Proposal to Establish a Priority Program

Robust Assessment & Safe Applicability of Language Modeling: Foundations for a New Field of Language Science & Technology (LaSTing)

Programme committee

Prof. Dr. **Michael Franke**, University of Tübingen, Tübingen, Professor (W3), General Linguistics / Pragmatics / Cognitive Science (**coordinator**)

Prof. Dr. **Vera Demberg**, Saarland University / MPI for Informatics, Saarbrücken, Professor (W3), Computational Linguisics / Psycholinguistics

Prof. Dr. **Gerhard Jäger**, University of Tübingen, Tübingen, Professor (W3), General Linguistics / Computational Historical Linguistics

Prof. Dr. **Barbara Plank**, Ludwig Maximilian University, Munich, Professor (W3), Artificial Intelligence / Computational Linguistics

Prof. Dr. **David Schlangen**, University of Potsdam, Potsdam, Professor (W3), Computational Linguistics / Artificial Intelligence / Dialogue Processing

1 Summary

The recent rapid increase in performance of neural language models has surprised even leading experts at the frontier of development. Though in large part not fully understood, modern language technology is already permeating industry, education, personal life, and scientific research in multiple ways. These developments bring particular challenges and chances for theoretical linguistics and the wider field of the (cognitive / computational) language sciences. For one, the language sciences are challenged to solve deep foundational issues regarding the role of language models in theoretical and empirical studies on human language. For another, there are immense opportunities for using language technology as a tool in linguistic inquiry, if its implications are understood well enough. Moreover, the rapid development of language technology has outpaced our ability to develop well-grounded methods for understanding and safely applying language models, whether in science or elsewhere. To help solve these methodological problems, grown-and-tested theoretical concepts and established standards for empirical research from the language sciences readily suggest themselves, but require further, dedicated development.

These issues are urgent and felt strongly across traditional disciplinary boundaries. Indeed, researchers from diverse fields, spanning theoretical and experimental linguistics, computational linguistics, psycholinguistics, natural language processing and, more generally, the cognitive sciences are troubled by very similar problems in very different contexts of research. We are seeing a new emerging field of interdisciplinary and methodologically diverse work at the interface between the (cognitive) language sciences (broadly construed) and language technology (focused on neural language models, but not exclusively so). At the heart of this field are foundational issues, touching on methodology and core concepts, which are well-nigh impossible to be solved within a few disparate research projects. What is required is an umbrella structure to unite scattered research efforts, to enable cross-fertilizing dialogue, and to provide the grounds on which research communities can agree on shared values, concepts and methods as the foundation for new field of language research pivoting around language modeling. We therefore propose a Priority Program, LaSTing, which provides such a platform for community-wide, interdisciplinary effort in order to provide common solutions for common problems.

2 State of the art and preliminary work

We live in very exciting times for anyone interested in language. The methodological basis for theorizing about language and for explaining data pertaining to human linguistic processing and communication has

been broadening steadily and fruitfully. For example, the last two decades have brought an **experimental turn in linguistics** especially in subfields, such as syntax, semantics and pragmatics, which had so far mostly relied on formal and analytic methods (Meibauer and Steinbach, 2011; Noveck, 2018; Sprouse, 2023). Traditionally experimental fields in the language sciences, like psycholinguistics, have seen a rise in the use of **computational models** for more insightful analyses of subtle quantitative aspects of empirical data (Crocker, 2010; Dotlačil and Brasoveanu, 2020; Erk, 2022). Yet, the most striking methodological advances clearly happened in language technology with the advent of extremely powerful **Language Models** (LMs), which not only seem to be able to predict language behaviour very accurately but also —thanks to this very property— show astonishing practical value and are now beginning to permeate many routines in business, education, science, and personal life; and promise to be highly impactful in the future as well.

The trained instances of modern LMs which attracted wide-spread public attention have been developed at enormous costs mainly by major players in industry. The **main motivation behind LM development was applicability, not transparency of the models**, let alone scientific curiosity into the workings of human language. Indeed, the formula for success of recent language technology is strikingly simple and in part fortunate historical happenstance. While the mathematical foundations for models of neural networks for language processing are several decades old (Elman, 1990; Hochreiter and Schmidhuber, 1997; Bengio et al., 2003), and the routines for pre-training current models build on long established ideas of backpropagation, three main factors have arguably led to a fundamental breakthrough in the applicability of statistical learning models for language during the last years. First, **the invention of the transformer architecture** (Vaswani et al., 2017) allowed neural models of sequence processing to be trained in parallel, thus allowing for scalability, i.e., the ability to efficiently train much larger models. Second, driven largely by the gaming industry, we have hardware in the form **high-performing processing units (GPUs, TPUs) for massive parallel computation** which but turned out to be excellent for training the artificial neural network architectures underlying current LMs. Third, the internet made available **vast amounts of (unlabelled, unstructured) data for training** these models.

The above description of the development and employment of modern LMs seems to suggest that scientific fields concerned with human language on a theoretical or empirical basis, like linguistics, the (cognitive) language sciences, and computational linguistics (all of which will be addressed with the umbrella term *language sciences* for ease of reference in the following), have played no role at all and may have —in part— even become obsolete. This initial impression is natural, but emphatically incorrect. Eclipsed by the focus of public attention, scientists from many different fields have started to investigate the behavioral and theoretical underpinnings of modern language technology (Wahle et al., 2023), including early work from the cognitive language sciences (e.g., Linzen et al., 2016; Hu, Gauthier, et al., 2020). Recent years have seen **an emerging vibrant research strand of "LM-ology"**, i.e., the scientific study of language models. LM-ology is variably concerned with interpreting the behavior of LMs at an input-output level (e.g., BIG-bench authors, 2023), shedding light on the inner representations and mechanisms ("un-black-boxing") of trained models (e.g., K. Clark et al., 2019), the impact of training and fine-tuning routines (Ouyang et al., 2022), and characterizing potential limits of the capability of LMs from an experimental, conceptual or mathematical point of view (e.g., Hahn, 2020; Binz and Schulz, 2023), or by comparing models to humans at different levels of description (Shiffrin and Mitchell, 2023). However, as with any emerging new field of interdisciplinary research, the proper goals, concepts and methods for LM-ology are yet to be made precise. This is where **input from the languages sciences is crucial**. A first major motive behind the present proposal is the necessity to **sharpen the scientific profile of the study of language models with established concepts and methods from the language sciences**.

This kind of methodological, interdisciplinary effort is necessary, because **current LM-ology lacks a proper methodological foundation**. Given the speed of development of models, the development of methods that give lasting, generalizable results has not been able to keep up. This shows in many

diverse areas, key examples of which are discussed in detail below. To put it in provocative terms, the relevance of studies investigating the performance of the currently most prominent model on a quickly assembled benchmark loosely resembling patterns of reasoning potentially relatable to some phenomenon of interest has a half-life of just a couple of months and teaches us little about the inherent capabilities of a *class* of models, let alone human language or cognition. Instead, what is required for the next couple of years is communal effort to converge on proper standards for **robust assessment of language models**. Methodology is robust, in the sense intended here, if its results are **generalizable** (carrying over with sufficient certainty to other models and data sets), **transferable** (insightful beyond the purposes of LM-ology), and **reproducible** (with the same or different models and data sets). Robust methodology also aspires to be as **future-proof** as possible, i.e., likely relevant for the next generation of models or antagonistic examples. Concepts and methods from the language sciences promise to provide exactly this if brought to bear appropriately.

Understanding the nature, workings and capabilities of LMs is relevant also for delineating the proper way of using LMs in downstream applications. In the ideal case, we would be certain that for any possible input, the resulting output will meet the criteria of adequacy relevant for the current use-case. But such certainty is presently out of reach for most purposes given the highly complex and largely intransparent nature of modern LMs; a fact which highlights the relevance of methods for robust assessment. Nevertheless, it is clear that we must strive towards what we here address as **safe applicability of language models**, which we take to subsume critical aspects such as being **conceptually sound** (e.g., anchored in "first principles" or established empirical knowledge), ideally **validated** (e.g., by mathematical proof or other rigorous derivation) or at least **stress-tested** across a near-exhaustive traversal of possible conditions of use, **ethical** (e.g., bias- and harm-free, or privacy-respecting), and also **economical** (i.e., minimizing data requirements and energy consumption).

Concern for safe applicability applies not only to practical applications (e.g., user-facing products), but also to **applications of language technology in science**. While industry-driven applications may naturally be very concerned about ethical and economical aspects, applications of language technology for knowledge gain put particular emphasis on the soundness and validity aspects of safe applicability. Indeed, LMs have been suggested to be helpful at almost all stages of common research cycles. To begin with, LMs can assist during simple, auxiliary tasks, such as data processing, programming assistance, or automatic visualization. More controversially, **LMs could possibly be used in scientific research to support or even replace human input**. Again, this can be reasonably benign, as for example in the case of assisting or replacing humans in simple data annotation tasks (Ziems et al., 2023). Slightly more critical is the potential for automatic creation of experimental stimuli (Gandhi et al., 2023). LMs may also be used inside of larger models, be it for applied or explanatory purposes, such as in neuro-symbolic models (M. C. Frank, 2023). In the most extreme and controversially discussed cases, however, LMs would be used as complete stand-ins for human intuition or human responses, as is the case when LMs *replace* human experimental participants (Aher et al., 2023; Dillion et al., 2023; Harding et al., 2023; Bavaresco, Bernardi, et al., 2024). It is obvious that the more important the correctness of the LM's performance is for the validity of the research it supports, the heavier the burden is on solid understanding of the capabilities of the technology. A second major motivation for this SPP is therefore to **lay the methodological foundation for using novel language technology in theoretically-informed applications**, be that practical applications that draw on insights from the language sciences, or direct applications *within* the language sciences for some explanatory purpose.

Computational, theoretical and empirical research into human language is particularly well positioned to contribute to a better understanding of modern language technology and its safe applicability — and indeed, linguists and language scientists have been among the first to contribute to assessment of the linguistic capabilities in language models and the human-likeness of their behavior (e.g., Linzen et al., 2016; Hu, Gauthier, et al., 2020; Wilcox et al., 2021). But **modern language technology also raises**

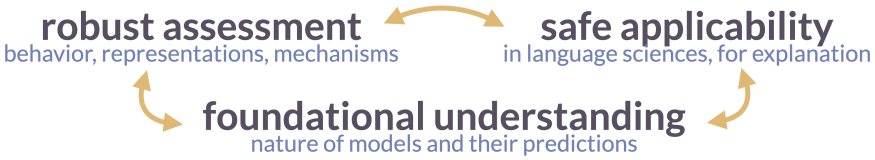

Figure 1: Three pillars of research on language models in a new field of language science & technology.

deep foundational issues specific to the language sciences which are now surfacing pressingly in relation to theoretical and empirical work (Fox and Katzir, 2024; Mahowald et al., 2024). These issues must be addressed widely and explicitly to make progress on robust assessment and safe applicability of the technology. Central issues revolve around the nature of LMs as a class of abstract models (What are LMs models of?) and their proper role in the scientific research into human language (How can LMs be used as explanatory tools for understanding human language?).

In sum, the present research landscape at the interface between the language sciences and language technology consist of three tightly connected areas (see Figure 1): (i) advancing understanding of the relevant technology through theoretically informed and methodologically sound **robust assessment**, (ii) honing in on standards for **safe applicability** which are conceptually anchored in established knowledge from the language sciences, and (iii) developing a deeper **foundational understanding** concerning the nature of language modeling and its place in the language sciences. Progress in each of these areas is largely dependent on progress in the others. It is impossible to solve everything at once with a single-team, dedicated research project. Rather, distributed, interdisciplinary research efforts, which are anchored in concrete research traditions or address concrete problems of practical relevance, will need to be synergized to **provide the basis for a new field of language science and technology for research *on* and *with* language models**. This, in a nutshell, is the main goal behind the proposal of LaSTing.

To demonstrate that the structural support for this rising research field via a Priority Programme is timely, relevant and possible, the following paragraphs will survey a number of concrete **core issues** (see Figure 2) that loom large in the current research landscape along the interface between the language sciences and language technology and that should be addressed by LaSTing. We emphasize that these core issues are often highly connected, thus requiring cooperative interdisciplinary exchange. We also document how the core issues arise in concrete instances of current research in rather diverse fields of (applied) research and linguistic sub-disciplines, which have traditionally not interacted closely. Our examples cover (computational) psycholinguistics, the cognitive neuroscience of language, language acquisition, syntax, semantics, pragmatics, historical linguistics, discourse and dialogue, as well as the philosophy of language and linguistics. Nevertheless, it needs to be emphasized that the given examples are by no means exhaustive, but only intended to demonstrate that **there is a crisp set of shared problems that re-occur, over and over again, across a wide range of seemingly unrelated research areas**, which clearly shows that a new interdisciplinary field of research on and around language modeling is emerging and requires shaping as provided by LaSTing.

❖ **Behavioral Assessment.** The most obvious way of investigating the input-output behavior of generative models is to use what we here call *behavioral assessment* (in allusion to behaviorism from psychology, which is also only concerned by observable behavior and shuns speculation about internal information processing). Behavioral assessment usually consists of benchmark testing when performance issues are relevant (e.g., BIG-bench authors, 2023) or of treating models like human subjects in a behavioral psychological experiment when we care for human-likeness of their input-output behavior (e.g., Binz and Schulz, 2023). Yet, in either domain, there are a lot of seemingly arbitrary researcher degrees of freedom, which as such should already be alarming from the point of view of good practices of scientific inquiry (Chambers, 2017; Wieling et al., 2018). It is well-known that predictions of LMs depend

Robust Assessment

❖ Behavioral Assessment (❖BehAss)

What are adequate, robust methods of experimentally assessing the (abstracted, linguistic) capability of an LM based on its input-output behavior? What is a valid comparison of machine predictions to human behavior?

❖ Representations & Mechanisms (❖RepMec)

Which information is reliably retrievable from LMs' latent representations (embeddings) for linguistic/explanatory purposes or for understanding the inner workings of LMs? How can we distill the abstract computational processes that generate an LM's behavior?

❖ Training & Optimization (❖TrainOpt)

How can we understand LMs in terms of their optimization, e.g., in terms of properties of the training data, their internal inductive biases, the training objective etc.? How does that compare to human language learning?

Safe Applicability

❖ Task Decomposition Models (❖TaskDec)

What are best practices for using LMs as part of a larger (theoretically informed) composition of the task to be solved (e.g., in agent models, applications like RAG, or explanatory, neuro-symbolic cognitive models)?

❖ Resource Efficiency (❖ResEff)

How can we solve problems of data-hunger and computational costs (training and inference), e.g., by taking human-like inductive biases into account, or using more informative curated data? How can we use synthetic data and machine judgements to solve theoretical issues?

❖ Alternative Models (❖AltMod)

How can language science benefit from alternative models beyond text-to-text LMs, e.g., by embracing multimodality, interaction, dialogue, or more cognitively plausible model architectures?

Foundations

❖ Ontological Status (❖OntStat)

Are LMs models or theories of language? What exactly does an LM predict (occurrences frequencies, behavior of idealized speaker, individualized speakers, …)?

❖ Explanatory Potential (❖ExplPot)

How can novel language technology be used as or in support of explanations, e.g., of linguistic phenomena, empirical or experimental data in the language sciences?

Figure 2: Core issues at the interface between the language sciences and language technology.

in non-systematic ways on properties of the input prompts (e.g., Webson and Pavlick, 2022; Leidinger et al., 2023). Looking at the common task of multiple choice selection (of which the standard language modeling task —missing- or next-word prediction— is arguably a special case), **there is presently no consensus on how to precisely determine the predictions of a given model**. There is even leeway in the way the probability of the next word in a sequence is to be computed precisely if complex models use particular tokenization schemes (Oh and W. Schuler, 2024; Pimentel and Meister, 2024). Naive ways of calculating predictions for multiple-choice tasks are known to be biased in non-human ways (Zhao et al., 2021; Holtzman et al., 2021). While disparate (ad hoc) solutions for de-biasing or prompt-engineering abound, different methods of assessment can given different results for different models (Hu and Levy, 2023; Tsvilodub, H. Wang, et al., 2024). The kind of assessment chosen may relate in intricate ways to aspects similar to performance-related factors or task-demands in human studies (Hu and M. C. Frank, 2024). At the heart of these problems is, arguably, the lack of foundational consensus on what LMs are capable of predicting due to their training and formal design (see ❖OntStat). Moreover, there are good reasons to judge system performance or abilities on more than just accuracy of output (Shiffrin and Mitchell, 2023; Mondorf and Plank, 2024). In short, all of these considerations call for **community-wide reflection on best practices of conducting behavioral assessment** in a way that is fair to both humans and machines, in parallel to established considerations in fields like Comparative Psychology (Hagendorff, 2023; Lampinen, 2023).

A concrete example for the importance of reflection on how to assess predictions of neural language models come from recent research in **computational psycholinguistics**. A prominent theory of processing difficulty is *surprisal theory*, which maintains that human effort in incremental processing of language can be predicted well by a measure of next-word surprisal (Levy, 2008). Concrete predictors for empirical validation of this theory usually come from language models and have been shown to provide good fit to measures from self-paced reading (Smith and Levy, 2013), EEG (S. L. Frank et al., 2015), or eye-tracking (Demberg and Keller, 2008). However, as recent research has shown, **predictions from larger and generally better performing language models provide worse predictions for human data** (Oh and W. Schuler, 2023). This raises fundamental questions of how to obtain predictions from language

models and how to define adequate link functions to map these predictions to (explanatory variables for) experimental data from humans. In the specific case of surprisal theory, it seems that we require a more sophisticated, parameterized link function to accommodate for the usually too confident predictions of high-performing models (Liu et al., 2024). This issue, while arising here in the concrete context of established research in computational psycholinguistics, is structurally similar to the problem of overconfidence in the context of other multi-choice prediction tasks (e.g., Kumar, 2022; Si et al., 2022). In sum, the example shows how tightly related issues of behavioral assessment affect otherwise unrelated research areas, thus requiring structures that foster common solutions for common problems.

❖ **Representations & Mechanisms.** Vector embeddings have emerged as a very powerful representational format for word and sentence meanings, from static non-contextualized (e.g., Mikolov et al., 2013) to contextualized representations (e.g., Peters et al., 2018). Yet, deep lingering questions remain, such as in how far embeddings can capture subtle aspects of *linguistic productivity* and *compositional meaning* (e.g., Erk, 2012; Hupkes et al., 2020; Y. Yao and Koller, 2022). In the context of LMs, a prominent open research question is which kinds of (conceptual or linguistic) knowledge are represented in the internal representations during a forward pass. Many works have therefore investigated specifically the mechanisms underlying language models with an eye to discovering processing steps and representations that link to known linguistic categories or processes, in particular related to parsing and syntax (K. Clark et al., 2019; Hewitt and Manning, 2019; Tenney et al., 2019; Müller-Eberstein et al., 2022; Waldis et al., 2024).

To fully understand internal representations in a neural network, we arguably have to understand the information provided by the representations for the larger computational process that is performed by the network (Rahwan et al., 2019; Harding, to appear). Issues of **mechanistic interpretability** are therefore a very prominent topic of ongoing research in machine learning and NLP. Yet, owing to relative novelty of these approaches, current methods may yet require conceptual ripening, additional validation and stress-testing. Some currently popular methods, e.g., the *logit lense* (nostalgebraist, 2020), may yield interesting and seemingly interpretable results, but currently lack a theoretical or mathematical justification. Other methods promise clearer paths towards conceptual grounding, as they revolve around counterfactual notions of alternativeness, a topic well-researched in theoretical and experimental linguistics (Gotzner and Romoli, 2022) and relevant for causal explanation (e.g., Geiger et al., 2021). For example, mechanistic attribution methods like *contrastive explanations* (Yin and Neubig, 2022), *amnesic probing* (Elazar et al., 2021), *causal mediation analysis* (Vig et al., 2020), or *activation or path patching* (Meng et al., 2024), all entail comparison with a counterfactual input or activation pattern. These alternatives are currently chosen primarily based on technical requirements (e.g., for "empty" or "neutral" activation patterns) or selected based on pre-theoretic intuition. Yet, with an eye towards conceptual soundness of robust methodology, these should ideally be anchored in theories of alternative expressions (Katzir, 2007; Rohde and Kurumada, 2018) and meanings (Beaver et al., 2017; Buccola et al., 2021), drawing on ideas from syntax, semantics and pragmatics. Similarly, applications of methods like *circuit analysis* (e.g., K. Wang et al., 2022; Merullo et al., 2024) often require decomposition of the task which is presumed to be performed, so that for many non-trivial cases of language processing extant linguistic analyses and empirical results should serve as guidance (see ❖TaskDec). In sum, we expect that **linguistic concepts play a crucial role for interpretability of language models**, just as concepts that are intelligible for humans matter for interpretation of deep neural networks in general (Kim et al., 2018; Marks et al., 2024).

An exciting example of vibrant current work where modern language technology and the language sciences have largely overlapping interests concerning internal representations and interpretability comes from **computational cognitive neuroscience of language**. There is a growing interest in comparing latent representations from LMs with human brain activity during language processing (Gauthier and Levy, 2019; Caucheteux and King, 2022; Schrimpf et al., 2021). Yet controversies exist, be it general methodological (Antonello and Huth, 2023) or specific linguistic issues, such as whether correlation is driven by

syntactic or more semantic features (Kauf et al., 2024; Fresen et al., 2024). Settling such issues requires identifying *where* LMs perform *which* operation, which in turn requires better grasp on what embeddings represent and how LMs compute with the available information. Concretely, the prominent method of *neural decoding* uses latent representations from LMs to predict human brain activation patterns for identical stimuli (Gerven et al., 2019). This mapping relies on machine learning tools, but frequently resorts to simple linear regression. To ensure safe methodology, the choice of adequate mapping should be informed by robust results from interpretability studies, which is what, in turn, requires cross-disciplinary exchange (cf., Beinborn et al., 2023). Looking ahead, the currently prevalent transformer-based architecture may not be the ideal comparison to human language processing, since it assumes that all prior linguistic input is available *verbatim* at all times (but needs to be selectively attended to). Human processing, on the other hand, relies on compressed representations of prior input in memory. Whence that recently developed alternative model architectures, like selective state space models (Gu and Dao, 2023) or extended long-short term memory models (Beck et al., 2024), might provide a clearer analogue to human representation during language processing (❖AltMod).

❖ **Training & Optimization.** To fully understand a trained LM for safe applications, we must also consider a functional, teleological perspective (Rahwan et al., 2019; McCoy et al., 2024), i.e., the complex interaction between: (i) the system's architecture, (ii) the training task (objective function), (iii) the algorithmic optimization strategy, and (iv) composition of the training data. In the ideal case, mathematical results may give us certainty that particular properties must or cannot possibly ensue (e.g., Hahn, 2020). But matters are complex since modern LMs are often optimized iteratively, using different training objectives, data sets and optimization techniques (e.g., Ouyang et al., 2022), and may be closed source. To ensure safe applicability, it is therefore important to strive for generalizable and transferable, if not future-proof insights into the effects of training and optimization.

An important contact area between the languages sciences and language technology where these issues are prominently discussed in recent work is **language acquisition**, with possible implications for **syntactic theory**. One common approach is to study learnability under controlled variation of the training data, the model size etc., similar to Artificial Language Learning tasks (Culbertson and K. Schuler, 2019). To study the impact of training regimes and model architecture, the popular BabyLM Challenge, first held in 2023, set the task to train neural language models with training data roughly commensurable with the amount of language input which children are exposed to during first-language acquisition (Warstadt, Mueller, et al., 2023). Many contributions explored potentially more efficient training regimes (see ❖ResEff) inspired by human language learning (e.g., Bunzeck and Zarrieß, 2023). Training LMs on realistically sized, but systematically manipulated input also helps to shed light on the inductive learning biases implicit in LMs. In this way, LMs might serve, effectively, as miniature models for studying language learning under controlled laboratory conditions (Warstadt and Bowman, 2024). Highly interesting questions at the syntax-technology interface arise, such as whether LMs can learn rare syntactic constructions from generalization or only from memorization (Misra and Mahowald, 2024) or whether it is harder for an LM to learn an artificially constructed language which is considered harder to learn in common theories of syntax (Kallini et al., 2024). Importantly, concerns of (the limits of) learnability quickly also lead to considering alternative model architectures (see ❖AltMod).

❖ **Task Decomposition Models.** When expensive fine-tuning is not an option, an attractive strategy to coerce LMs to perform a difficult task, is *prompt engineering* (e.g., Kojima et al., 2022; Wei et al., 2022). Yet, more recently, there is growing awareness that a safe application of LMs may require more control over the reasoning process behind the output generation than can be guaranteed by a single forward pass. Consequently, LMs are increasingly used as parts of larger applications, ranging from relatively simple, general-purpose systems like *retrieval augmented generation* (RAG) (Li et al., 2022), via more open-ended problem-solving strategies like *tree of thought reasoning* (S. Yao et al., 2023), to full fledged

agent models for planning, interaction or robotic control (Huang et al., 2022; Park et al., 2023; Richter et al., 2023). These LM-fueled applications are, essentially, constitutive of (yet) another wave of interest in hybrid neuro-symbolic models (Garcez and Lamb, 2020).

This development is highly relevant for research at the interface between NLP and the language sciences, for example, in the contact area of **linguistic pragmatics**. For one, specialized complex tasks may benefit from theoretically-informed task decomposition, as supplied by recent pragmatic models (e.g., M. C. Frank and Goodman, 2012; Franke and Jäger, 2016). Relevant applications combining linguistic task-analysis with neural network components include structured question-answering (Bosselut et al., 2021) and pragmatic language generation (Andreas and Klein, 2016; Shen et al., 2019; Zarrieß and Schlangen, 2019). For another, it becomes attractive to explore **neuro-symbolic cognitive models designed to predict or explain human reasoning and choice behavior**. Such hybrid models have a long history and have recently started to integrate LMs as generative components, e.g., for generation of categorical contingencies for probabilistic reasoning (Lew et al., 2020; Tsvilodub, Franke, and Carcassi, 2024), translation of natural language to code for structured inference (Wong et al., 2023), or as scoring functions, supplying human-like notions, e.g., of relevance, for downstream numerical computation (Park et al., 2023; Zhang et al., 2023). Yet, for each contribution of an LM inside a neuro-symbolic model we must ascertain that its contribution is what we expect it do be (see ❖BehAss). Indeed, it is not obvious how predictions from LMs should be supplied inside of models that aspire to predict likelihoods for empirical data (Franke, Tsvilodub, et al., 2024), so that it requires careful methodological reflection on how trustworthy the evidence accrued by such hybrid models is for scientific inquiry (see ❖ExplPot). Finally, since neuro-symbolic models often have a modular internal structure (Fodor, 1983), this ties in to recent discussions after modularity in human (pragmatic) reasoning (Allott, 2023) and the explanatory potential of inherently modular alternative neural architectures as ways of achieving higher cognitive plausibility, generalizability and mechanistic interpretability (Ponti et al., 2023) (❖AltMod).

❖ **Resource Efficiency.** Modern LMs are resource-demanding, requiring huge data sets for training, as well as massive computational power, memory, and energy for training and inference. While scaling laws predict that increasing the size of models and training data will lead to increased performance, there is also a growing emphasis on developing smaller, more efficient models at similar performance with reduced computational costs (e.g., Touvron et al., 2023). Yet, there are areas where resource demands arguably cannot easily be solved by engineering solutions alone, e.g., in cases of low-resource languages with a small number of speakers or less recorded data in digital form.

Sparse resources are a chief concern for **applications in comparative-historical linguistics and typological linguistics**. There are only about 6,000 extant languages, and the number of languages with sizable documentation is at least an order of magnitude smaller. Also, data availability across languages follows a Zipfian distribution, with the vast majority being concentrated on a small number of languages. Yet it is also in these areas that integration of LM-based research methods promises to be very fruitful. For example, the prevalent practice in computational historical linguistics for automatically inferring language family trees heavily relies on manual annotation of cognate words, which creates a bottleneck of data sparseness (Jäger, 2019). Machine Learning methods hold great promise in **automatizing annotation of cognate words** (e.g., Jäger et al., 2017; List et al., 2021). Recent research has shown that transformer-based deep learning models improve upon the state of the art (Akavarapu and Bhattacharya, 2024). Going forward, it is promising to develop methods for **extracting phylogenetically informative features** from data sources that are both ecologically more realistic and easier to obtain than word lists, such as texts and sound recordings. Moreover, language technology can also aid the **documentation of endangered and low-resource languages** (e.g., ImaniGooghari et al., 2023; Tanzer et al., 2023). An essential step in language documentation is the compilation of reference grammars which are understandable by humans and can be used as stepping stone for downstream tasks such as language preservation/revitalization efforts. LMs hold a great potential in assisting grammar compilation.

Fine-tuning an LM for a target language involves the implicit induction of grammatical regularities, which can be extracted explicitly in a later step. This second step of inferring explicit grammatical generalizations from interviews with native speakers / trained LMs can likewise be assisted by LM technology. Yet, all of these examples naturally presuppose that the tools are appropriate for the specific purposes (see ❖BehAss) and are used appropriately for knowledge gain (see ❖ExplPot).

❖ **Alternative Models.** The primary focus of recent language models has been on text-to-text tasks (Touvron et al., 2023; Brown et al., 2020). A common criticism of text-only LMs is that they may lack genuine understanding due to the absence of grounding in real-world experiences (Bender and Koller, 2020; Yiu et al., 2023). Earlier, theoretically driven applications have incorporated visual elements, such as images or video, as forms of textual grounding (e.g., Mao et al., 2016), and customer-facing products like ChatGPT and Claude have integrated recognition and generation of other modalities, like pictures as well. But **recent vision-and-language models** (like CLIP Radford et al., 2021) bring their own challenges for important linguistic issues, such as compositionality (Thrush et al., 2022) or pragmatic understanding (Bavaresco, Testoni, et al., 2024) (see ❖BehAss ❖RepMec), and it remains unclear whether these models can meet the philosophical criteria for understanding (Chalmers, 2023; Schlangen, 2023b) or more accurately represent human language abilities (Mahowald et al., 2024).

An exciting area of productive work at the interface between language technology and the language sciences is research on **multi-modality, interaction and dialogue**. Indeed, a significant limitation of text-to-text models, when compared to human language processing, is the omission of supplementary cues such as intonation, prosody, or accompanying gestures and facial expressions (Vigliocco et al., 2014). Furthermore, current chat models are far from functioning as full-fledged dialogue agents (Kopp and Krämer, 2021; Schlangen, 2022, 2023c). Therefore, **it is important to consider alternative model architectures, data sources, and training objectives, which may be more conducive to the scientific investigation of human language than current standard language models**. Alternative approaches may explore different training objectives, such as based on interaction with a static environment (Hill et al., 2020) or genuine interaction (Chalamalasetti et al., 2023; Schlangen, 2023a). Another direction is to study language abilities emerging in the context of simulated linguistic interaction (Lazaridou, Potapenko, et al., 2020; Lazaridou and Baroni, 2020; Ohmer et al., 2022; Tsvilodub and Franke, 2023). In sum, tying in with the general reflection on the nature of current language technology (see ❖OntStat) it requires critical reflection on scientifically more useful modeling architectures (see related remarks on ❖RepMec and ❖TaskDec).

❖ **Ontological Status.** Many concrete practical problems of understanding and applying modern language technology root in foundational questions of what exactly LMs are supposed to be. There is controversy about whether current LMs could possibly be considered models or theories of language (Pater, 2019; Potts, 2019; Piantadosi, 2023; Kodner et al., 2023; Katzir, 2023). LaSTing will contribute to foundational conceptual questions relevant for the language sciences, such as: What does modern linguistic theory aspire to explain, and which of these questions are (partially) addressable by current language technology (❖ExplPot)? Which alternative models or technologies would we need, which may not be application-efficient, but would serve as better tools for knowledge gain (see ❖AltMod)? How much must/may we anthropomorphize generative AI in research or in public-facing science communication (Shanahan, 2023; Shanahan et al., 2023)?

These high-level questions impact also very concrete applications and empirical case studies, in particular by asking: **if some safe use of language technology is to be explanatory or predictive, what exactly is it that we are explaining or predicting?** In many cases, this is highly non-trivial. For example, an LM pre-trained on (written) text may be said to capture the occurrence frequencies of (written) text. Later steps of fine-tuning, however, will change the predictions. In what way? Are modern, fine-tuned models predicting, or capable of predicting, behavior of idealized speakers, average speakers, or

stereotypes? These issues show **parallels with foundational questions known from experimental psychology**, when it comes to learning about individual behavior from average performances (e.g., Estes and Maddox, 2005). Questions of individual differences are increasingly in the focus also in empirical studies on language acquisition and processing (Kidd et al., 2018). Recent studies into **impersonation** are beginning to explore similar issues also for language models (e.g., Deshpande et al., 2023; Salewski et al., 2023; Škrjanec et al., 2023). This will likely become even more important also for applications, when it comes to studying how human speakers quickly adapt to their interlocutors (Brennan and Hanna, 2009) and how this can be mirrored in machines for personalized assistants. All of these issues have an empirical and a technical component, and they reach deep into concerns of other core issues (see ❖BehAss and ❖RepMec), showing that structures that facilitate cross-disciplinary effort, like the proposed Priority Programme, are required for progress.

❖ **Explanatory Potential.** As described with several concrete examples above, language technology can be used as a tool in linguistic research in multiple ways. Looking into future possibilities, LMs could be used, for instance, to automatize the creation of (text-based) experimental stimuli, e.g., for use in reading studies. Creation of experimental stimuli by human experts can introduce biases (cf., H. H. Clark, 1973), so we should be open-minded but careful concerning the prospects of synthetically generating experimental materials. Another context in which LMs may aid, but also influence empirical studies in the language sciences is where LMs are used as partners for dynamic interaction with human participants. Naturally, all of these use cases must all be scrutinized, conceptually and empirically, for machine-induced biases that may affect the eventual empirical results.

The case for careful scrutiny is even more pressing when data is created that itself is analyzed, such as to overcome problems of data-sparsity in low-resource cases (see ❖ResEff). An important foundational question that the field needs to address is **when to accept machine judgements as scientific evidence** for or against a theoretical idea, hypothesis or explanation. For example, the experimental turn in linguistics has brought a vivid discussion about whether, e.g., in syntax, empirical acceptability judgements should inform theoretical accounts (Sprouse, 2023). Similar questions will have to be answered by the concerned community about the evidential value of machine output, too (e.g., Tsvilodub, Marty, et al., 2024). This necessary foundational discussion will partly overlap with the question of whether it is acceptable to replace human subjects with LMs in psychological experiments (Aher et al., 2023; Dillion et al., 2023; Harding et al., 2023), but the case is much more subtle and important for the language sciences, where that which the technology is aimed to capture is (part of) the object of inquiry of the scientific discipline. Depending on what we may safely take a model's prediction to be about (see ❖OntStat), it may be prudent to accept the judgements of a powerful statistical learner as (additional) evidence for claims about, say, language structure. However, even if we should *not* consider language model predictions as theoretical evidence, these are matters that require gradual consensus formation in the community, which must be triggered by increased awareness of these issues and structures that facilitate repeated exchange, as promoted by this Priority Programme.

Aher, Gati et al. (2023). "Using large language models to simulate multiple humans and replicate human subject studies". In: *Proceedings of the 40th International Conference on Machine Learning*.

Akavarapu, V.S.D.S.Mahesh and Arnab Bhattacharya (2024). "Automated Cognate Detection as a Supervised Link Prediction Task with Cognate Transformer". In: *Proceedings of the EACL 18*, pp. 965–975.

Allott, Nicholas (2023). "Encapsulation, inference and utterance interpretation". In: *Inquiry*, pp. 1–35. doi: 10.1080/0020174x.2023.2267084.

Andreas, Jacob and Dan Klein (2016). "Reasoning about pragmatics with neural listeners and speakers". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1173–1182.

Antonello, Richard and Alexander Huth (2023). "Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data". In: *Neurobiology of Language*, pp. 1–16. doi: 10.1162/nol_a_00087.

Bavaresco, Anna, Raffaella Bernardi, et al. (2024). *LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks*. arXiv: 2406.18403.

Bavaresco, Anna, Alberto Testoni, et al. (2024). “Don’t Buy it! Reassessing the Ad Understanding Abilities of Contrastive Multimodal Models”. In: *Proceedings ACL 62 (Short Papers)*, pp. 870–879. doi: 10.18653/v1/2024.acl-short.77.

Beaver, David I. et al. (2017). “Questions Under Discussion: Where Information Structure Meets Projective Content”. In: *Annual Review of Linguistics* 3.1, pp. 265–284. doi: 10.1146/annurev-linguistics-011516-033952.

Beck, Maximilian et al. (2024). *xLSTM: Extended Long Short-Term Memory*. doi: 10.48550/ARXIV.2405.04517.

Beinborn, Lisa et al. (2023). “Robust Evaluation of Language–Brain Encoding Experiments”. In: *Computational Linguistics and Intelligent Text Processing*, pp. 44–61.

Bender, Emily M. and Alexander Koller (2020). “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*, pp. 5185–5198. doi: 10.18653/v1/2020.acl-main.463.

Bengio, Yoshua et al. (2003). “A Neural Probabilistic Language Model”. In: *Journal of Artificial Intelligence Research* 3, pp. 1137–1155.

BIG-bench authors (2023). “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. In: *Transactions on Machine Learning Research*.

Binz, Marcel and Eric Schulz (2023). “Using cognitive psychology to understand GPT-3”. In: *Proceedings of the National Academy of Sciences* 120.6, e2218523120. doi: 10.1073/pnas.2218523120. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2218523120.

Bosselut, Antoine et al. (2021). “Dynamic Neuro-Symbolic Knowledge Graph Construction for Zero-shot Commonsense Question Answering”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.6, pp. 4923–4931. doi: 10.1609/aaai.v35i6.16625.

Brennan, Susan E. and Joy E. Hanna (2009). “Partner-Specific Adaptation in Dialog”. In: *Topics in Cognitive Science* 1.2, pp. 274–291. doi: 10.1111/j.1756-8765.2009.01019.x.

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165.

Buccola, Brian et al. (2021). “Conceptual alternatives: Competition in language and beyond”. In: *Linguistics and Philosophy* 45.2, pp. 265–291. doi: 10.1007/s10988-021-09327-w.

Bunzeck, Bastian and Sina Zarrieß (2023). “GPT-wee: How Small Can a Small Language Model Really Get?” In: *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pp. 35–46. doi: 10.18653/v1/2023.conll-babylm.2.

Caucheteux, Charlotte and Jean-Rémi King (2022). “Brains and algorithms partially converge in natural language processing”. In: *Communications Biology* 5.1. doi: 10.1038/s42003-022-03036-1.

Chalamalasetti, Kranti et al. (2023). “clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11174–11219. doi: 10.18653/v1/2023.emnlp-main.689.

Chalmers, David J. (2023). *Could a Large Language Model be Conscious?* arXiv: 2303.07103 [cs.AI].

Chambers, Chris (2017). *The Seven Deadly Sins of Psychology*. Princeton University Press. doi: 10.2307/j.ctvc779w5.

Clark, Herbert H. (1973). “The language-as-fixed-effect fallacy: A critique of language statistics in psychological research”. In: *Journal of Verbal Learning and Verbal Behavior* 12.4, pp. 335–359. doi: 10.1016/s0022-5371(73)80014-3.

Clark, Kevin et al. (2019). “What Does BERT Look at? An Analysis of BERT’s Attention”. In: *BlackboxNLP@ACL*.

Crocker, Matthew (2010). “Computational Psycholinguistics”. In: *Computational Linguistics and Natural Language Processing*, pp. 482–513.

Culbertson, Jennifer and Kathryn Schuler (2019). “Artificial Language Learning in Children”. In: *Annual Review of Linguistics* 5.1, pp. 353–373. doi: 10.1146/annurev-linguistics-011718-012329.

Demberg, Vera and Frank Keller (2008). “Data from eye-tracking corpora as evidence for theories of syntactic processing complexity”. In: *Cognition* 109.2, pp. 193–210. doi: https://doi.org/10.1016/j.cognition.2008.07.008.

Deshpande, Ameet et al. (2023). “Toxicity in chatgpt: Analyzing persona-assigned language models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1236–1270. doi: 10.18653/v1/2023.findings-emnlp.88.

Dillion, Danica et al. (2023). “Can AI language models replace human participants?” In: *Trends in Cognitive Sciences* 27.7, pp. 597–600. doi: 10.1016/j.tics.2023.04.008.

Dotlačil, Jakub and Adrian Brasoveanu (2020). *Computational Cognitive Modeling and Linguistic Theory*. Springer.

Elazar, Yanai et al. (2021). “Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals”. In: *Transactions of the Association for Computational Linguistics* 9, pp. 160–175. doi: 10.1162/tacl_a_00359.

Elman, Jeffrey L. (1990). “Finding Structure in Time”. In: *Cognitive Science* 14.2, pp. 179–211. doi: 10.1207/s15516709cog1402_1.

Erk, Katrin (2012). “Vector Space Models of Word Meaning and Phrase Meaning: A Survey”. In: *Language and Linguistics Compass* 6.10, 635â□□653. doi: 10.1002/lnco.362.

— (2022). “The Probabilistic Turn in Semantics and Pragmatics”. In: *Annual Review of Linguistics* 8.1. **Erk**, pp. 101–121. doi: 10.1146/annurev-linguistics-031120-015515.

Estes, William K. and W. Todd Maddox (2005). “Risks of Drawing Inferences about Cognitive Processes from Model Fits to Individual versus Average Performance”. In: *Psychonomic Bulletin & Review* 12.3, pp. 403–408.

Fodor, Jerry (1983). *The Modularity of Mind*. MIT Press.
Fox, Danny and Roni Katzir (2024). "Large Language Models and theoretical linguistics". In: *Theoretical Linguistics* 50.1–2, pp. 71–76. doi: 10.1515/tl-2024-2005.
Frank, Michael C. (2023). "Large language models as models of human cognition". doi: 10.31234/osf.io/wxt69.
Frank, Michael C. and Noah D. Goodman (2012). "Predicting Pragmatic Reasoning in Language Games". In: *Science* 336.6084, p. 998. doi: 10.1126/science.1218633.
Frank, Stefan L. et al. (2015). "The ERP response to the amount of information conveyed by words in sentences". In: *Brain and Language* 140, pp. 1–11. doi: 10.1016/j.bandl.2014.10.006.
Franke, Michael and Gerhard Jäger (2016). "Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics". In: *Zeitschrift für Sprachwissenschaft* 35.1, pp. 3–44. doi: 10.1515/zfs-2016-0002.
Franke, Michael, Polina Tsvilodub, et al. (2024). *Bayesian Statistical Modeling with Predictors from LLMs*. arXiv: 2406.09012.
Fresen, Abraham Jacob et al. (2024). "Language Models That Accurately Represent Syntactic Structure Exhibit Higher Representational Similarity To Brain Activity". In: *Proceedings of CogSci 46*, pp. 675–683.
Gandhi, Kanishk et al. (2023). *Understanding Social Reasoning in Language Models with Language Models*. arXiv: 2306.15448.
Garcez, Artur d'Avila and Luis C. Lamb (2020). *Neurosymbolic AI: The 3rd Wave*. arXiv: 2012.05876 [cs.AI].
Gauthier, Jon and Roger Levy (2019). "Linking artificial and human neural representations of language". 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)*, pp. 529–539. doi: 10.18653/v1/D19-1050.
Geiger, Atticus et al. (2021). *Causal Abstractions of Neural Networks*. arXiv: 2106.02997 [cs.AI].
Gerven, Marcel A. J. van et al. (2019). "Current Advances in Neural Decoding". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 379–394. doi: 10.1007/978-3-030-28954-6_21.
Gotzner, Nicole and Jacopo Romoli (2022). "Meaning and Alternatives". In: *Annual Review of Linguistics* 8.1, pp. 213–234. doi: 10.1146/annurev-linguistics-031220-012013.
Gu, Albert and Tri Dao (2023). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. doi: 10.48550/ARXIV.2312.00752.
Hagendorff, Thilo (2023). *Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods*. arXiv: 2303.13988.
Hahn, Michael (2020). "Theoretical Limitations of Self-Attention in Neural Sequence Models". English. In: *Transactions of the Association for Computational Linguistics* 8, pp. 156–171. doi: 10.1162/tacl_a_00306.
Harding, Jacqueline (to appear). "Operationalising Representation in Natural Language Processing". In: *The British Journal for the Philosophy of Science*. doi: 10.1086/728685. eprint: https://doi.org/10.1086/728685.
Harding, Jacqueline et al. (2023). "AI language models cannot replace human research participants". In: *AI & SOCIETY*. doi: 10.1007/s00146-023-01725-x.
Hewitt, John and Christopher D. Manning (2019). "A Structural Probe for Finding Syntax in Word Representations". In: *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)*, pp. 4129–4138. doi: 10.18653/v1/N19-1419.
Hill, Felix et al. (2020). "Grounded Language Learning Fast and Slow". Technical report.
Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735.
Holtzman, Ari et al. (2021). "Surface Form Competition: Why the Highest Probability Answer Isn't Always Right". In: *Proceedings of EMNLP*, pp. 7038–7051. doi: 10.18653/v1/2021.emnlp-main.564.
Hu, Jennifer and Michael C. Frank (2024). *Auxiliary task demands mask the capabilities of smaller language models*. arXiv: 2404.02418.
Hu, Jennifer, Jon Gauthier, et al. (2020). "A Systematic Assessment of Syntactic Generalization in Neural Language Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1725–1744. doi: 10.18653/v1/2020.acl-main.158.
Hu, Jennifer and Roger Levy (2023). *Prompt-based methods may underestimate large language models' linguistic generalizations*. arXiv: 2305.13264.
Huang, Wenlong et al. (2022). *Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents*. arXiv: 2201.07207 [cs.LG].
Hupkes, Dieuwke et al. (2020). "Compositionality decomposed: how do neural networks generalise?" In: *Journal of Artificial Intelligence Research* 67.
ImaniGooghari, Ayyoob et al. (2023). "Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages". In: *Proceedings of ACL 61 (Long Papers)*, pp. 1082–1117.
Jäger, Gerhard (2019). "Computational historical linguistics". In: *Theoretical Linguistics* 45.3-4, pp. 151–182. doi: doi:10.1515/tl-2019-0011.
Jäger, Gerhard et al. (2017). "Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1205–1216.

Kallini, Julie et al. (2024). “Mission: Impossible Language Models”. In: *Proceedings of ACL 62 (Long Papers)*, pp. 14691–14714. doi: 10.18653/v1/2024.acl-long.787.

Katzir, Roni (2007). “Structurally-Defined Alternatives”. In: *Linguistics and Philosophy* 30.6, pp. 669–690. doi: doi:10.1007/s10988-008-9029-y.

— (2023). “Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi”. In: *Biolinguistics* 17. doi: 10.5964/bioling.13153.

Kauf, Carina et al. (2024). “Lexical-Semantic Content, Not Syntactic Structure, Is the Main Contributor to ANN-Brain Similarity of fMRI Responses in the Language Network”. In: *Neurobiology of Language* 5.1, pp. 7–42. doi: 10.1162/nol_a_00116.

Kidd, Evan et al. (2018). “Individual Differences in Language Acquisition and Processing”. In: *Trends in Cognitive Sciences* 22.2, pp. 154–169. doi: 10.1016/j.tics.2017.11.006.

Kim, Been et al. (2018). “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80, pp. 2668–2677.

Kodner, Jordan et al. (2023). *Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi (2023)*. arXiv: 2308.03228.

Kojima, Takeshi et al. (2022). *Large Language Models are Zero-Shot Reasoners*. doi: 10.48550/ARXIV.2205.11916.

Kopp, Stefan and Nicole Krämer (2021). “Revisiting Human-Agent Communication: The Importance of Joint Co-construction and Understanding Mental States”. In: *Frontiers in Psychology* 12. doi: 10.3389/fpsyg.2021.580955.

Kumar, Sawan (2022). “Answer-level Calibration for Free-form Multiple Choice Question Answering”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 665–679. doi: 10.18653/v1/2022.acl-long.49.

Lampinen, Andrew Kyle (2023). *Can language models handle recursively nested grammatical structures? A case study on comparing models and humans*. arXiv: 2210.15303.

Lazaridou, Angeliki and Marco Baroni (2020). *Emergent Multi-Agent Communication in the Deep Learning Era*. arXiv: 2006.02419.

Lazaridou, Angeliki, Anna Potapenko, et al. (2020). “Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7663–7674.

Leidinger, Alina et al. (2023). “The language of prompting: What linguistic properties make a prompt successful?” In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9210–9232. doi: 10.18653/v1/2023.findings-emnlp.618.

Levy, Roger (2008). “Expectation-Based Syntactic Comprehension”. In: *Cognition* 106, pp. 1126–1177. doi: 10.1016/j.cognition.2007.05.006.

Lew, Alexander K. et al. (2020). “Leveraging Unstructured Statistical Knowledge in a Probabilistic Language of Thought”. In: *Proceedings of CogSci 42*, pp. 2223–2229.

Li, Huayang et al. (2022). *A Survey on Retrieval-Augmented Text Generation*. arXiv: 2202.01110.

Linzen, Tal et al. (2016). “Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies”. In: *Transactions of the Association for Computational Linguistics* 4, pp. 521–535. doi: 10.1162/tacl_a_00115.

List, Johann-Mattis et al. (2021). “Lexibank, a public repository of standardized wordlists with computed phonological and lexical features”. In: *Scientific Data* 9.

Liu, Tong et al. (2024). “Temperature-scaling surprisal estimates improve fit to human reading times – but does it do so for the “right reasons”?” In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9598–9619. doi: 10.18653/v1/2024.acl-long.519.

Mahowald, Kyle et al. (2024). “Dissociating language and thought in large language models”. In: *Trends in Cognitive Sciences* 28.6, pp. 517–540. doi: 10.1016/j.tics.2024.01.011.

Mao, Junhua et al. (2016). “Generation and Comprehension of Unambiguous Object Descriptions”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20. doi: 10.1109/CVPR.2016.9.

Marks, Samuel et al. (2024). *Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models*. arXiv: 2403.19647 [cs.LG].

McCoy, R. Thomas et al. (2024). “Embers of autoregression show how large language models are shaped by the problem they are trained to solve”. In: *Proceedings of the National Academy of Sciences* 121.41. doi: 10.1073/pnas.2322420121.

Meibauer, Jörg and Markus Steinbach, eds. (2011). *Experimental Pragmatics/Semantics*. John Benjamins.

Meng, Kevin et al. (2024). “Locating and editing factual associations in GPT”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Merullo, Jack et al. (2024). *Circuit Component Reuse Across Tasks in Transformer Language Models*. arXiv: 2310.08744.

Mikolov, Tomas et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781.

Misra, Kanishka and Kyle Mahowald (2024). *Language Models Learn Rare Phenomena from Less Rare Phenomena: The Case of the Missing AANNs*. arXiv: 2403.19827 [cs.CL].

Mondorf, Philipp and Barbara Plank (2024). *Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models – A Survey*. arXiv: 2404.01869.

Müller-Eberstein, Max et al. (2022). “Probing for Labeled Dependency Trees”. In: *Proceedings of ACL 60th (Long Papers)*, pp. 7711–7726. doi: 10.18653/v1/2022.acl-long.532.
nostalgebraist (2020). “Interpreting GTP: The Logit Lens”. Blog post on *Less Wrong*.
Noveck, Ira A. (2018). *Experimental Pragmatics: The Making of a Cognitive Science*. Cambridge University Press.
Oh, Byung-Doh and William Schuler (2023). “Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?” In: *Transactions of the Association for Computational Linguistics* 11, pp. 336–350. doi: 10.1162/tacl_a_00548.
— (2024). *Leading Whitespaces of Language Models’ Subword Vocabulary Poses a Confound for Calculating Word Probabilities*. arXiv: 2406.10851.
Ohmer, Xenia et al. (2022). “Mutual influence between language and perception in multi-agent communication games”. In: *PLoS Computational Biology* 18.10, e1010658. doi: 10.1371/journal.pcbi.1010658.
Ouyang, Long et al. (2022). “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 27730–27744.
Park, Joon Sung et al. (2023). *Generative Agents: Interactive Simulacra of Human Behavior*. arXiv: 2304.03442 [cs.HC].
Pater, Joe (2019). “Generative linguistics and neural networks at 60: Foundation, friction, and fusion”. In: *Language* 95.1, e41–e74. doi: 10.1353/lan.2019.0009.
Peters, Matthew E. et al. (2018). “Deep Contextualized Word Representations”. In: *Proceedings of NACL: Human Language Technologies*, pp. 2227–2237. doi: 10.18653/v1/N18-1202.
Piantadosi, Steven T. (2023). *Modern language models refute Chomsky’s approach to language*. lingbuzz/007180.
Pimentel, Tiago and Clara Meister (2024). *How to Compute the Probability of a Word*. arXiv: 2406.14561.
Ponti, Edoardo Maria et al. (2023). “Combining Parameter-efficient Modules for Task-level Generalisation”. In: *Proceedings EACL 17*, pp. 687–702. doi: 10.18653/v1/2023.eacl-main.49.
Potts, Christopher (2019). “A case for deep learning in semantics: Response to Pater”. In: *Language* 1, e115–e124. doi: https://doi.org/10.1353/lan.2019.0019.
Radford, Alec et al. (2021). “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139, pp. 8748–8763.
Rahwan, Iyad et al. (2019). “Machine behaviour”. In: *Nature* 568.7753, pp. 477–486. doi: 10.1038/s41586-019-1138-y.
Richter, brian et al. (2023). “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances”. In: *Proceedings of The 6th Conference on Robot Learning*. Vol. 205, pp. 287–318.
Rohde, Hannah and Chigusa Kurumada (2018). “Alternatives and inferences in the communication of meaning”. In: *Current Topics in Language*, pp. 215–261. doi: 10.1016/bs.plm.2018.08.012.
Salewski, Leonard et al. (2023). “In-Context Impersonation Reveals Large Language Models’ Strengths and Biases”. In: *Thirty-seventh Conference on Neural Information Processing Systems*.
Schlangen, David (2022). “Norm Participation Grounds Language”. In: *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pp. 62–69.
— (2023a). “Dialogue Games for Benchmarking Language Understanding: Motivation, Taxonomy, Strategy”. In: *CoRR*. doi: 10.48550/arXiv.2304.07007. arXiv: 2304.07007.
— (2023b). *On General Language Understanding*. arXiv: 2310.18038.
— (2023c). “What A Situated Language-Using Agent Must be Able to Do: A Top-Down Analysis”. In: *CoRR*. doi: 10.48550/arXiv.2302.08590. arXiv: 2302.08590.
Schrimpf, Martin et al. (2021). “The neural architecture of language: Integrative modeling converges on predictive processing”. In: *Proceedings of the National Academy of Sciences* 118.45. doi: 10.1073/pnas.2105646118.
Shanahan, Murray (2023). *Talking About Large Language Models*. arXiv: 2212.03551.
Shanahan, Murray et al. (2023). “Role play with large language models”. In: *Nature* 623.7987, pp. 493–498. doi: 10.1038/s41586-023-06647-8.
Shen, Sheng et al. (2019). “Pragmatically Informative Text Generation”. In: *Proceedings of NAACL*.
Shiffrin, Richard and Melanie Mitchell (2023). “Probing the psychology of AI models”. In: *Proceedings of the National Academy of Sciences* 120.10, e2300963120. doi: 10.1073/pnas.2300963120.
Si, Chenglei et al. (2022). “Re-Examining Calibration: The Case of Question Answering”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2814–2829. doi: 10.18653/v1/2022.findings-emnlp.204.
Škrjanec, Iza et al. (2023). “Expert-adapted language models improve the fit to reading times”. In: *Procedia Computer Science* 225, pp. 3488–3497.
Smith, Nathaniel J. and Roger Levy (2013). “The Effect of Word Predictability on Reading Time is Logarithmic”. In: *Cognition* 128, pp. 302–319. doi: 10.1016/j.cognition.2013.02.013.
Sprouse, Jon, ed. (2023). *The Oxford Handbook of Experimental Syntax*. Oxford University Press.
Tanzer, Garrett et al. (2023). “A benchmark for learning to translate a new language from one grammar book”. In: *arXiv preprint arXiv:2309.16575*.
Tenney, Ian et al. (2019). “BERT Rediscovers the Classical NLP Pipeline”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601. doi: 10.18653/v1/P19-1452.

Thrush, Tristan et al. (2022). “Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238.

Touvron, Hugo et al. (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971.

Tsvilodub, Polina and Michael Franke (2023). “Evaluating pragmatic abilities of image captioners on A3DS”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1277–1285.

Tsvilodub, Polina, Michael Franke, and Fausto Carcassi (2024). *Cognitive Modeling with Scaffolded LLMs: A Case Study of Referential Expression Generation*. arXiv: 2407.03805.

Tsvilodub, Polina, Paul Marty, et al. (2024). “Experimental Pragmatics with Machines: Testing LLM Predictions for the Inferences of Plain and Embedded Disjunctions”. In: *Proceedings of CogSci*, pp. 3960–3967.

Tsvilodub, Polina, Hening Wang, et al. (2024). *Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods*. arXiv: 2403.00998.

Vaswani, Ashish et al. (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30.

Vig, Jesse et al. (2020). “Investigating Gender Bias in Language Models Using Causal Mediation Analysis”. In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 12388–12401.

Vigliocco, Gabriella et al. (2014). “Language as a multimodal phenomenon: implications for language learning, processing and evolution”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1651, p. 20130292. doi: 10.1098/rstb.2013.0292.

Wahle, Jan Philip et al. (2023). “We are Who We Cite: Bridges of Influence Between Natural Language Processing and Other Academic Fields”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12896–12913. doi: 10.18653/v1/2023.emnlp-main.797.

Waldis, Andreas et al. (2024). *Holmes: Benchmark the Linguistic Competence of Language Models*. arXiv: 2404.18923.

Wang, Kevin et al. (2022). *Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small*. arXiv: 2211.00593 [cs.LG].

Warstadt, Alex and Samuel R. Bowman (2024). *What Artificial Neural Networks Can Tell Us About Human Language Acquisition*. arXiv: 2208.07998.

Warstadt, Alex, Aaron Mueller, et al. (2023). “Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora”. In: *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pp. 1–34. doi: 10.18653/v1/2023.conll-babylm.1.

Webson, Albert and Ellie Pavlick (2022). “Do Prompt-Based Models Really Understand the Meaning of Their Prompts?” In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2300–2344. doi: 10.18653/v1/2022.naacl-main.167.

Wei, Jason et al. (2022). “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 24824–24837.

Wieling, Martijn et al. (2018). “Reproducibility in Computational Linguistics: Are We Willing to Share?” In: *Computational Linguistics* 44.4, pp. 641–649. doi: 10.1162/coli_a_00330.

Wilcox, Ethan et al. (2021). “A Targeted Assessment of Incremental Processing in Neural Language Models and Humans”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 939–952. doi: 10.18653/v1/2021.acl-long.76.

Wong, Lionel et al. (2023). *From Word Models to World Models: Translating from Natural Language to the Probabilistic Language of Thought*. arXiv: 2306.12672.

Yao, Shunyu et al. (2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. arXiv: 2305.10601.

Yao, Yuekun and Alexander Koller (2022). “Structural generalization is hard for sequence-to-sequence models”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5048–5062. doi: 10.18653/v1/2022.emnlp-main.337.

Yin, Kayo and Graham Neubig (2022). “Interpreting Language Models with Contrastive Explanations”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 184–198.

Yiu, Eunice et al. (2023). *Imitation versus Innovation: What children can do that large language and language-and-vision models cannot (yet)?* arXiv: 2305.07666 [cs.AI].

Zarrieß, Sina and David Schlangen (2019). “Know What You Don’t Know: Modeling a Pragmatic Speaker that Refers to Objects of Unknown Categories”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 654–659. doi: 10.18653/v1/P19-1063.

Zhang, Cedegao E. et al. (2023). “Grounded physical language understanding with probabilistic programs and simulated worlds”. In: *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, pp. 3476–3483.

Zhao, Zihao et al. (2021). “Calibrate Before Use: Improving Few-Shot Performance of Language Models”. In: *Proceedings of the 38th International Conference on Machine Learning*. eprint: 2102.09690.

Ziems, Caleb et al. (2023). *Can Large Language Models Transform Computational Social Science?* arXiv: 2305.03514.