

ThumbWar – a call for RFPs

András Kornai

V0.12

0. Background The idea of *mechanization* is already present at the very beginnings of computer science (Post, 1936; Turing, 1937; Kolmogorov, 1953; McCulloch and Pitts, 1943; Kleene, 1956). Until Chomsky, 1956 introduced systems of intermediate capacity, there were only two choices: infinite or finite memory. Since arithmetic was a primary target of mechanization, finite automata received considerably less attention than the more powerful Turing-equivalent formulations. The outstanding exception was text processing, where grep-style utilities (Thompson, 1968) continue to play a dominant role to this day.

Given the arithmetic focus it is unsurprising that recursive functions were developed primarily in the $\mathbb{N} \rightarrow \mathbb{N}$ setting (see Rogers, 1967 for a classic summary). Here we argue for bringing sequence-to-sequence transducers (Sutskever, Vinyals, and Le, 2014) into the focus of theoretical work, not just because with ChatGPT they already take center stage in applied work, but also because they display a far broader range of human capabilities than arithmetic ever could. With 3-4 bit quantization (Frantar et al., 2023) they become finite state transducers, and can now be fit on a thumb drive and run on ordinary laptops.

1. Goal The classic two-level approach (Koskenniemi, 1983), based on running FSTs in parallel, is well-understood and well automated in one direction, going from two-level rules to large automata, often tens of millions of states. A variety of both proprietary and open source environments exist that support weighted generalizations as well (see this list). Such systems leverage the linguistic understanding of the problem into efficient computational blocks, both in two-level morphology and in constraint-based syntax.

The opposite direction, going from automata (often trained on noisy data) to some linguistic rules/constraints, is not nearly as well developed, in fact it is fair to say that this appears a rather hard instance of the ‘blackbox AI’ problem. Yet the linguistic evidence that there are rules (or constraints)

at play, starting with Berko, 1958, seems incontrovertible, and LLMs begin to show remarkable analogical/one shot/few shot capabilities (Brown et al., 2020), exactly the kind that are probed by the psycholinguistic experiments.

Our goal is to encourage theoretical work on the decomposition of large FSTs of the seq2seq class. We see a dangerous imbalance in the current situation, where the basic building blocks, such as LSTMs, are already opaque (Greff et al., 2015), and the composition methods, ranging from the tensorflow-style architecture diagrams Vaswani et al., 2017 to LoRA (Hu et al., 2021) were arrived at by trial and error, and do not, for the most part, reflect the functionality of the parts being composed. (For a clear description of the same problem on a much smaller scale, see Jonas and Kording, 2017.) Without an understanding of regularization, and no way to reverse composition, the danger is that we do not understand, and cannot control, emerging LLM behavior (Wei et al., 2022).

2. Approach The connection between finite automata, semigroups, and formal languages is very well known (for classic summaries, see Kuich and Salomaa, 1985; Pin, 1995), with remarkably powerful decomposition techniques such as Krohn-Rhodes (Diekert, Kufleitner, and Steinberg, 2011) available for over half a century. This kind of work, now summarized in handbooks such as Droste, Kuich, and Vogler, 2009, should play a guiding role in ‘explainable AI’ (XAI). We list here, without aiming at completeness, some key areas where research is urgently needed.

2.1 Understanding leading linguistic generalizations There is a broad consensus among workers in phonetics, phonology, and morphophonology that DISTINCTIVE FEATURES are fundamental to the organization of human speech. With end-to-end neural speech transducers now performing on a par with the best Hidden Markov Models of the previous generation, the question of how these features are encoded in the trained systems becomes quite urgent. Similar questions can be raised about all aspects of phonological organization, including AUTOSEGMENTAL STRUCTURE (Goldsmith, 1976), FEATURE GEOMETRY (Clements, 1985), HARMONIC/OT CONSTRAINTS (Smolensky, 1986), LEVEL ORDERING (Kiparsky, 1982), etc. It is perhaps worth noting that the subregular hierarchy is very much in the focus of computational phonology work (Rogers et al., 2013; Yli-Jyrä, 2015; Chandlee and Jardine, 2019; Rawski and Dolatian, 2020), which makes attempts at XAI in this domain considerably easier.

The same can be said about the vast majority of grammatical categories, from PARTS OF SPEECH (a common target in computational linguistics), GRAMMATICAL FUNCTIONS, SUBCATEGORIZATION FRAMES, etc.

Few linguists or computational linguists doubt that such generalizations are essential for the explanation of linguistic behavior, very much including the increasingly human-like linguistic behavior of LLMs (Tenney, Das, and Pavlick, 2019). We see all theoretical constructs given in SMALL CAPS as leading candidates for higher structure in the networks. Surely, if we can't find the distinctive features we are nowhere near understanding how LLMs work.

2.2 Understanding leading machine learning techniques Again, there is a broad consensus among workers in ML that certain techniques work remarkably well: the list includes reinforcement learning, now playing a critical role in AI alignment, regularization, and of course gradient optimization. But we still don't have a firm understanding what optimization, viewed in the quantized (discrete) setting, is actually doing for us as a machine operation! In other words, we are calling for careful micro-analysis to complement the large-scale behavioral work that approaches the problem from the physics side (Maloney, Roberts, and Sully, 2022).

2.3 From strings to graphs In many cases, the linguistic generalizations are better stated in terms of (hyper)graphs than in terms of strings – this is particularly clear at the higher levels of semantic structure. The situation is largely parallel to that found for strings: graph transduction has remarkable theoretical depth (Courcelle and Engelfriet, 2012) and efficient computational tools (Gontrum et al., 2017) yet today's GNNs (graph neural networks) actually work on string input.

3. Summary and conclusions The task of explaining LLMs is a large one, and for now a disparate set of junior, mid-career, and senior researchers appear to be working on what appears to be various body parts of the same elephant. We are calling on the funding agencies, not just NSF and the ERC, but also the national grant organizations and the more specialized funders, to create programs that transcend the seniority issue and national boundaries. The funders should do what they do best: promote high quality long-term theoretical work to counterbalance the enormous sums now pouring into short-term applied work.

References

Berko, Jean (1958). “The child's learning of English morphology”. In: *Word* 14, pp. 150–177.

- Brown, Tom et al. (2020). “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- Chandlee, Jane and Adam Jardine (Mar. 2019). “Autosegmental Input Strictly Local Functions”. In: *Transactions of the Association for Computational Linguistics* 7, pp. 157–168. DOI: 10.1162/tacl_a_00260. URL: <https://aclanthology.org/Q19-1010>.
- Chomsky, Noam (1956). “Three models for the description of language”. In: *IRE Transactions on Information Theory* 2, pp. 113–124.
- Clements, George N. (1985). “The geometry of phonological features”. In: *Phonology Yearbook* 2, pp. 225–252.
- Courcelle, Bruno and Joost Engelfriet (2012). *Graph structure and monadic second-order logic*. Cambridge University Press.
- Diekert, Volker, Manfred Kufleitner, and Benjamin Steinberg (2011). *The Krohn-Rhodes Theorem and Local Divisors*. arXiv: 1111.1585 [math.GR].
- Droste, Manfred, Werner Kuich, and Heiko Vogler, eds. (2009). *Handbook of Weighted Automata*. Monographs in Theoretical Computer Science. Springer.
- Frantar, Elias et al. (2023). *GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers*. arXiv: 2210.17323 [cs.LG].
- Goldsmith, John A. (1976). *Autosegmental Phonology*. PhD thesis MIT.
- Gontrum, Johannes et al. (2017). “Alto: Rapid Prototyping for Parsing and Translation”. In: *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: Association for Computational Linguistics, pp. 29–32. URL: <https://www.aclweb.org/anthology/E17-3008>.
- Greff, Klaus et al. (2015). “LSTM: A search space odyssey”. In: arXiv: 1503.04069 [cs.NE].
- Hu, Edward J. et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv: 2106.09685 [cs.CL].
- Jonas, Eric and Konrad Paul Kording (2017). “Could a Neuroscientist Understand a Microprocessor?” In: *PLOS Computational Biology* 13.1, e1005268. DOI: 10.1371/journal.pcbi.1005268.
- Kiparsky, Paul (1982). “From cyclic phonology to lexical phonology”. In: *The structure of phonological representations, I*. Ed. by H. van der Hulst and N. Smith. Dordrecht: Foris, pp. 131–175.

- Kleene, Stephen C. (1956). “Representation of events in nerve nets and finite automata”. In: *Automata Studies*. Ed. by C. Shannon and J. McCarthy. Princeton University Press, pp. 3–41.
- Kolmogorov, Andrei N. (1953). “O ponyatii algoritma”. In: *Uspehi matematicheskikh nauk* 8.4, pp. 175–176.
- Koskenniemi, Kimmo (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. University of Helsinki: PhD thesis.
- Kuich, W. and A. Salomaa (1985). *Semirings, Automata and Languages*. Springer-Verlag New York, Inc. Secaucus, NJ, USA. ISBN: 0387137165.
- Maloney, Alexander, Daniel A. Roberts, and James Sully (2022). *A Solvable Model of Neural Scaling Laws*. arXiv: 2210.16859 [cs.LG].
- McCulloch, W.S. and W. Pitts (1943). “A logical calculus of the ideas immanent in nervous activity”. In: *Bulletin of mathematical biophysics* 5, pp. 115–133.
- Pin, Jean-Eric (1995). “Finite semigroups and recognizable languages: an introduction”. In: *Semigroups, formal languages and groups*. Ed. by John Fountain. Kluwer Academic, pp. 1–32.
- Post, Emil (1936). “Finite Combinatory Processes-Formulation 1”. In: *The Journal of Symbolic Logic* 1.3.
- Rawski, Jonathan and Hossep Dolatian (2020). “Multi-Input Strict Local Functions for Tonal Phonology”. In: *Proceedings of the Society for Computation in Linguistics*. Vol. 3. 25.
- Rogers, Hartley (1967). *The theory of recursive functions and effective computability*. McGraw-Hill.
- Rogers, James et al. (2013). “Cognitive and sub-regular complexity”. In: *Formal Grammar*. Vol. 8036. Lecture Notes in Computer Science. Springer, pp. 90–108.
- Smolensky, Paul (1986). “Information Processing in Dynamical Systems: Foundations of Harmony Theory”. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Ed. by James Lloyd McClelland and David Everett Rumelhart. Cambridge, MA, USA: MIT Press, pp. 194–281. ISBN: 026268053X.
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). “Sequence to Sequence Learning with Neural Networks”. In: *Proc. NIPS*. Montreal, CA, pp. 3104–3112. URL: <http://arxiv.org/abs/1409.3215>.
- Tenney, Ian, Dipanjan Das, and Ellie Pavlick (July 2019). “BERT Rediscovered the Classical NLP Pipeline”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: As-

- sociation for Computational Linguistics, pp. 4593–4601. DOI: 10.18653/v1/P19-1452. URL: <https://www.aclweb.org/anthology/P19-1452>.
- Thompson, K. (1968). “Regular Expression Search Algorithm”. In: *Communications of the ACM* 11.6, pp. 419–422.
- Turing, Alan (1937). “On computable numbers, with an application to the Entscheidungsproblem”. In: *Proceedings of the London Mathematical Society* 42, pp. 230–265, 544–546.
- Vaswani, Ashish et al. (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 5998–6008. arXiv: 1706.03762 [cs.CL]. URL: <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>.
- Wei, Jason et al. (2022). *Emergent Abilities of Large Language Models*. DOI: 10.48550/ARXIV.2206.07682. URL: <https://arxiv.org/abs/2206.07682>.
- Yli-Jyrä, Anssi (2015). “Three Equivalent Codes for Autosegmental Representations”. In: *Proceedings of the 12th International Conference on Finite-State Methods and Natural Language Processing 2015 FSMNLP 2015 Düsseldorf*.