of how to manipulate knwoledge: the current **empiric** trend in AI and the **nativist** theory of Chomsky[4–6]. The first ones, the empiricist models, do not pay attention to form such as the difference between a noun or a verb because they do not have a built-in symbolic base structure. A different philosophy for nativists, who emphasise form in order to appreciate substance. Nativists claim that there are innate structures from the moment of birth. The current trend of AI seems to agree with the empiricists, who are building performing models. An example of an empiricist model is the Transformers-models[7] which has no specific a priori knowledge about space, time or objects other than what is represented in the training corpus. This basic idea is precisely the antithesis of Chomsky who, through universal grammars [4], provided an explanation for the phenomenon of innate linguistics.

However, the idea that thoughts and sentences can be represented as vectors, flat and meaningless, rather than as complex symbolic structures such as syntactic trees [4] makes the Transformers-models a very good tester of the empiricist hypothesis.

Recently, many works have shown that knowledge acquired only from experience, as is done among the Transformers-models, is superficial and unreliable [8,9] and they adhere to hooks that emerge from the statistical nature of the model [10]. Nevertheless this is nothing new, since had already been pointed out by Lake and Baroni in [11]. They argued that RNNs, which are after all the parents of Transformers-models, generalize well discretely when differences between training set and test set are small but if generalization requires systematic compositional skills, RNNs fail spectacularly. Rather than supporting the empiricist view, Transformers-models seem to be an incidental counter-evidence to it. Summing up, it does not seem like great news for the symbol-free thought-vector view [12]. Vector-based systems can predict word categories, but they do not embody thoughts in a sufficiently reliable way. As far as the underlying symbolic structures are concerned, they seem to be captured in a discrete but imperfect way [13,14]. Transformers-models are a triumph for empiricism and in light of the massive resources of data and computation that have been poured into them. However, this is a clear sign that it is time to dust off old approaches.

With the aim of studing these phenomena, we propose KERMIT*system* [1] (Kernel-inspired Encoder with Recursive Mechanism for Interpretable Trees) as a meeting point between the two theories of knowledge. The purpose is to embed the long symbolic-syntactic history in the modern Transformer architecture. In order to investigate the origin of knowledge and the achievable performance, a long reflection on the role of knowledge in artificial intelligence is made during this work.

The contributions of our paper are as follows: firstly, we will try to give *"form"* to the origins and development of artificial intelligence. (sec.2). After investigating the long history, KERMIT*system* will be introduced as a unit of investigation (sec.3). Finally, limitations, weaknesses and future developments will be highlighted in a wide-ranging discussion (sec.4).

2. Related Works

In the last few years, there has been an increasing interest in the digitization of Cultural Heritage collections. Many public and private digitization campaigns have made billions of documents accessible by using online tools. This gives rise to new tools for the end-user of this data where the artificial intelligence (AI) domain plays a key role. Indeed, volume and size of historical data leads to critical factors [2]. The most important one concerns knowledge extraction and inference. To solve this problem the application of Machine Learning (ML) gives the opportunity to improve historical archives and heritage information processing [3]. Although very accurate and performing, ML-based models fail to bring with them a long history of linguistic theory and the symbols are merely meaningless numbers. In order to preserve the richness and uniqueness of the symbols transmitted over a long period of time, it is important to trace the path and ideas behind symbolic minds and data abstraction. Modern currents of thought have classical origins, with both nativist and empiricist theories going back to Plato and Aristotle.

Plato, the father of nativism, is at the origin of innateness, and most of his works deal with the theory of inborn knowledge. More specifically, the Athenian philosopher showed his thinking in a dialogue called Meno. By posing a mathematical puzzle (known as The Learner's Paradox) to a slave who does not know the principles of geometry, Plato demonstrated that concepts and ideas are present in the human mind before birth. Therefore, *"seeking and learning are in fact nothing but recollection"*. Plato tried to explain the innactivity of knowledge in the human brain by defining it as a *"receptacle of all that comes to be"* [15]. In this space, matter takes form and symbols take on meaning thanks to the ideas and thoughts innately embedded in the human brain.

Centuries later, Noam Chomsky, studying symbols in the form of linguistic phenomena and following nativist theories, argued that human children could not acquire human language unless they were born with a "language acquisition device" [4], or what Steven Pinker [16] called a "language instinct". From this school of thought derive the theories for symbolic syntactic interpretations, Chomsky grammars [5,6]. Over the years, theories that confirm the universal structural basis of language [17] have evolved and provided increasingly advanced models, like rule-based and statistical parsers [18,19]. Until a few years ago these were the only means of working with language. These tools required a lot of manpower, tokenization, lemmatization and a series of repetitive and low-value additional operations performed by humans. These approaches bring success but require that the modelling is well done and requires a lot of time and resources [20].

On the other side of the coin, Aristotle in the "Physics" revised the ideas of his mentor Plato on the difference between "matter" and "form" [21]. Aristotle broke down the theories

doi:10.20944/preprints202109.0062.v1

of innatism and in later works defined theories based on experience. If we think of Aristotle's "Logic" [22] we find a closeness to the ideas of today AI, because he interpreted the state that the study of thought is the basis of knowledge. In fact, Aristotle, thinking about the processes of forming concrete proofs, developed a non-formalised system of syllogisms and used them in the design of proof procedures. Aristotle's ideas were the founding pillars for studying the formal axiomatization of logical reasoning, which, added to a *"tabula-rasa"* knowledge, allows the human being to think, and it can be seen as a physical system, a precursor idea of the ML. In this philosophy of thought, matter takes shape through experience, so symbols take on meaning after a process of acquisition. These ideas have been handed down over the centuries through the thought of Roger Bacon, Thomas Aquinas, John Locke and Immanuel Kant. In modern times, the developmental psychologist Elizabeth Bates [23] and the cognitive scientist Jeff Elman [24]. The thought that unites them is something like a *"tabula-rasa"*, for which our knowledge comes from experience, provided through the senses, arguments reminiscent of the ML approach. With the aim of making the text computable [25] assumed to encode the presence of the text by counting the occurrences. This approach turned out to be very superficial; in fact, some time later [26–28], using purpose-built neural networks and large corpora made large distributed representations of the text. Recently, these representations have been built with Transformers-models [7]. Architectures such as BERT[29] and GPT[30] are based on several layers of encoders or decoders. In BERT, there are appropriate mechanisms aimed at learning universal language representations that are independent of the task. Although these models achieve extraordinary results, knowledge from experience alone seems to be not enough. The statistical learners are very good students as long as we talk about superficial and "simple" tasks. However, when the bar is raised and the task becomes more difficult, the inability of the statistical learners emerges [10]. Moreover, it seems that the knowledge acquired by the Transformers-models is superficial and unreliable [8,9].

To briefly summarise, there seems to be a huge gap between nativist and empiricist theory. The gap between nativists and empiricists is also transmitted in the representation of the world because while the former make strong use of symbols, the latter use dense vectors. In the field of representation, this gap is nothing new, as also Zanzotto et al. [31] describes a clear division between the symbols that were used for the older representations and the numbers and vectors that are used for the newer representations.

Actually, we will find out that it is not entirely true, because the gap is not so huge. In fact, it seems that human beings have an innate mechanism, ready to adapt, as we will see in sec.3.3 and consequently also representations are not so radical and firm as several representations can coexist at the same time.

In order to test this hypotheses, we propose KERMIT*system* [1]. In fact, KERMIT*system* could unite the two great visions of language development and knowledge. In section 3, after extensively reviewing the theories on the origin of knowledge, knowledge in the form of function will be proposed (3.4) and KERMIT*system* will be studied as a unit of enquiry (3.5) for hybrid knowledge between symbols and experience.

3. Our point of view

Digitalization techniques are degrading the uniqueness of cultural works. Although the ongoing analysis and study of artworks seems to be more accurate by discovering new links, the same does not apply to the richness and uniqueness of the work. Before the advent of AI the only means of study was the human mind dominated by symbols while with modern AI technologies everything has been cancelled out. To investigate the origins and future developments of Machine Learning and Natural Language Processing applied in the study of cultural heritage works, the following sections will examine the nativist and empiricist viewpoints, the representations through which the theories manifest themselves and a point of intersection.

3.1. Innatness

There is much evidence to suggest the presence of innatness. Despite languages vary, they share many universal structural properties [4,17]. For this reason we asked whether we are heading in the right direction or whether there is a need for innate mechanisms. To answer this question we went to dust off the ancient linguistic theories. The theories of innateness were started, unconsciously, by Plato and transmitted through the centuries to the present day with the studies of Chomsky [4,5]. Chomsky in his studies claimed that there are universal grammars unique to everyone [4] and humans at birth are endowed with a kind of innate machine already initialized [5]. Therefore, to mitigate the hard view of innateness, humans - unlike animals - are predisposed to learn, starting with the innate mechanism with which they are endowed from birth. Even if there is a part of experience in the processing of linguistic knowledge, the underlying structures can be defined by recurrent and universal patterns [32].

In the early years of the 2000s, these theories were advanced by Cristianini et al. [33], Moschitti [34] and Collins et al. [35]. In order to understand the relationships between the underlying structures, they focused on the structures by working on kernel functions capable of counting common subtrees in order to extrapolate similarities. While long studies on syntactic theories have led to the production of very good representations that are quite light from a computational point of view but unfortunately not ready because they are not computable. In the years that followed, a number of methodologies have been developed to make syntactic structures computable [36–38].

3.2. No-Innatness

On the other side of the coin, there is the theory derived from the thoughts of Aristotle, who can be classed as a *"tabula rasa"* empiricist, for he rejects the claim that we have innate ideas or principles of reasoning. These principles were widely shared by supporters of statistical learners.

Nowadays, we can see what is happening with the Transformers-models [7], they are only based on experience and seem to be achieving state-of-the-art results in many downstream tasks. Transformers-models rely only on the knowledge derived from the experience they learned on huge corpus. This way of working follows the empiricist theories widely studied since Aristotle. Thus, the Transformers-models are the proven proof that these theories can work, as they claim that knowledge can be constructed by experience alone. All this is great if and only if things work out.

Unfortunately, this is not always the case, considering that these architectures, although very good learners, tend to adapt very much to shallow heuristics [10]. The knowledge they learn is very superficial [8,9], in fact, in hostile contexts, where an acception can totally change the meaning of a sentence, they do not work; however, they perform well in long contexts, only in the presence of many resources[39]. These things stem from the fact that working with these means the only important thing is to maximize or minimize a cost function. Thus, with Transformers-models we have a representation ready to be used by neural networks, it seems to encode syntactic and semantic information [14,40] but unfortunately in unclear ways and are very computationally expensive.

3.3. The truth lies in the middle

All thinkers of every time would recognise that genius and experience are not separate but work together. As the nativists, empiricists would not doubt that we were born with a specific biological machinery that enables us to learn. Indeed, Chomsky's famous "Language Acquisition Device"[4], in another view, should be seen precisely as an **innate learning mechanism**. At the end of the day, it has been supported the stance for which a significant part of the innate knowledge provided to humans consists of learning mechanisms, that is a form of innateness that enables learning [16,41,42]. From these insights it is clear how the two theories are related in human minds.

An interesting question relates to the basic argument of this article: do artificially intelligent systems need to be equipped with significant amounts of innate machinery, or is it sufficient, given the powerful Machine Learning systems recently developed, for such systems to work in *"tabula-rasa"* mode. An answer to this question has been investigated for years by researchers in various fields: from psychology to neuroscience as well as traditional linguistics and today's natural language processing.

doi:10.20944/preprints202109.0062.v1

There are studies that claim that children are endowed early in life with a "knowledge base", as stated by Spelke [43]. Indeed, children have some ability to trace and reason about objects, which is unlikely to arise through associative learning. It seems that the brain of an 8-month-old child can learn and identify abstract syntactic rules with only a few minutes of exposure to artificial grammars [44]. Other work has suggested that deaf children can invent language without any direct model [45], and that language can be selectively impaired even in children with normal cognitive function [46].

Humans are not precise machines, so many questions remain unanswered. We do not know if at the moment of birth we are equipped with Chomsky's machine or if we have a machine set up for learning language. On the other hand, even if we could demonstrate the evidence of innatism, it would not mean that there is no learning.

Probably the presence of innatism means that any learning takes place against a background of certain mechanisms that precede learning. On the other hand, can the *"innate machinery"* be an ingredient for a human-like artificial intelligence?

3.4. Knowledge as function

There is much evidence to suggest the presence of empiricism or nativism, but there is also evidence to suggest that it is only the starting point for the development of knowledge and actually there is a middle way between the two streams of thought as mentioned in sec. 3.3.

In order to understand the ingredients in depth we can define the knowledge as a function:

$$K = f(k, e, r, m, i, t) \tag{1}$$

In function 1 there is a kind of background knowledge, called initialization knowledge denoted by k, in contrast, e denoting knowledge derived from experience. Regarding r and t are the representation of the innate part and the representation non innate part. Finally, m indicates the mechanisms and the underlying algorithms used to arrive at K.

In a tabula rasa radical scenario, it would set k,i, t and r to zero, set a to some extremely minimal value (e.g., an operation for adjusting weights relative to reinforcement signals), and leave the rest to experience. This very radical view is also held by the father of deep learning in LeCun et al. [12] and is applied in Transformers[7]. The same thing but with exchanged variables could be argued by following Chomsky's ideas. In this view the variables r and e are almost zero.

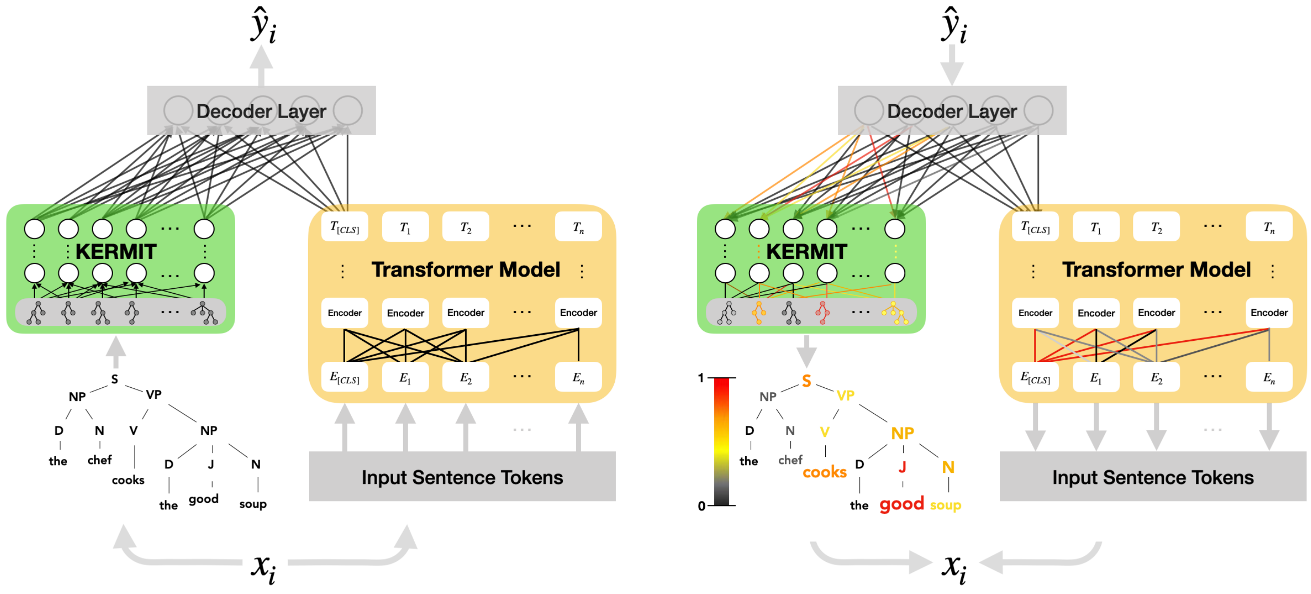
What AI researchers in general think about the values of k,e,r,m,i and t, is largely unknown, in part because few researchers ever explicitly discuss the question, probably because people get wrapped up in the practicalities of their immediate research, and do not really tend to think about nativism at all, as claim Marcus [42].

Many researchers believe that methods for incorporating prior knowledge in the form of symbolic rules are too cumbersome and, while they are very useful from an engineering point of view, do not contribute to a plausible theory of general intelligence. However, there are some conditions that delineate the boundary. The *"No Free Lunch"* theorem [47,48] effectively shows that the above variables cannot be absolute zero. Each system will generalise in different ways, depending on which initial algorithm is specified, and no algorithm is uniquely the best. This idea ties in with what was said in section 3.3 , because although on paper the scenario looks radical enough to set some variables of the function 1 to zero in practice this is not exactly. Although it seems that no variable is truly absolute zero, at the same time it is rare to find an architecture in which all these variables coexist at the same time.

3.5. KERMIT

For this reason, it would be interesting to understand and experience all these variables combined in one system. An answer can be found in *KEMRIT* $\bowtie$ *Transformers*. KERMIT(Kernel inspired Encoder with Recursive Mechanism for Interpretable Trees) [1], combined $\bowtie$, with the powerful Transformer [7] architecture.

This system is the trade-off between long linguistic theory and modern universal sentence embedding representations. By using this framework, it is possible to combine these two theories on the origin of linguistic knowledge. Equation 1 is satisfied and all variables are given the same



(a) Forward pass

(b) Interpretation pass

Figure 1. The *KEMRIT* ⋈ *Transformers* architecture. During the interpretation pass KERMIT*viz* is used to produce *heat parse trees*, while a transformer's activation visualizer is used for the remainder of the network.

weight. Basic knowledge, denoted by k, is provided by the symbolic part. Experience, denoted by e, comes from the part. The representation, denoted by r, comes directly from the encoding of experience. While the symbolic representation is obtained thanks to the underlying mechanisms allow the encoding of the tree representation t in a computable form.

Fig.1 shows how *KEMRIT* ⋈ *Transformers* works. Specifically, in subfigure1a we can observe the input of *KEMRIT* ⋈ *Transformers* in symbolic and in Transformers-models form, the input is then joined to a special decoder layer denoted with $\hat{y}_i$. In subfigure1b, another mechanism can be seen that makes it possible to observe the most activated parts of the symbolic input, i.e. the parts that contributed most to the classifier's final decision. The provided source code [1] enables users to test the system on all classification tasks because it provides the architecture ready to do training and testing phase.

To support the hypothesis of the union of the two knowledge we can look at the results obtained in the official paper [1]. *KEMRIT* ⋈ *Transformers* has been tested in different configurations. For the Transformers-models part, BERT[29] was used. BERT and KERMIT are only used as encoders without training the internal weights. This choice was made in order to fix the part derived from innate knowledge and the part derived from experience without building up further experience at learning-time. Finally, the two parts have been merged into one layer which chooses the most important.

In *KEMRIT* ⋈ *Transformers*, we can see how the nativist and empiricist aspects can work together to produce new, holistic and enriched forms of knowledge. Despite the excellent results, important questions remain. One crucial question concerns finding out how much knowledge is required by the AI of the future and how much must already be encoded and how much must be left to be learned.

3.6. Measuring knowledge

After defining the basic ingredients for knowledge through the Equation 1 and after identifying *KEMRIT* ⋈ *Transformers* as a possible complete and balanced embodiment of the function, important questions arose: how much knowledge must be present, how much must be entered by humans and how much must be left to be learned by a system?

[1] The code is available at https://github.com/ART-Group-it/KERMIT

To answer these questions we propose an approach to discern how much innateness might be required for AI. In order to find an answer we would be to create synthetic agents that do difficult tasks, with some initial degree of innateness, achieve state of the art performance with those tasks, and then iterate, reducing as much innateness as possible, ultimately converging on some minimal amount of innate machinery. This strategy is close to the one proposed by Silver et al. [49] and taken up by Marcus [42].

In order to carry out this test, we need to build machines that are partially innate and therefore, able to learn from experience, as machine learning paradigm, but at the same time act by human hand. This architecture should be similar to Pat-in-the-Loop[50], a system that allows humans to input rules into a neural network. The results must be interpretable, so humans must be able to understand the decisions made by the system. In order to guarantee the explainability of the *intereptation pass* (see Fig.1b) present in the KERMIT architecture introduced in the previous section could be used.

This could preserve the uniqueness of cultural works such as an image or a portion of text, which is being lost lately.

4. Conclusions

Studies on the documentation of Cultural Heritage play an important role both in preserving memory and in feeding future knowledge from the past. Over the centuries, symbolic documents have been extensively studied by humanist scholars with their symbolic minds. Today, humanist scholars are seeking the help of artificial intelligence (AI) to accelerate their analyses. Unfortunately, current AI technologies use non-symbolic representations representations learned in obscure and complex ways.

After a long view on the computational application of nativist and empiricist theories, we defined knowledge in the form of a function from which we proposed KERMIT[1] as a point of intersection between symbolic theories and recent statistical learning.

Finally, we concluded this work by proposing an innateness test. This test will enable researchers to investigate the amount of innateness needed to achieve better performance on a task. Furthermore, once you have quantified the necessary innateness you can start testing on KERMIT and no longer just by using these two parts as encoders, but by using them with learnable weights.

The point of this paper is that, on the one hand, the issue is difficult to solve, and on the other, the balance between the two approaches has become, across the whole field of machine learning, seriously distorted. It is time for AI to take nativism more seriously.

References

1. Zanzotto, F.M.; Santilli, A.; Ranaldi, L.; Onorati, D.; Tommasino, P.; Fallucchi, F. KERMIT: Complementing Transformer Architectures with Encoders of Explicit Syntactic Interpretations. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Association for Computational Linguistics: Online, 2020; pp. 256–267. doi:10.18653/v1/2020.emnlp-main.18.
2. Condorelli, F.; Rinaudo, F.; Salvadore, F.; Tagliaventi, S. A Neural Networks Approach to Detecting Lost Heritage in Historical Video. *ISPRS International Journal of Geo-Information* **2020**, *9*. doi:10.3390/ijgi9050297.
3. Fiorucci, M.; Khoroshiltseva, M.; Pontil, M.; Traviglia, A.; Del Bue, A.; James, S. Machine Learning for Cultural Heritage: A Survey. *Pattern Recognition Letters* **2020**, *133*, 102–108. doi:https://doi.org/10.1016/j.patrec.2020.02.017.
4. Chomsky, N. *Aspects of the Theory of Syntax*; The MIT Press: Cambridge, 1965.
5. Chomsky, N. On certain formal properties of grammars. *Information and Control* **1959**, *2*, 137–167. doi:https://doi.org/10.1016/S0019-9958(59)90362-6.
6. Chomsky, N. *Syntactic Structures*; Mouton, 1957.
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need, 2017, [arXiv:cs.CL/1706.03762].
8. Sinha, K.; Sodhani, S.; Dong, J.; Pineau, J.; Hamilton, W.L. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text, 2019, [arXiv:cs.LG/1908.06177].
9. Talmor, A.; Elazar, Y.; Goldberg, Y.; Berant, J. oLMpics – On what Language Model Pre-training Captures, 2020, [arXiv:cs.CL/1912.13283].

10. McCoy, T.; Pavlick, E.; Linzen, T. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics: Florence, Italy, 2019; pp. 3428–3448. doi:10.18653/v1/P19-1334.
11. Lake, B.M.; Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks, 2018, [arXiv:cs.CL/1711.00350].
12. LeCun, Y.; Bengio, Y.; Hinton, G.E. Deep learning. *Nat.* **2015**, *521*, 436–444. doi:10.1038/nature14539.
13. Goldberg, Y. Assessing BERT's Syntactic Abilities, 2019, [arXiv:cs.CL/1901.05287].
14. Hewitt, J.; Manning, C.D. A Structural Probe for Finding Syntax in Word Representations. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 4129–4138. doi:10.18653/v1/N19-1419.
15. Zeyl, D.; Sattler, B. Plato's Timaeus. In *The Stanford Encyclopedia of Philosophy*, Summer 2019 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University, 2019.
16. Pinker, S.; Jackendoff, R. The faculty of language: What's special about it? *Cognition* **2005**, *95*, 201–36. doi:10.1016/j.cognition.2004.08.004.
17. Newmeyer, F.J. Explaining language universals. *Journal of Linguistics* **1990**, *26*, 203–222. doi:10.1017/S002222670001450X.
18. Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing*; The MIT Press: Cambridge, Massachusetts, 1999.
19. Collins, M. Head-Driven Statistical Models for Natural Language Parsing. *Computational Linguistics* **2003**, *29*, 589–637. doi:10.1162/089120103322753356.
20. Settles, B.; Craven, M.; Friedland, L.A. Active Learning with Real Annotation Costs. 2008.
21. Bodnar, I. Aristotle's Natural Philosophy. In *The Stanford Encyclopedia of Philosophy*, Spring 2018 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University, 2018.
22. Smith, R. Aristotle's Logic. In *The Stanford Encyclopedia of Philosophy*, Fall 2020 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University, 2020.
23. Bates, E.; Thal, D.; Finlay, B.; Clancy, B. Early Language Development And Its Neural Correlates. *Handbook of Neuropsychology* **1970**, *Vol. 6*.
24. Elman, J.L.; Bates, E.A.; Johnson, M.H.; Karmiloff-Smith, A.; Parisi, D.; Plunkett, K. *Rethinking Innateness: A Connectionist Perspective on Development*; MIT Press, 1996.
25. Salton, G. Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley* **1989**, *169*.
26. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space, 2013, [arXiv:cs.CL/1301.3781].
27. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1532–1543. doi:10.3115/v1/D14-1162.
28. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* **2016**.
29. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019, [arXiv:cs.CL/1810.04805].
30. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models are Few-Shot Learners, 2020, [arXiv:cs.CL/2005.14165].
31. Ferrone, L.; Zanzotto, F.M. Symbolic, Distributed, and Distributional Representations for Natural Language Processing in the Era of Deep Learning: A Survey. *Frontiers in Robotics and AI* **2020**, *6*, 153. doi:10.3389/frobt.2019.00153.
32. White, L. Second Language Acquisition and Universal Grammar. *Studies in Second Language Acquisition* **1990**, *12*, 121–133. doi:10.1017/S0272263100009049.
33. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; Cambridge University Press, 2000.
34. Moschitti, A. Making Tree Kernels practical for Natural Language Learning. In *Proceedings of EACL'06*; 2006.

35. Collins, M.; Duffy, N. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. Proceedings of {ACL}02, 2002.
36. Culotta, A.; Sorensen, J. Dependency Tree Kernels for Relation Extraction. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics; Association for Computational Linguistics: USA, 2004; ACL '04, p. 423–es. doi:10.3115/1218955.1219009.
37. Pighin, D.; Moschitti, A. On Reverse Feature Engineering of Syntactic Tree Kernels. Proceedings of the Fourteenth Conference on Computational Natural Language Learning; Association for Computational Linguistics: Uppsala, Sweden, 2010; pp. 223–233.
38. Zanzotto, F.M.; Dell'Arciprete, L. Distributed Tree Kernels, 2012, [arXiv:cs.LG/1206.4607].
39. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context, 2019, [arXiv:cs.LG/1901.02860].
40. Puccetti, G.; Miaschi, A.; Dell'Orletta, F. How Do BERT Embeddings Organize Linguistic Knowledge? Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Association for Computational Linguistics: Online, 2021; pp. 48–57.
41. Marler, P. Innateness and the instinct to learn. *Anais da Academia Brasileira de Ciências* **2004**, *76*, 189–200. doi:10.1590/S0001-37652004000200002.
42. Marcus, G. Innateness, AlphaZero, and Artificial Intelligence, 2018, [arXiv:cs.AI/1801.05667].
43. Spelke, E.S.; Kinzler, K.D. Core knowledge. *Developmental Science* **2007**, *10*, 89–96.
44. Gervain, J.; Berent, I.; Werker, J.F. Binding at Birth: The Newborn Brain Detects Identity Relations and Sequential Position in Speech. *J. Cognitive Neuroscience* **2012**, *24*, 564–574. doi:10.1162/jocn_a_00157.
45. Senghas, A.; Kita, S.; Özyürek, A. Children Creating Core Properties of Language: Evidence from an Emerging Sign Language in Nicaragua. *Science (New York, N.Y.)* **2004**, *305*, 1779–82. doi:10.1126/science.1100199.
46. Lely, H.; Pinker, S. The biological basis of language: Insight from developmental grammatical impairments. *Trends in Cognitive Sciences* **2014**, *18*. doi:10.1016/j.tics.2014.07.001.
47. Geman, S.; Bienenstock, E.; Doursat, R. Neural Networks and the Bias/Variance Dilemma. *Neural Computation* **1992**, *4*, 1–58. doi:10.1162/neco.1992.4.1.1.
48. Wolpert, D.H. The Lack of a Priori Distinctions between Learning Algorithms. *Neural Comput.* **1996**, *8*, 1341–1390. doi:10.1162/neco.1996.8.7.1341.
49. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; Driessche, G.; Graepel, T.; Hassabis, D. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354–359. doi:10.1038/nature24270.
50. Zanzotto, F.M.; Onorati, D.; Tommasino, P.; Ranaldi, L.; Fallucchi, F. Pat-in-the-Loop: Declarative Knowledge for Controlling Neural Networks. *Future Internet* **2020**, *12*. doi:10.3390/fi12120218.