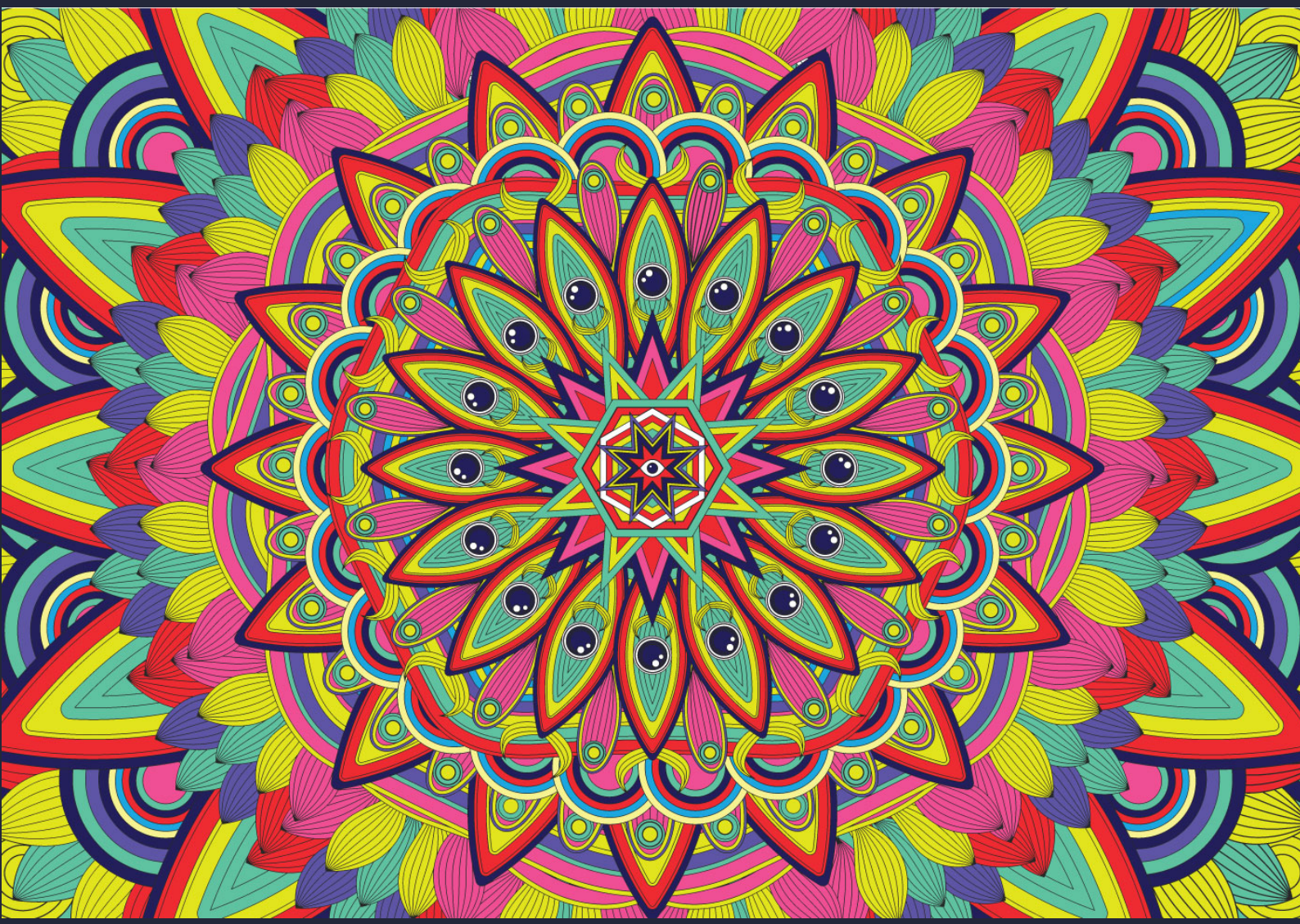
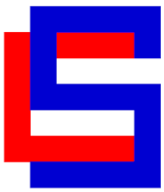
LSD 2023

Proceedings of the 2023 CLASP Conference on Learning with Small Data

Editors: Ellen Breitholtz, Shalom Lappin, Sharid Loáiciga, Nikolai Ilinykh, and Simon Dobnik



Gothenburg and online
11–12 September 2023



CLASP centre for linguistic theory and studies in probability

CLASP Papers in Computational Linguistics, Volume 5
University of Gothenburg

CLASP Conference Proceedings, Volume 2

Front-cover photo:
psychedelic pattern, https://www.freevector.com/psychedelic-pattern-28724.

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISSN 2002-9764
ISBN 979-8-89176-000-4

Message from the organisers

We are happy to welcome you to the CLASP Conference on Learning with Small Data (LSD 2023)! This volume consists of the papers presented at the LSD conference held at the Department of Philosophy, Linguistics and Theory of Science (FLoV), University of Gothenburg on September 11–12, 2023.

The purpose of our conference is to bring together researchers from several areas of NLP, addressing datasets, methods and limits of *effective* (machine) learning with **small data** containing natural language and associated multi-modal information. The conference covers areas such as machine learning, natural language processing, language technology, computational linguistics, theoretical linguistics, psycholinguistics, as well as artificial intelligence, cognitive science, ethics, and policy.

Current deep learning systems require large amounts of data in order to yield optimal results. Despite ever-increasing model and data size, these systems have achieved remarkable success across a wide range of tasks in NLP, and AI in general. However, these systems possess a number of limitations. Firstly, the models require a significant amount of time for pre-training, and modifying them proves to be challenging. As a result, much NLP research is shaped by what can be achieved with large transformers. This has marginalised important computational learning questions for which they are not well suited. Second, due to the substantial resources necessary for their development, they have become the preserve of technological companies. Researchers are now positioned as consumers of these systems, restricted to fine-tuning them for experimental work on downstream tasks. Thirdly, the complexity, size, and mode of computation of transformers have obscured the process through which they derive generalisations from data. This opacity has created a challenge in comprehending precisely the reasons behind their success or failure in different scenarios. Finally, comparison with human learning and representation has become increasingly difficult, given the large disparity in accessible data and learning time between transformers and humans. Therefore, the cognitive interest of deep learning has receded.

Papers were invited on topics from these and closely related areas, including (but not limited to): small-scale neural language modelling, both text and multi-modal; training corpus and test task development; visual, dialogue and multi-modal inference systems; neurolinguistic and psycho-linguistic experimental approaches to human language processing; semantics and pragmatics in neural models; dialogue modelling and linguistic interaction; formal and theoretical approaches to language production and comprehension; language acquisition in the context of computational linguistics; statistical, machine learning, reinforcement learning, and information theoretic approaches that embrace small data; methodologies and practices for annotating datasets; visual, dialogue and multi-modal generation; text generation in both the dialogue and document settings; semantics-pragmatics interface; social and ethical implications of the development and application of large or small neural language models, as well as relevant policy implications and debates.

This conference aims to initiate a discussion about these related topics and to examine various approaches and how they can mutually inform each other. The event includes 4 keynote talks, 10 peer-reviewed long papers, 5 peer-reviewed short papers, 5 peer-reviewed student papers, and 9 non-archival presentations.

We would like to thank all our contributors and programme committee members, with special thanks to CLASP for organising the hybrid conference and the Swedish Research Council for funding CLASP.

Ellen Breitholtz, Shalom Lappin, Sharid Loáiciga, Nikolai Ilinykh, and Simon Dobnik

Gothenburg

September 2023

Organising Committee

Program Chairs

Ellen Breitholtz, University of Gothenburg, Sweden
Shalom Lappin, University of Gothenburg, Sweden
Sharid Loáiciga, University of Gothenburg, Sweden

Proceedings Chairs

Simon Dobnik, University of Gothenburg, Sweden
Nikolai Ilinykh, University of Gothenburg, Sweden

Local Arrangements Organisers

Asad Sayeed, University of Gothenburg, Sweden
Susanna Myyry, University of Gothenburg, Sweden

Programme Committee

Maxime Amblard	University of Lorraine
Alexander Berman	University of Gothenburg
Raffaella Bernardi	University of Trento
Jean-Philippe Bernardy	Chalmers Technical University
Ellen Breitholtz	University of Gothenburg
Stergios Chatzikyriakidis	University of Crete
Rui Chaves	University at Buffalo
Alexander Clark	University of Gothenburg
Ariel Cohen	Ben-Gurion University of the Negev
Robin Cooper	University of Gothenburg
Philippe de Groote	Inria Nancy - Grand Est
Simon Dobnik	University of Gothenburg
Markus Egg	Humboldt University
Adam Ek	University of Gothenburg
Katrin Erk	University of Texas at Austin
Arash Eshghi	Heriot-Watt University
Chris Fox	University of Gothenburg
Jonathan Ginzburg	Université Paris-Diderot
Eleni Gregoromichelaki	University of Gothenburg
Xudong Hong	Saarland University
Julian Hough	Queen Mary University of London
Christine Howes	University of Gothenburg
Nikolai Ilinykh	University of Gothenburg
Elisabetta Jezek	University of Pavia
Richard Johansson	Chalmers Technical University
Aram Karimi	University of Gothenburg
Ruth Kempson	King's College London
Nikhil Krishnaswamy	Colorado State University
Shalom Lappin	University of Gothenburg
Staffan Larsson	University of Gothenburg
Sharid Loáiciga	University of Gothenburg
Vladislav Maraev	University of Gothenburg
Yuval Marton	University of Washington
Elin McCready	Aoyama Gakuin University
Louise McNally	Universitat Pompeu Fabra
Gregory Mills	Kingston University
Marie-Francine Moens	KU Leuven
Joakim Nivre	Uppsala University
Bill Noble	University of Gothenburg
Manfred Pinkal	Saarland University
Massimo Poesio	Queen Mary University of London
Sameer Pradhan	cemantix.org & UPenn (Linguistic Data Consortium)
Violaine Prince	University of Montpellier
Matthew Purver	Queen Mary University of London
James Pustejovsky	Brandeis University
Chatrine Qwaider	Chalmers Technical University
Christian Retoré	University of Montpellier
Mehrnoosh Sadrzadeh	University College London

Asad Sayeed	University of Gothenburg
David Schlangen	University of Potsdam
William Schuler	The Ohio State University
Sabine Schulte im Walde	University of Stuttgart
Vidya Somashekarappa	University of Gothenburg
Tim Van de Cruys	KU Leuven
Carl Vogel	Trinity College Dublin
Alessandra Zarcone	Augsburg University of Applied Sciences
Sina Zarrieß	University of Bielefeld

Invited Speakers

Aurélie Herbelot, University of Trento
Tal Linzen, New York University & Google
Danielle Matthews, University of Sheffield
Shalom Lappin, University of Gothenburg

Invited talk: Aurélie Herbelot

Decentralised semantics

Large Language Models (LLMs) are currently the dominant paradigm in the field of Natural Language Processing. But their enormous architecture, coupled with an insatiable hunger for training data, makes them ill-suited for many purposes, ranging from fundamental linguistic research to small business applications. The main argument of this talk is that the monolithic architecture of LLMs, and by extension their reliance on big data, is a direct consequence of a lack of semantic theory in the underlying model. As an alternative, I will explore a modular architecture based on concepts from model theory, which lends itself to decentralised training over small data. Starting from research in linguistics and cognitive science, I will summarise evidence against the view that language competence should "live" in a single high-dimensional space. I will then review various computational models of meaning at the junction between formal and distributional approaches, and show how they can be combined into a modular system. Finally, I will present a possible implementation where learning takes place over individual situation types, at low dimensionality. This decentralised approach has natural benefits in terms of accessibility and energy efficiency.

Invited talk: Danielle Matthews

How children learn to use language through interaction

This talk will chart out pragmatic development with a focus on the real-world experiences that allow infants to start using language for social communication and permit children to use it at ever more complex levels. Following a working definition of pragmatics in the context of human ontogeny, we will trace the early steps of development, from a dyadic phase, through to intentional triadic communication and early word use before briefly sketching out later developments that support adult-like communication at the sentential and multi-sentential levels and in literal and non-literal ways. Evidence will be provided regarding the experiential basis of learning from the study of individual differences, from randomised controlled trials and from deaf infants growing up in families with little prior experience of deafness (and who are thus at risk of reduced access to interaction). This will provide a summary of elements from a forthcoming book: *Pragmatic Development: How children learn to use language for social communication*.

Invited talk: Tal Linzen

How much data do neural networks need for syntactic generalisation?

I will discuss work that examines the syntactic generalisation capabilities of contemporary neural network models such as transformers. When trained from scratch to perform tasks such as transforming a declarative sentence to a question, models generalise in ways that are very different from humans. Following self-supervised pre-training (word prediction), however, transformers generalise in line with syntactic structure. Robust syntactic generalisation emerges only after exposure to a very large amount of data, but even more moderate amounts of pre-training data begin to steer the models away from their linear inductive biases. Perhaps surprisingly, pre-training on simpler child-directed speech is more data-efficient than on other genres; at the same time, this bias is insufficient for a transformer to learn to form questions correctly just from the data available in child-directed speech.

Invited talk: Shalom Lappin

Assessing the Strengths and Weaknesses of Large Language Models

The transformers that drive chatbots and other AI systems constitute large language models (LLMs). These are currently the focus of a lively discussion in both the scientific literature and the popular media. This discussion ranges from hyperbolic claims that attribute general intelligence and sentience to LLMs, to the skeptical view that these devices are no more than "stochastic parrots". In this talk I will present an overview of some of the weak arguments that have been presented against LLMs, and I will consider several more compelling criticisms of these devices. The former significantly underestimate the capacity of transformers to achieve subtle inductive inferences required for high levels of performance on complex, cognitively significant tasks. In some instances, these arguments misconstrue the nature of deep learning. The latter criticisms identify significant limitations in the way in which transformers learn and represent patterns in data. They also point out important differences between the procedures through which deep neural networks and humans acquire knowledge of natural language. It is necessary to look carefully at both sets of arguments in order to achieve a balanced assessment of the potential and the limitations of LLMs.

Table of Contents