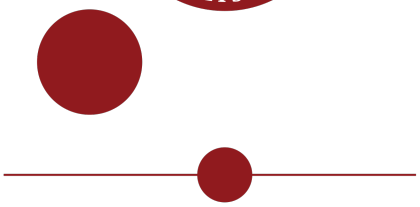
LANGUAGE UNDERSTANDING IN HUMANS AND ARTIFICIAL NEURAL NETWORKS: PARALLELS AND CONTRASTS

MOSTAFA ABDOU

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen.

LANGUAGE UNDERSTANDING IN HUMANS AND ARTIFICIAL NEURAL NETWORKS: PARALLELS AND CONTRASTS

MOSTAFA ABDOU



UNIVERSITY OF
COPENHAGEN

PhD Thesis

November 2021

Mostafa Abdou: *Language Understanding in Humans and Artificial Neural Networks: Parallels and Contrasts*

Thesis Supervisor:
Anders Søgaard

Assessment Committee: - , University of Copenhagen
Tal Linzen, New York University
Leila Wehbe, Carnegie Mellon University

Affiliation:
Department of Computer Science
Faculty of Science
University of Copenhagen

Thesis submitted:
November 30th, 2021

ABSTRACT

Modern artificial neural networks (ANNs) are loosely inspired by the human brain. This begs the question: can they help us model aspects of (human) language understanding by serving as plausible hypotheses for its underlying representations and mechanisms? Although the way modern ANN language models (LMs) learn — via training on immense amounts of text to predict either future or masked tokens — is manifestly not human-like, they have demonstrated a remarkable ability to simulate human understanding on a wide range of tasks. In addition, they have recently been shown to predict or align to a variety of cognitive measurements. This dissertation presents research that investigates ANN LMs, examining the linguistic properties of the representations they acquire and exploring parallels and disparities between their language processing capacities and those of humans. This is done based on three classes of analytical comparisons, viz. (i) behavioural data, (ii) linguistic theory, and (iii) neural response measurements.

For the first class, we (a) analyse the representational similarity between ANN language model activation patterns and eye-tracking data and (b) evaluate the structural alignment of humans' perceptual color space with LM-derived color name representations. In both cases, we find interesting correspondences.

For the second, we show that ANN LMs (a) when finetuned on task-specific data, are robust to linguistic perturbations that minimally affect human understanding, (b) can learn attention patterns that reflect linguistic structure, and (c) trained on sentences with scrambled word order, still retain a notion of word order derived from statistical cues that persist in the scrambled data, offering an explanation for why they still perform well on language understanding tasks.

For the final class of comparisons, we present a literature review of research linking computational models of language with human neural response measurements and conclude by introducing a framework for leveraging ANN LMs to enable the evaluation of targeted hypotheses about the composition of meaning in the human brain.

Overall, this dissertation works towards furthering our understanding of ANN LMs through comparisons to what we already know about how humans process language and, reciprocally, towards developing frameworks where LMs can help provide insights into human language.

ABSTRACT IN DANISH

Moderne kunstige neurale netværk (KNN) er løst inspireret af den menneskelige hjerne. Dette rejser spørgsmålet: Kan de hjælpe os med at modellere aspekter af (menneskelig) sprogforståelse ved at fungere som plausible hypoteser for dets underliggende repræsentationer og mekanismer? Selvom den måde, moderne KNN-baserede sprogmodeller lærer — via træning på enorme mængder tekst ved at forudsige enten fremtidige eller maskerede sproglige enheder — åbenlyst ikke er menneskelignende, har de vist en bemærkelsesværdig evne til at simulere menneskelig forståelse på en bred vifte af opgaver. Derudover har de for nylig vist sig at forudsige eller tilpasse sig en række kognitive målinger. Denne afhandling præsenterer forskning, der undersøger KNN-baserede sprogmodeller, de sproglige egenskaber ved de lærte repræsentationer og udforsker ligheder og forskelle mellem deres sprogbehandlingskapacitet og menneskers. Dette gøres ud fra tre klasser af analytiske sammenligninger: (i) Adfærdsdata, (ii) lingvistisk teori og (iii) neurale responsmålinger. For den første klasse analyserer vi (a) den repræsentative lighed mellem KNN-baserede sprogmodellers aktiveringsmønstre og eye-tracking data og (b) evaluerer den strukturelle tilpasning af menneskers perceptuelle farverum med repræsentation af farvenavne afledt fra sprogmodeller. I begge tilfælde finder vi interessante paralleller. For det andet viser vi, at KNN-baserede sprogmodeller (a) er robuste over for sproglige forstyrrelser, der påvirker den menneskelige forståelse minimalt, når de er finjusteret på opgavespecifik data, (b) kan lære opmærksomhedsmønstre, der afspejler sproglig struktur og (c) trænet på sætninger med permuteret ordrækkefølge bevarer information om ordrækkefølge, der er afledt af statistiske afhængigheder og signaler, der fortsætter i det permuterede data, hvilket forklarer hvorfor de stadig klarer sig godt på opgaver, der kræver menneskelig sprogforståelse. I den sidste klasse af sammenligninger præsenterer vi en litteraturgennemgang af forskning, der forbinder sprogmodeller med neurale responsmålinger på mennesker og afslutter med at introducere en ramme for udnyttelse af KNN-baserede sprogmodeller for at muliggøre evaluering af mere målrettede hypoteser om sammensætningen af mening i den menneskelige hjerne. Overordnet arbejder denne afhandling mod at fremme vores forståelse af KNN-baserede sprogmodeller gennem sammenligninger med det vi allerede ved om menneskers sprogforståelse og, gensidigt, mod at udvikle rammer, hvor sprogmodeller kan hjælpe med at give indsigt i menneskeligt sprog.

PUBLICATIONS

The papers marked by ∗ are not part of this thesis but were authored during and contributed to the ideas that led to its completion.

Ravishankar, Vinit, Artur Kulmizev, **Abdou, Mostafa**, Anders Søgaard, and Joakim Nivre (Apr. 2021). "Attention Can Reflect Syntactic Structure (If You Let It)." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 3031–3045. DOI: 10.18653/v1/2021.eacl-main.264. URL: https://aclanthology.org/2021.eacl-main.264.

∗Aralikatte, Rahul, **Abdou, Mostafa**, Heather C Lent, Daniel Hershcovich, and Anders Søgaard (2021). "Joint Semantic Analysis with Document-Level Cross-Task Coherence Rewards." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.14, pp. 12516–12525. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17484.

∗Kulmizev, Artur, Vinit Ravishankar, **Abdou, Mostafa**, and Joakim Nivre (July 2020). "Do Neural Language Models Show Preferences for Syntactic Formalisms?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4077–4091. DOI: 10.18653/v1/2020.acl-main.375. URL: https://aclanthology.org/2020.acl-main.375.

∗Liétard, Bastien, Mostafa Abdou, and Anders Søgaard (Nov. 2021). "Do Language Models Know the Way to Rome?" In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 510–517. URL: https://aclanthology.org/2021.blackboxnlp-1.40.

∗Nikolaus, Mitja, **Abdou, Mostafa**, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott (Nov. 2019). "Compositional Generalization in Image Captioning." In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 87–98. DOI: 10.18653/v1/K19-1009. URL: https://aclanthology.org/K19-1009.

∗Vilares, David, **Abdou, Mostafa**, and Anders Søgaard (June 2019). "Better, Faster, Stronger Sequence Tagging Constituent Parsers." 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3372–3383.

DOI: 10.18653/v1/N19-1341. URL: https://aclanthology.org/N19-1341.

Abdou, Mostafa (2021). "Connecting Neural Response measurements & Computational Models of language: a non-comprehensive guide." In: *Currently under review.*

Abdou, Mostafa, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard (Nov. 2021a). "Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color." In: *Proceedings of the 25th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 109–132. URL: https://aclanthology.org/2021.conll-1.9.

Abdou, Mostafa, Artur Kulmizev, Felix Hill, Daniel M. Low, and Anders Søgaard (Nov. 2019). "Higher-order Comparisons of Sentence Encoder Representations." 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 5838–5845. DOI: 10.18653/v1/D19-1593. URL: https://aclanthology.org/D19-1593.

Abdou, Mostafa, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard (July 2020). "The Sensitivity of Language Models and Humans to Winograd Schema Perturbations." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7590–7604. DOI: 10.18653/v1/2020.acl-main.679. URL: https://aclanthology.org/2020.acl-main.679.

Abdou, Mostafa, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard (2021b). "Word Order Does Matter and Shuffled Language Models Know It." In: *Currently under review.*

Abdou, Mostafa, Ana Valeria Gonzalez, Maria Toneva, Daniel Hershcovich, and Anders Søgaard (2021c). "Does injecting linguistic structure into language models lead to better alignment with brain recordings?" In: *Currently under review.*

∗**Abdou, Mostafa**, Cezar Sas, Rahul Aralikatte, Isabelle Augenstein, and Anders Søgaard (Nov. 2019). "X-WikiRE: A Large, Multilingual Resource for Relation Extraction as Machine Comprehension." In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, pp. 265–274. DOI: 10.18653/v1/D19-6130. URL: https://aclanthology.org/D19-6130.

"...and they passed on the painful news by the mail of the sun
in all of the city,
"The moon was killed!"

— Amal Donqol

ACKNOWLEDGMENTS

I left this part to the very end, because in some ways, it is the hardest. I am grateful to countless people for the last three years in ways which are impossible to express in this format (or at all).

I am grateful to everyone at CoAStAL for the challenging and enriching but always amicable environment: Ana, Victor, Lasse, Matt, Marreike, Daniel, Rahul, Yova, Stephanie, Emanuelle, and everyone else! Foremost, I am thankful to my supervisor, Anders, who offered constant guidance, support, perspective, and the freedom to pursue my research interests. To Desmond who always offered wholesome, measured advice and mentorship, I am also particularly thankful.

I am grateful to all my collaborators over these years who introduced me to new ideas and helped clarify and refine my work.

I am grateful to those who read this manuscript and offered comments and thoughts: Vinit, Artur, Ana, Victor, Mitja, and Stephanie.

I am grateful to Vinit and Artur with whom I shared thoughts, took roadtrips, wrote papers, ran, rowed, devised theories, and somehow spoke almost everyday for the past four years: "*I arrived to the windy city as a tourist and I leave as a reactionary...*".

I am grateful to those who made this city a home: Anna, Mourad, Lisa, Lucia, Lea, Jona, Pascal, David, Diego, Daphne, Karim, Andrea, Francesco, Davide, Martina, Lydia, and to those with home I share a home: Dan, David, Astrid, Kirsten and to friends everywhere else who are nevertheless always present: Mitja, Jean, Dani, Sabri, Hamada, Etman, Hazem, Tata, Yehia, Titi, and countless others.

And finally, I am more than ever grateful for my family whom I continue to be indebted for the support, comfort, care, grounding and everything else: Mariam, Ibrahim, my endlessly generous dad and mom, to teta, to 3ameto, khalto, 3amo, khalo, and to my cousins.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

NLP Natural Language Processing

ANN Artificial Neural Network

LM Language Model

NLU Natural Language Understanding

RSA Representational Similarity Analysis

WSC Winograd Schema Challenge

fMRI Functional magnetic resonance imaging

PMI Pointwise Mutual Information

UD Universal Dependencies

Part I

INTRODUCTION AND BACKGROUND

1 INTRODUCTION

The systematic study of language dates back to some time around the 6th century BC, from which time we have records of the grammarian Panini's formal description of Sanskrit, and of documents describing Sumerian, various Chinese languages, and somewhat later, Arabic, Hebrew, and Greek.

Modern linguistics can be traced back to the 19th century, and was marked by a shift in focus towards the structural features of language, putting emphasis on identifying, isolating, and analysing the different units which carry linguistic significance, e.g. phonemes, morphemes, syntactic constituents, etc. These "mechanistic" analytical procedures (Bloomfield, 1926) used for the description of language structures were followed by Chomsky (1957)'s "generative grammar" which consisted of a precisely formulated set of rules whose output is all (and only) the sentences of a language. These sets of rules can be seen as *models* of language in some sense, making predictions about what can be considered a grammatical utterance. Such models, and much of the work that followed, was primarily concerned with the careful description linguistic structure and syntax, and did not take the *meaning* of an utterance into consideration.[1]

Leaping forward to the present day, we find that the field of Natural Language Processing (NLP) has developed models of language which are able to simulate (aspects of) human understanding in a variety of settings. Unlike the models of syntax mentioned above, these models incorporate no explicit rules. Instead, they rely on *distributional information*: the statistical patterns and dependencies present in a corpus of text (Brown et al., 1990; Mikolov et al., 2013; Schütze, 1993). This class of model, loosely inspired by biological brains, is known as Artificial Neural Networks (ANNs).[2] The way they encode information involves the optimization of millions (or billions) of weights over a large number of observed examples. To optimize these weights, the model is "trained" to perform a certain task, such as predicting the upcoming words in a paragraph, or labelling the part of speech category of a word in a sentence. An *objective function* is defined that, when applied to a set of observations and the corresponding predictions a model makes, defines a trajectory towards a minimum in the

1 Chomsky argues, for instance, that it is "relatively useless" to use meaning "as a basis for grammatical description"(Chomsky, 1957).

2 Note that there are other classes of NLP models which also rely on distributional information, but we focus on ANN-based models as they are the most relevant to the work carried out in this dissertation.

high-dimensional vector space parametrized by the model's weights, minimizing the error in in its predictions.

The grounding of such ANN language models (LMs) in linguistic theory is tenuous.[3] Indeed, they hardcode very little to nothing in terms of linguistic rules, definitions, or formal structures. Nevertheless, they have (a) demonstrated an impressive ability to simulate understanding (Wang et al., 2019b, 2018a); (b) been found to encode a breadth of linguistic information (Belinkov and Glass, 2019a; Manning et al., 2020); and (c) been shown to show representational alignment to cognitive measurements, such as the brain activation recordings of human subjects processing linguistic stimuli (Caucheteux and King, 2020; Schrimpf et al., 2020c). Although the way these models learn is clearly unlike the way in which humans do — through text-only data, optimizing for a small number of fixed objectives, and requiring many more observations — the solutions to which they converge do, overall, provide the most accurate predictions (among existing models) for various aspects of language. This naturally leads to the question of whether and where there is potential for their employment as plausible models of the linguistic phenomena studied in theoretical, psycho-, and neuro- linguistics. For this to happen, an in-depth understanding of ANN LM's linguistic capacities, and where or how they might correspond to or diverge from humans, must first be established.

1.1 SCOPE

Accordingly, the work carried out in this dissertation has two broad aims: (a) furthering our understanding of of ANN LMs through comparisons to what we already know about how humans process and understand language and (b) developing an understanding of where LMs might help provide insights into the human language faculty.

To this end, this dissertation presents an investigation along three related lines of questioning:

- How well do ANN LMs align to behavioural data (e.g. eye-tracking data, language understanding tasks, etc.)?
- How well do ANN LMs agree with insights from linguistic theory about language processing and understanding in humans?
- Can LMs, through comparisons to measurements of neural response, be used for the study of meaning composition in the human brain?

3 There are, of course, exceptions to this, such as models which incorporate tree-like, linguistic structure as part of models' inductive biases (Dyer et al., 2016; Tai et al., 2015).

1.2 KEY QUESTIONS

For each of the threads of inquiry listed above, this thesis examines several specific questions, which are enumerated below:

- Behavioural Data:
  - Are there structural correspondences between eye-tracking fixation patterns and LM representations?
  - Can language models encode perceptual structure without grounding?
- Linguistic Theory:
  - How sensitive are humans and LMs to linguistic perturbations of Winograd Schema Challenge[4] examples?
  - Can the attention patterns learned by LMs reflect linguistic structure (in the form of dependency trees)?
  - Why do LMs trained on sentences with shuffled words still perform well on Natural Language Understanding (NLU) tasks? Do they still encode some word order information? Are there NLU tasks where degrading word order information has a stronger effect on performance?
- Neural Response Measurements:
  - Can LMs be used to enable the evaluation of targeted hypotheses about the composition of meaning in the brain?
  - Does injecting linguistic structure into LM representations lead to better alignment with brain activation measurements?

1.3 OVERVIEW AND CONTRIBUTIONS

Addressing the questions listed above, the contribution of this dissertation is to analyse numerous specific parallels and divergences between LM and human language comprehension, aiming to establish where they might be utilised for linguistic theorizing or a plausible cognitive models of language. Figure 1.1 shows an illustrative overview of the work to be presented.

This thesis is divided into four parts. Part i is made up of the current introduction followed by a background chapter, which contextualizes the work which is to be presented in the rest of the thesis. Part ii is made up of Chapters 3 & 4, which present studies that make of use behavioural data to study ANN LMs. Part iii includes Chapters 6

4 A widely-employed test of commonsense reasoning ability, proposed by Levesque et al. (2012)

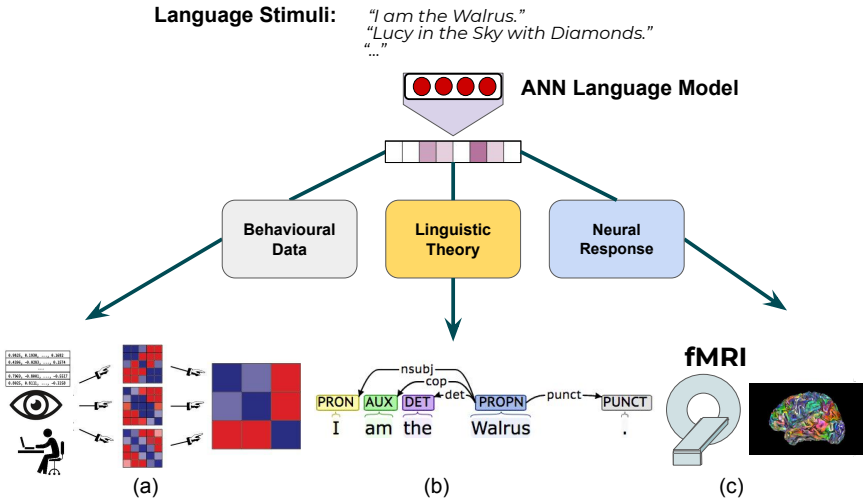


Figure 1.1: Research carried out in this dissertation worked towards linking ANN LM representations to (a) behavioural data such as eye-tracking data or similarity judgements, (b) linguistic theory, e.g. dependency tree structure, and (c) neural response measurements, e.g. fMRI.

& 7, which detail analyses of LMs that are based on various aspects of linguistic theory. In Part iv, Chapter 8 presents a survey of work linking computational models of language with neural response measurements, and Chapter 9 introduces a framework for the use of LMs to study linguistic composition in the human brain. Finally, in Part v, a discussion and conclusion are presented.

Note that the division of studies along the lines of behavioural data, linguistic theory, and neural response measurements is done primarily for organizational purposes, as there is, in any case, significant inherent interdependence and overlap between research done on the three topics.

A brief description of the work to be presented in each of the chapters is shown below:

- The work presented in Chapter 3 links gaze features (from reading) with LM representations, revealing a previously unknown correspondence between LM representation divergence and human processing difficulty.

- In Chapter 4, we show that even though they are trained on textual data only, LM representations can (partially) reflect the topology of humans' perceptual color space. In analysis, we find that this topological alignment is, in part, mediated by collocationality and differences in syntactic usage. We also establish a connection to findings from recent work on efficient communication in color naming.

- Chapter 5 analyses the robustness of both LMs and humans to linguistic perturbations of Winograd Schema Challenge examples. The findings presented show that the latter are more robust than the former, but that through finetuning on task-specific data LMs become more robust.

- Chapter 6 presents experiments on decoding dependency trees from the attention heads of a multilingual LM. The results presented span 18 languages and show that: (a) full trees can be decoded above baseline accuracy from single attention heads, and that that particular relations are often tracked by the same heads across languages and (b) steering the representations towards explicit linguistic structure by finetuning with a supervised parsing objective leads to the same structure being reflected in the resulting attention patterns.

- Chapter 7 presents an examination of recent results showing that LMs pretrained and/or finetuned on sentences with shuffled words still exhibit competitive performance on NLU benchmarks. In this work, we demonstrate that: (a) shuffled models retain information pertaining to the original natural word order due to statistical cues present in the data they are trained on and (b) there are more rigorous NLU tasks where degradation of word order information has a stronger effect on performance.

- Chapter 8 presents a survey of research that has worked to link computational models of language with neural response measurements. The survey traces a line from early research combining Event Related Potentials with complexity measures derived from simple LMs, to contemporary studies employing ANN LMs in combination with neural response recordings from multiple modalities and using naturalistic stimuli.

- Finally, Chapter 9 proposes a framework where LMs are employed for the evaluation of targeted hypotheses about the composition of meaning in the brain. Using this, we show that, across two fMRI datasets, LM representations align better with brain recordings, when their attention is biased to match linguistic annotations from three syntaco-semantic formalisms.

2

BACKGROUND

This chapter presents a brief background that is designed to contextualise the work to be presented in this dissertation. As outlined in the previous chapter, this dissertation studies ANN language models, comparing their linguistic processing and comprehension capacities to those of humans, through the lens of (i) behavioural data, (ii) linguistic theory, and, (iii) neural response measurements.

This chapter is divided into in four sections. The first offers a brief history and description of ANN LMs. The following three sections each presents a survey of work that connects ANN LMs to one of the three types of resources listed above.

2.1 ARTIFICIAL NEURAL NETWORK LANGUAGE MODELS

2.1.1 *A brief history*

The Modern ANN LM which has become so predominant in NLP can be seen as the heir to the legacy of multiple research traditions.

From statistical language modelling, it inherits the probabilistic and information theoretic frameworks which formally define what we have come to think of as a language model — a probability distribution over sequences of words $P(w_1, \ldots, w_n) = P(w_1) \times P(w_2|w_1) \times \cdots \times P(w_n|w_1 \ldots w_n - 1)$, where probabilities are computed using smoothed counts from a corpus of text — and how we might evaluate the fitness of one (e.g. perplexity) (Brown et al., 1990). Statistical language models from this tradition are often called n-gram models, due to their incorporation of a limited number n of previous units of context (e.g. words). Models of this kind have been employed to good effect in a variety of NLP applications since the late 1980s (Barzilay and Lee, 2004; Brown et al., 1990; Church, 1989; Jelinek et al., 1991; Kemighan et al., 1990; Roark, 2001). Among the main distinctions between this class of models and ANN LMs is the treatment of words or other linguistic units as discrete rather than continuous entities and because of this, the practical impossibility of including previous contexts longer than five or six sequence units due to data sparsity.

From the distributional semantics and the information retrieval literature, ANN LMs inherit the idea of representing linguistic units via their distributional properties. Even before the advent of ANN LMs, modern NLP was heavily reliant on this approach where a word is typically represented by a high-dimensional vector capturing its co-occurrence statistics in a corpus, through a variety of context-counting

and matrix factorisation models (Bullinaria and Levy, 2007; Schütze, 1993; Turney and Pantel, 2010).

Finally, from the connectionist paradigm, current day LMs inherit the basic principles of and the motivation for the ANN architecture. This paradigm takes loose inspiration from the manner in which information is processed by the biological brain. Its core idea is that information should be processed via the propagation of activation among simple units linked to each other through weighted connections representing synapses or groups thereof (McClelland, 1988). Connectionist models date back at least to work such as McCulloch and Pitts (1943) and Rosenblatt (1958), which showed that simple models like the Perceptron could implement (some) logical operations. Rumelhart and McClelland (1985) presented one of the first applications of connectionist models to language. Training simple ANN models to learn the morphological inflection patterns of past-tense English verbs, they found them to reflect (to some extent) the patterns of children learning. In the early 2000s, seminal work like Schwenk and Gauvain (2002) and Bengio et al. (2003) presented the first *continuous space* language models, which can be seen as the direct precursors to today's language models. By representing words using distributed feature vectors and training neural network models to maximize training corpus log-likelihood, they could fight the *curse of dimensionality* problem caused by data sparsity in statistical language models, enabling generalization to unseen sequences of words.

2.1.2 *Where things stand*

It took a decade or so of software and hardware advances till these models became widely adopted and deployed by the NLP community where they have recently been described as "foundation models" which are "critically central yet of incomplete character" (Bommasani et al., 2021). Most current-day LMs are based either on gated reccurent neural network architectures like the *Long Short Term Memory* (Hochreiter and Schmidhuber, 1997) or on Transformer (Vaswani et al., 2017) architectures. The latter, in particular, has become the architecture of choice for the pretraining and transfer learning paradigm.

This paradigm involves the self-supervised training of a large (billions of parameters) ANN LM on huge amounts of data (billions of tokens). These models are trained to predict future tokens (Brown et al., 2020; Radford et al., 2018, 2019), randomly masked tokens (Devlin et al., 2019a; Liu et al., 2019b), or via a variety of other denoising objectives (Lewis et al., 2020; Raffel et al., 2019). After this pretraining stage, the information acquired by these LMs can then be "transfered" to virtually any NLP task either by finetunig them with a supervised learning objective on small or medium-sized task-specific datasets or

in a zero-shot fashion. This dissertation focuses on this class of models.

2.2 LINKING ARTIFICIAL AND HUMAN COMPREHENSION OF LANGUAGE

Besides measures like perplexity which evaluate the fitness of an LM on the language modelling task itself, a significant amount of research has gone into understanding what types of information might be encoded in LMs' internal representations and what tasks they might be useful for. Below, we survey approaches of doing this via comparisons to human behaviour, linguistic theory, and neural response. Since this dissertation is not only concerned with the evaluation of LMs but also with the enabling of cross-pollination between NLP and (psycho, cognitive, and neuro-) linguistics, we also highlight work that has gone in the other direction, leveraging LMs in the study of the human language faculty.

2.2.1 *Behavioural Data*

One major approach towards evaluating LMs has relied on the construction of benchmarks that measure their capacity to emulate human behaviour on a broad range of NLU tasks. Examples of this include tasks such as: judging the grammatical acceptability of a sentence (Warstadt et al., 2019); Reading Comprehension, where the task is to read a passage and answer questions based on it (Lai et al., 2017; Rajpurkar et al., 2018, 2016); the WSC (see Chapter 5), which involves the identification of an ambiguous pronoun's correct referent (Levesque et al., 2012); and Natural Language Inference, where given a *premise* statement and a *hypothesis* statement, an answerer must classify the relationship between them as one of entailment, contradiction, or neutrality (Bowman et al., 2015; Dagan et al., 2005). Beyond linguistic capacity, these tasks are often considered to require other competencies, such as "reasoning ability" or "world knowledge".

In an effort to standardize evaluation, the community developed suites which consolidate together numerous NLU tasks and present the results in the form of overall rankings or leaderboards, such as GLUE, SuperGLUE, and decaNLP (McCann et al., 2018a; Wang et al., 2019b, 2018b). Performance on these suites, it was thought, would guide the development of models, providing a reliable indication of where progress was being made. And indeed, through various schemes involving the pretraining and finetuning of LMs, rapid progress was made on the benchmarks, with the best performing LMs reaching near-human scores on most tasks.

Although the ability of LMs to closely match human answers for these tasks is remarkable, caution is due when interpreting what this

might mean with regards to their linguistic (or reasoning) capacities. Indeed, a significant body of work has aimed to understand the extent to which LMs are solving these tasks by using what we expect them to as opposed to leveraging statistical artifacts and biases. This thread of work has to date revealed significant problems with reading comprehension datasets (Chen et al., 2016a; Kaushik and Lipton, 2018), Natural Language Inference datasets (Gururangan et al., 2018a; McCoy et al., 2019; Poliak et al., 2018a; Tsuchiya, 2018), and tasks like the WSC (Trichelair et al., 2018), among others. In connection to this, Chapter 5, presents a new diagnostic dataset, testing how sensitive (compared to humans) LMs are to minimal linguistic perturbations of WSC examples.

More recently, researchers found that the performance of LMs on most tasks which make up benchmarks like GLUE, is not significantly impacted by applying permutations of word order (and other various perturbations such as sorting, duplicating, and dropping tokens) to pretraining, finetuning, or test-time data (Gupta, 2003; Pham et al., 2020; Sinha et al., 2021, 2020), casting doubt on the extent to which deeper linguistic knowledge is actually required for these tasks. This thread of research is addressed in Chapter 7 of this dissertation.

Other types of behavioural data that have been used to evaluate LMs include lexical similarity or relatedness judgements (Chronis and Erk, 2020) and eye-tracking measurements, as is presented in Hollenstein et al. (2021) and Wilcox et al. (2020) and in in Chapter 3 of this thesis. In relation to the latter, several studies have also used eye-tracking fixation patterns as a signal to improve LM performance on NLP tasks (Barrett et al., 2018; Gonzalez-Garduno and Søgaard, 2017; Hollenstein et al., 2019; Hollenstein and Zhang, 2019; Malmaud et al., 2020).

Examples of work employing ANN LMs towards linguistic or psycholinguistic research that is based on behavioural data are less common, but do exist. Indeed, recent investigations of various aspects of syntactic processing in humans based on eye-tracking measurements have used word surprisal scores derived from LMs (Van Schijndel and Linzen, 2018, 2021). Boyce et al. (2020), employs LM predictions to automate the generation of probable and improbable sentence continuations in the Maze task (Forster et al., 2009), a standard task used in sentence processing research where a subject reads a sentence word by word, and at each word position is presented with a forced choice between a word that correctly continues the sentence and a distractor. The subject's reaction time is recorded as they choose the continuation.

2.2.2 *Linguistic Theory*

ANN LMs are opaque, complex, monolithic models. In order to interpret and analyse the information encoded by their representations and the human-like generalizations they make, researchers have relied on aspects of linguistic theory, in what has become known as the probing paradigm. Broadly speaking, this involves using parametrized or parameter-free probing tasks to answer questions about the types of (linguistic) information encoded in a model's internal representation and which component of a model's architecture or what part of its training objective and inductive biases lead to it acquiring certain properties or making certain decisions (Belinkov and Glass, 2019a; Kulmizev and Nivre, 2021).

Parametrized probing relies on a fairly straightforward procedure: a LM is used to induce representations for a set of examples annotated for a given class of linguistic information (e.g. part-of-speech), then a classifier is trained to map from the representations to the labels. If the classifier does well at predicting the labels of an unseen test set, the LM-dervied representations are said to encode for this linguistic property. Using this framework, research has shown that LMs encode for morphosyntactic information (Belinkov et al., 2017a; Shi et al., 2016), hierarchical linguistic structure (Conneau et al., 2018; Hewitt and Manning, 2019; Hupkes et al., 2017), and common sense or factual knowledge (Feldman et al., 2019; Petroni et al., 2019), with some research even suggesting that pretrained LMs "rediscover the classic NLP pipeline", with lower layers encoding for local syntax and higher layers capturing more complex semantic information (Tenney et al., 2019). Soon however, it became recognized that parametrized-probe performance could be affected by factors which are not directly related to how well a class of information is encoded for or is readable from a given representation, such as probe expressivity (Hewitt and Liang, 2019b).

To address this, several approaches have been proposed, such as: comparing probe performance to a control task, encouraging *selective* probes that simultaneously achieve high linguistic task and low control task accuracies (Hewitt and Liang, 2019b; Ravichander et al., 2020); controlling for or quantifying probe expressivity (Pimentel et al., 2020a; Voita and Titov, 2020), and various other information-theoretic operationalizations of parametrized probing (Hewitt et al., 2021; Hou and Sachan, 2021; Pimentel et al., 2020c). Many of the methods described above are utilised in this dissertation (Chapters 4, 5, 6, 7 & 9). The reader is referred to Belinkov (2021) for a detailed treatment of the topic.

Parameter-free probing, on the other hand, circumvents many of the issues described above, relying instead on carefully curated sets of stimuli that are based on phenomena described in the syntax lit-

erature. A prominent example of this is the Targeted Syntactic Evaluation framework or variants of thereof (Chowdhury and Zamparelli, 2019; Futrell and Levy, 2019; Linzen et al., 2016) where minimal pairs of examples — one grammatical and one not — are designed to test if the probability distribution defined by an LM conforms to the grammar of the language. An example of this, taken from Marvin and Linzen (2018) is:

1. The bankers knew the officer smiles.
2. *The bankers knew the officer smile.

Here, a model that is able to correctly analyse the syntactic structure of the sentence would be expected to assign a higher probability to (1) than to (2), based on the need for the verb *smiles* to agree with the embedded subject *officer* rather than the subject of the main clause (*bankers*).

Recently, Warstadt et al. (2020) and Hu et al. (2020a) presented evaluation suites which follow this framework. Grouping over different classes of syntactic phenomena that occur in the English language (Agreement, Licensing, Ellipsis, etc.), these suites enable a systematic evaluation of the syntactic knowledge in LMs. We refer the reader to Linzen and Baroni (2021) for a detailed survey of this line of research and to Kulmizev and Nivre (2021) for a discussion of its pitfalls.

Targeted syntactic evaluation is applied to measure the syntactic capacities of structurally-biased models in Chapter 9 of this dissertation. Other approaches for parameter-free probing include methods that rely on similarity structure like Representational Similarity Analysis, which this thesis discusses in detail in Chapter 3.

While the influence of linguistic theory on ANN LM research is undeniable, the opposite has so far not been true. Baroni (2021) argues that ANN LMs should be treated as linguistic theories, and that the lack of interest in these models within theoretical linguistics research can be attributed to two main obstacles: (a) the lack of methodical consistency and theoretically grounding with regards to how architecture, hyperparameters, training data, etc. are chosen, which makes it difficult to "take this work seriously from the perspective of linguistic theorizing" and (b) the lack of interesting predictions about previously unexplored linguistic patterns, since the the field (i.e. NLP) has so far been conducting an "extensive sanity check", testing how well LMs can capture already recognized patterns and phenomena.

2.2.3 *Neural Response Measurements*

Finally, the reader is referred to Chapter 8 for a comprehensive survey of work linking ANN LMs with Neural Response measurements. Here, we find a reversal of the trend from the previous two categories,

where the use of LMs in neurolinguistic research has been more common than the converse.

Part II

BEHAVIOURAL DATA

3

HIGHER-ORDER COMPARISONS OF SENTENCE ENCODER REPRESENTATIONS

3.1 ABSTRACT

Representational Similarity Analysis is a technique developed by neuroscientists for comparing activity patterns of different measurement modalities (e.g., fMRI, electrophysiology, behavior). As a framework, RSA has several advantages over existing approaches to interpretation of language encoders based on probing or diagnostic classification: namely, it does not require large training samples, is not prone to overfitting, and it enables a more transparent comparison between the representational geometries of different models and modalities. We demonstrate the utility of RSA by establishing a previously unknown correspondence between widely-employed pretrained language encoders and human processing difficulty via eye-tracking data, showcasing its potential in the interpretability toolbox for neural models.

3.2 INTRODUCTION

Examining the parallels between human and machine learning is a natural way for us to better understand the former and track our progress in the latter. The "black box" aspect of neural networks has recently inspired a large body of work related to interpretability, i.e. understanding of representations that such models learn. In NLP, this push has been largely motivated by linguistic questions, such as: *what linguistic properties are captured by neural networks?* and *to what extent do decisions made by neural models reflect established linguistic theories?* Given the relative recency of such questions, much work in the domain so far has been focused on the context of models in isolation (e.g. *what does model* **X** *learn about linguistic phenomenon* **Y**?) In order to more broadly understand models' representational tendencies, however, it is vital that such questions be formed not only with other models in mind, but also other representational methods and modalities (e.g. behavioral data, fMRI measurements, etc.). In context of the latter concern, the present-day interpretability toolkit has not yet been able to afford a practical way of reconciling this.

In this work, we employ Representational Similarity Analysis as a simple method of interpreting neural models' representational spaces as they relate to other models and modalities. In particular, we conduct an experiment wherein we investigate the correspondence between human processing difficulty (as reflected by gaze fixation mea-

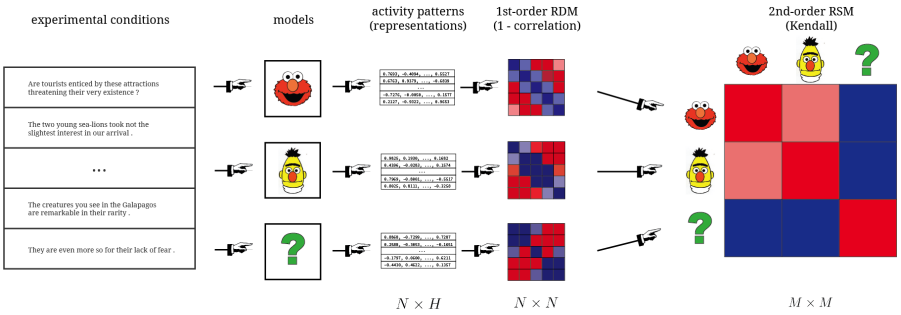


Figure 3.1: An example of first- and second-order analyses, where N = # of experimental conditions, M = # of models, and H = # of activity patterns observed for a given model (i.e. dimensionality). The right-most side of the figure depicts a representational similairty matrix (RSM) of correlations between RDMs.

surements) and the representations induced by popular pretrained language models. In our experiments, we hypothesize that there exists an overlap between the sentences which are difficult for humans to process and those for which per-layer encoder representations are least correlated.

Our intuition is that such sentences may exhibit factors such as low-frequency vocabulary, lexical ambiguity, and syntactic complexity (e.g. multiple embedded clauses), etc. that are uncommon in both standard language and, relatedly, the corpora employed in training large-scale language models. In the case of a human reader, encountering such a sentence may result in a number of processing delays, e.g. longer aggregate gaze duration. In the case of a sentence encoder, an uncommon sentence may lead to a degradation of representations in the encoder's layers, wherein a lower layer might learn to encode vastly different information than a higher one. Similarly, different models' representations may emphasize different aspects of these more complex sentences and therefore diverge from each other. With this in mind, our hypothesis is that sentences which are difficult for humans to process are likely to have divergent representations within models' internal layers and between different models' layers.

UNDERSTANDING AND ANALYSING LANGUAGE ENCODERS In recent years, some prominent efforts towards interpreting neural networks for NLP have included: developing suites that evaluate network representations through performance on downstream tasks (Conneau et al., 2017a; McCann et al., 2018b; Wang et al., 2018a); analyzing network predictions on carefully curated datasets (Dasgupta et al., 2018; Gulordava et al., 2018; Linzen et al., 2016; Loula et al., 2018; Marvin and Linzen, 2018; Tenney et al., 2018); and employing diagnostic classifiers to assess whether certain classes of information are encoded in a model's (intermediate) representations (Adi et al., 2016; Belinkov et al., 2017a; Chrupała et al., 2017; Hupkes et al., 2017).

While these approaches provide valuable insights into how neural networks process a large variety of phenomena, they rely on decoding accuracy as a probe for encoded linguistic information. If properly biased, this means that they can detect whether information is encoded in a representation or not. However, they do not allow for a direct comparison of representational structure between models. Consider a toy dataset of five sentences of interest and three encodings derived from quite different processing models; a hidden state of a trained neural language model, a *tf-idf* weighted bag-of-words representation, and measurements of fixation duration from an eye-tracking device. Probing methods do not allow us to quantify or visualise, for each of these encoding strategies, how the encoder's responses to the five sentences relate to each other. Moreover, probing methods would not directly reveal whether the fixations from the eye-tracking device aligned more closely with the tf-idf representation or the states of the neural language model. In short, while probing classifier methods can establish if phenomena are separable based on the provided representations, they do not tell us about the overall geometry of the representational spaces. RSA, on the other hand, provides a basis for higher-order comparisons between spaces of representations, and a way to visualise and quantify the extent to which they are isomorphic.

Indeed, RSA has seen a modest introduction within interpretable NLP in recent years. For example, Chrupała et al. (2017) employed RSA as a means of correlating encoder representations of speech, text, and images in a post-hoc analysis of a multi-task neural pipeline. Similarly, Bouchacourt and Baroni (2018) used the framework to measure the similarity between input image embeddings and the representations of the same image by an agent in an language game setting. More recently, Chrupała and Alishahi (2019) correlated activation patterns of sentence encoders with symbolic representations, such as syntax trees. Lastly, similar to our work here, Abnar et al. (2019a) proposed an extension to RSA that enables the comparison of a single model in the face of isolated, changing parameters, and employed this metric along with RSA to correlate NLP models' and human brains' respective representations of language. We hope to position our work among this brief survey and further demonstrate the flexibility of RSA across several levels of abstraction.

3.3 REPRESENTATIONAL SIMILARITY ANALYSIS

RSA was proposed by Kriegeskorte et al. (2008) as a method of relating the different representational modalities employed in neuroscientific studies. Due to the lack of correspondence between the activity patterns of disparate measurement modalities (e.g. brain activity via fMRI, behavioural responses), RSA aims to abstract away from the ac-

tivity patterns themselves and instead compute representational dissimilarity matrices (RDMs), which characterize the information carried by a given representation method through dissimilarity structure.

Given a set of representational methods (e.g., pretrained encoders) M and a set of experimental conditions (sentences) N, we can construct RDMs for each method in M. Each cell in an RDM corresponds to the dissimilarity between the activity patterns associated with pairs of experimental conditions $n_i, n_j \in N$, say, a pair of sentences. When $n_i = n_j$, the dissimilarity between an experimental condition and itself is intuitively 0, thus making the $N \times N$ RDM symmetric along a diagonal of zeros (Kriegeskorte et al., 2008).

The RDMs of the different representational methods in M can then be directly compared in a Representational *Similarity* Matrix (RSM). This comparison of RDMs is known as second-order analysis, which is broadly based on the idea of a *second-order isomorphism* (Shepard and Chipman, 1970). In such an analysis, the principal point of comparison is the match between the dissimilarity structure of the different representational methods. Intuitively, this can be expressed through the notion of *distance between distances*, and is thus related to Earth Mover's Distance (Rubner et al., 2000).[1] Figure 3.1 shows an illustration of the first and second order analyses for pretrained language encoders.

Note that RSA is meaningfully different from, and complementary to, methods that employ saturating functions of representation distances (e.g. decoding accuracy, mutual information), which suffer from **(a)** a ceiling effect: being able to distinguish experimental phenomenon A from B with with an accuracy of 100% and experimental phenomenon C from D with an accuracy of 100% does not mean that the distance between A and B is the same as that between C and D; and **(b)** discretization (Nili et al., 2014).

We follow Kriegeskorte et al. (2008) in using the correlation distance of experimental condition pairs $n_i, n_j \in N$ as a dissimilarity measure, where $\bar{n_i}$ is the mean of n_i's elements, $\cdot$ is the dot product, and $\|$ is the l_2 norm: $\text{corr}(x) = 1 - \frac{(n_i - \bar{n_i}) \cdot (n_j - \bar{n_j})}{\|(n_i - \bar{n_i}\|_2 \|(n_j - \bar{n_j}\|_2}$. Compared to other measures, correlation distance is preferable as it normalizes both the mean and variance of activity patterns over experimental conditions. Other popular measures include the Euclidean distance and the Malahanobis distance (Kriegeskorte et al., 2006).

3.4 FIXATION DURATION AND ENCODER DISAGREEMENT

Gaze fixation patterns have been shown to strongly reflect the online cognitive processing demands of human readers (Ashby et al., 2005;

1 More precisely, our measure of dissimilarity between experimental conditions is analogous to *ground distance* and dissimilarity between RDMs to *earth mover's distance.*

Raney et al., 2014) and to be dependent upon a number of linguistic factors (Van Gompel, 2007). Specifically, it has been demonstrated that word frequency, syntactic complexity, and lexical ambiguity play a strong part in determining which sentences are difficult for humans to process (Duffy et al., 1988; Levy, 2008; Rayner and Duffy, 1986).

Using the RSA framework, we aim to explore how gaze fixation patterns and the linguistic factors associated with sentence processing difficulty relate to the representational spaces of popular language encoders. Namely, we hypothesize that, for a given sentence, disagreement between hidden layers corresponds to processing difficulty. Because layer disagreement for a sentence measures the extent to which two layers (e.g. within BERT) disagree with each other about the pairwise similarity of the sentence (with other sentences in the corpus), a sentence with high layer disagreement will have unstable similarity relationships to other sentences in the corpus. This indicates that it has a degraded encoder representation. Going further, we also hypothesize that models' representations of said sentences may be confounded, in part, by factors that are known to influence humans.

EYE-TRACKING DATA For our experiments, we make use of the Dundee eye-tracking corpus (Kennedy et al., 2003), the English part of which consists of eye-movement data recorded as 10 native participants read 2,368 sentences from 20 newspaper articles. We consider the following fixation features: TOTAL FIXATION DURATION and FIRST PASS DURATION. For each of the features, we first take the average of the measurements recorded for all 10 participants per word, then obtain sentence-level annotations by summing the measurements of all words in a sentence and dividing by its length. The result of this is two vectors $V_{\mathtt{totfix}}$ and $V_{\mathtt{firstpass}}$ of length $2,368$, where each cell in the vector corresponds to a sentence's average total fixation and average first pass duration, respectively.

SYNTACTIC COMPLEXITY, WORD FREQUENCY, AND LEXICAL AMBIGUITY We also consider the three following linguistic features which affect processing difficulty. For each of the following the result is also a vector of length $2,368$ where each cell corresponds to a sentence:

a. the average word log frequency per sentence extracted from the British National Corpus (Leech, 1992), $V_{\mathtt{logFreq.}}$.

b. the average number of senses per word per sentence extracted from WordNet (Miller, 1995), $V_{\mathtt{wordSense}}$.

c. Yngve scores, a standard measure of syntactic complexity based on cognitive load (Yngve, 1960) , $V_{\mathtt{Yngve}}$.

PRETRAINED ENCODERS We conduct our analysis on pretrained BERT-large (Devlin et al., 2019b) and ELMo (Peters et al., 2018b), two widely employed contextual sentence encoders. To obtain a representation of a sentence from a given layer L, we perform mean-pooling over the time-steps which correspond to the words of a sentence, obtaining a vector representation of the sentence. Mean-pooling is a common approach for obtaining vector representations of sentences for downstream tasks (Conneau et al., 2017b; Peters et al., 2018b). We refer to ELMo's lowest layer as E1, BERT's 11th layer as B11, etc.

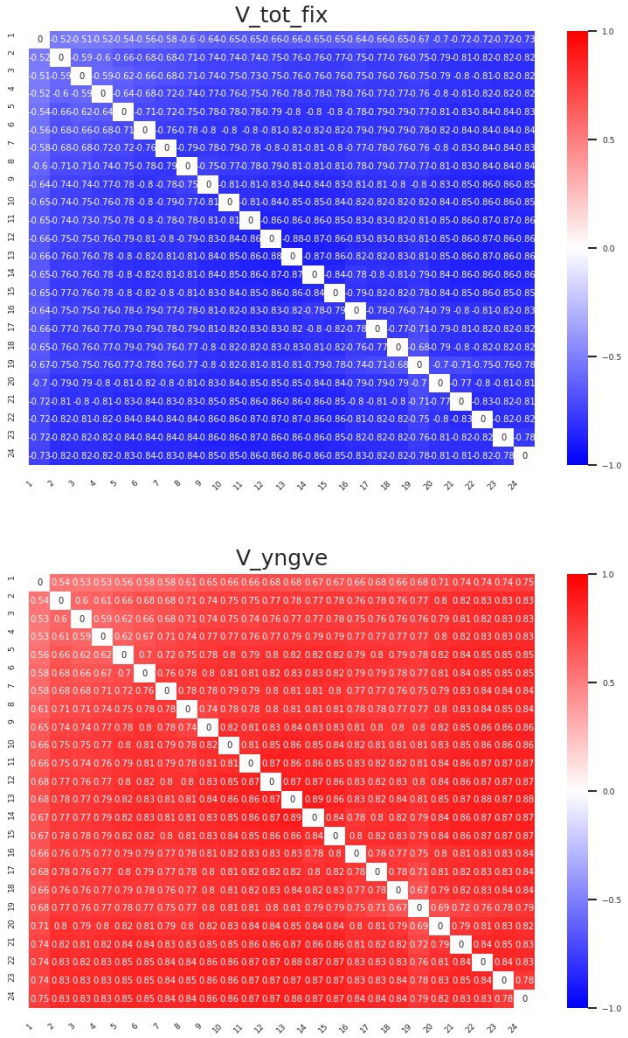


Figure 3.2: RSMs showing (Spearman's ρ) correlation between disagreement among layers i and j ($V_{\mathrm{Corr}_{L_i - L_j}}$) and V_{totfix} (left) and V_{Yngve} (Right). BERT layers are denoted with numbers from 1 (topmost) to 24 (lowest).

RDMS We construct an RDM (see §3.3) for each contextual encoder's layers. Each RDM is a $2,368 \times 2,368$ matrix which represents the dis-

similarity structure of the layer, (i.e., each row vector in the matrix contains the dissimilarity of a given sentence to every other sentence). We then compute the correlations between the two different RDMs. For our evaluation of how well the representational geometry of a layer correlates to another, we employ Kendall's τ_A as suggested in Nili et al. (2014), computing the pairwise correlation for each two corresponding rows in two RDMs. This second-order analysis gives us a pairwise relational similarity vector $V_{Corr_{L_i-L_j}}$ of length 2,368, which has the correlations between two layers L_i and L_j's RDMs for each of the sentences.

THIRD-ORDER ANALYSIS The final part of our analysis involves computing correlations of $\{V_{Corr_{L_i-L_j}}, V_{logFreq}, V_{Yngve}, V_{wordSense}\}$ with each of V_{totfix} and $V_{firstpass}$. The results from this are shown in Table 3.1. The top section of the table shows correlations when L_i and L_j are the three final adjacent layers in BERT and ELMo. The middle section shows the results for top three BERT layer pairs L_i and L_j which maximize the correlation scores. The final section shows correlation with the linguistic features. Finally, Figure 3.2 shows Spearman's ρ correlations between $V_{Corr_{L_i-L_j}}$ and each of V_{totfix}, and V_{Yngve} for all combinations of the 24 BERT layers.

3.5 DISCUSSION

Our results show highly significant negative correlations between $V_{Corr_{L_i-L_j}}$ and sentence gaze fixation times. These findings confirm the hypothesis that the sentences that are most challenging for humans to process, are the sentences **(a)** the layers of BERT disagree most on among themselves; and **(b)** that ELMo and BERT disagree most on, indicating that there may be common factors which affect human processing difficulty and result in disagreement between layers. By Layer disagreement we refer to the expression $1 - V_{Corr_{L_i-L_j}}$. It is important to note that these encoders are trained with a language modelling objective, unlike models where reading behaviour is explicitly modelled (Hahn and Keller, 2016) or predicted (Matthies and Søgaard, 2013). Indeed, the similarities here emerge naturally as a function of the task being performed. This can be seen as analogous to the case of similarities observed between neural networks trained to perform object recognition and spatio-temporal cortical dynamics (Cichy et al., 2016).

SYNTACTIC COMPLEXITY Figure 3.2 shows that, for all combinations of BERT layers, total fixation time and Yngve scores have strong negative and positive correlations (respectively) with layer disagreement. Furthermore, we observe that disagreement between middle layers seems to show the strongest correlation with Yngve scores.

Layer Disagreement	Total Fixation	First Pass Duration
E1-B22	-0.46	-0.46
E2-B23	-0.66	-0.67
E3-B24	-0.22	-0.23
B11-B12	-0.88	-0.87
B12-B13	-0.87	-0.85
B10-B21	-0.87	-0.86
Linguistic Features		
Log Freq.	-0.20	-0.19
Avg. Senses per Word	-0.007*	-0.004*
Yngve Score	0.66	0.66

Table 3.1: Spearman's ρ between $V_{\mathrm{Corr}_{L_i-L_j}}$, $V_{\mathtt{logFreq.}}$, $V_{\mathtt{wordSense}}$, $V_{\mathtt{Yngve}}$ and each of $V_{\mathtt{totfix}}$ and $V_{\mathtt{firstpass}}$. All correlations significant with $p < 0.0001$ after Bonferroni correction unless marked with *.

To confirm this, we split the correlations into four groups: "low" ($i, j \in [1, 8]$), "middle" ($i, j \in [9, 16]$), "high" ($i, j \in [17, 24]$), and "out" ($|i - j| > 7$), with the latter representing out-of-group correlations (e.g. $\mathrm{Corr}_{L_1-L_{24}}$). To account for correlations between disagreeing adjacent layers (e.g. $|i - j| = 1$) and Yngve scores being higher (as a possible confounding factor), we also distinguish layers as either "adjacent" or "non-adjacent". Considering these two factors as three- and two-leveled independent variables respectively, we conduct a two-way analysis of variance. The analysis reveals that the effect of group is significant at $F(3, 275) = 78.47, p < 0.0001$, with "low" ($\mu$ = 0.65, σ = 0.08), "middle" (μ = 0.84, σ = 0.03), "high" (μ = 0.80, σ = 0.05), and "out" (μ = 0.80, σ = 0.05). Neither the effect of adjacency nor its interaction with group proved to be significant.

This can be seen as (modest) support for the findings of previous work (Blevins et al., 2018; Tenney et al., 2019): namely, that the intermediate layers of neural language models encode the most syntax, and are therefore possibly more sensitive towards syntactic complexity. A very similar pattern is observed for total fixation time. When considered together with the correlation between $V_{\mathtt{Yngve}}$ and fixation times, this indicates a tripartite affinity between layer disagreement, syntactic complexity, and fixation.

LEXICAL AMBIGUITY AND WORD FREQUENCY Finally, we observe that $V_{\mathtt{logFreq.}}$ has a moderate correlation with both fixation time and layer disagreement and that $V_{\mathtt{wordSense}}$ is nearly uncorrelated to both. Detailed plots of the latter can be found in Appendix A.1.1.

3.6 conclusion

We presented a framework for analyzing neural network representations that allowed us to relate human sentence processing data with language encoder representations. In experiments conducted on two widely used encoders, our findings show that sentences which are difficult for humans to process have more divergent representations both intra-encoder and between different encoders. Furthermore, we lend modest support to the intuition that a model's middle layers encode comparatively more syntax. Our framework offers insight that is complimentary to decoding or probing approaches, and is particularly useful to compare representations from across modalities.

4

CAN LANGUAGE MODELS ENCODE PERCEPTUAL STRUCTURE WITHOUT GROUNDING? A CASE STUDY IN COLOR

4.1 ABSTRACT

Pretrained language models have been shown to encode relational information, such as the relations between entities or concepts in knowledge-bases — (Paris, Capital, France). However, simple relations of this type can often be recovered heuristically and the extent to which models implicitly reflect topological structure that is grounded in world, such as perceptual structure, is unknown. To explore this question, we conduct a thorough case study on color. Namely, we employ a dataset of monolexemic color terms and color chips represented in CIELAB, a color space with a perceptually meaningful distance metric. Using two methods of evaluating the structural alignment of colors in this space with text-derived color term representations, we find significant correspondence. Analyzing the differences in alignment across the color spectrum, we find that warmer colors are, on average, better aligned to the perceptual color space than cooler ones, suggesting an intriguing connection to findings from recent work on efficient communication in color naming. Further analysis suggests that differences in alignment are, in part, mediated by collocationality and differences in syntactic usage, posing questions as to the relationship between color perception and usage and context.

4.2 INTRODUCTION

Without grounding or interaction with the world, language models learn representations that encode various aspects of formal linguistic structure (e.g., morphosyntax (Tenney et al., 2019)) and semantic information (e.g., lexical similarity (Reif et al., 2019a)). Beyond this, it has been suggested that text-only training data is enough for LMs to also acquire factual and relational information about the world (Feldman et al., 2019; Petroni et al., 2019). This includes, for instance, some features of concrete and abstract concepts, such as objects' attributes and affordances (Forbes et al., 2019; Weir et al., 2020). Furthermore, the representational geometry of LMs has been found to naturally reflect human lexical similarity and relatedness judgements, as well as analogy relationships (Chronis and Erk, 2020). However, the extent to which these models reflect the structures that exist in humans' percep-

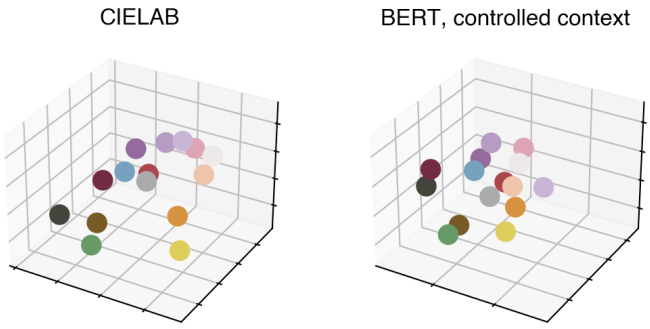


Figure 4.1: Right: Color orientation in 3d CIELAB space. Left: linear mapping from BERT (CC, see §4.3) color term embeddings to the CIELAB space.

tual world—such as the topology of visual perception (Chen, 1982), the structure of the color spectrum (Ennis and Zaidi, 2019; Provenzi, 2020), or of odour spaces (Chastrette, 1997; Rossiter, 1996)—is not well-understood. If LMs are indeed able to capture such topologies—in some domains, at least—it would mean that these structures are a) somehow reflected in language and, thereby, encoded in the textual training data on which models are trained, and b) learnable using models' current training objectives and architectural inductive biases. To the extent they are not, the question becomes whether the information is not there in the data, or whether model and training objective limitations are to blame. Certainly, this latter point relates to an ongoing debate regarding what exactly language models can be expected to learn from ungrounded form alone (Bender and Koller, 2020b; Bisk et al., 2020a; Merrill et al., 2021).

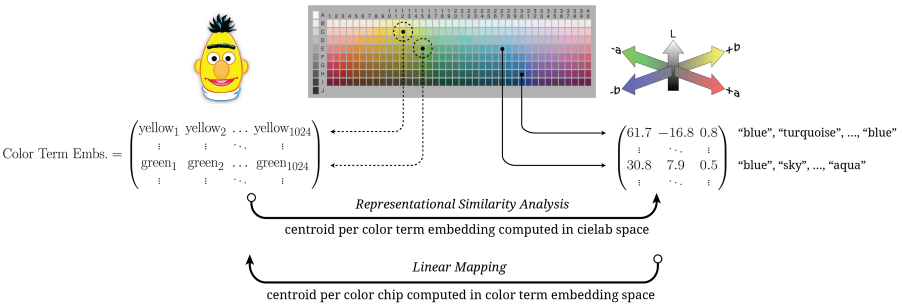


Figure 4.2: Our experimental setup. In the center is a Munsell color chart. Each chip in the chart is represented in the CIELAB space (right) and has 51 color term annotations. Color term embeddings are extracted through various methods. In the Representation Similarity Analysis experiments, a corresponding color chip centroid is computed in the CIELAB space. In the Linear Mapping experiments, a color term embedding centroid is computed per chip.

In this paper, we conduct a case study on color. Indeed, color perception in humans and its relation to speakers' use of color terms has

long been the subject of studies in cognitive science (Berlin and Kay, 1991; Kay et al., 2009; Kay and McDaniel, 1978; Regier et al., 2007). To this end, spaces have been defined in which Euclidean distances between related colors are correlated with reported perceptual differences.[1] In addition, the semantics of color terms have long been understood to hold particular linguistic significance, as they are theorised to be subject to universal constraints that arise directly from the neurophysiological mechanisms and properties underlying visual perception and cognition (Berlin and Kay, 1991; Kay et al., 1991; Kay and McDaniel, 1978).[2] Due to these factors, color offers a useful testbed for investigating whether or not structural information about the topology of the perceptual world might be encoded in linguistic representations.

To explore this in detail, we employ a dataset of English color terms and their corresponding color chips[3], the latter of which are represented in CIELAB — a perceptually uniform color space. In addition to the color chip CIELAB coordinates, we extract linguistic representations for the corresponding color terms. With these two representations in mind (see Figure 4.1 for a demonstrative plot from our experiments), we employ two methods of measuring structural correspondence, with which we evaluate the alignment between the two spaces. Figure 4.2 shows an illustration of the experimental setup. We find that the structures of various language model representations show alignment with the structure of the CIELAB space, demonstrating that some approximation of perceptual color space topology can indeed be learned from text alone. We also show that part of this distributional signal is learnable by simple models — e.g. models based on pointwise mutual information statistics — although large-scale language model pretraining (e.g., BERT) encodes the topology markedly better.

Analysis shows that larger language models align better than smaller ones and that much of the variance in CIELAB space can be explained by low-dimensional subspaces of LM-induced color term representations. To better understand the results, we also analyse the differences in alignment across the color spectrum, observing that that warm colors are generally better aligned than cool ones. Further investigation reveals a connection to findings reported in work on communication efficiency in color naming, which posits that warmer colors

1 The differences between color stimuli which are perceived by human observers.

2 These theories have been contested by work arguing for linguistic relativism (cf. the *Sapir–Whorf Hypothesis*), which emphasizes the arbitrariness of language and the relativity of semantic structures and minimizes the role of universals. Such critiques have, however, been accommodated for in the Berlin & Kay paradigm (Berlin and Kay, 1991), the basic assumptions of which, such as the existence of at least some perceptually-determined universal constraints on color naming, remain widely accepted.

3 Each chip is a unique color sample from the Munsell chart, which is made up of 330 such samples which cover the space of colors perceived by humans. See §4.3.

are communicated more efficiently. Finally, we investigate various corpus statistics which could influence alignment, finding that a measure of color term collocationality based on PMI statistics corresponds to lower alignment, while the entropy of a color term's dependency relation distribution (i.e. terms occurring as adjectival modifiers, nominal subjects, etc.) and how often it occurs as an adjectival modifier correspond to a stronger one.

4.3 METHODOLOGY

COLOR DATA We employ the Color Lexicon of American English, which provides extensive data on color naming. The lexicon consists of 51 monolexemic color name judgements for each of the 330 Munsell Chart color chips[4] (Lindsey and Brown, 2014). The color terms are solicited through a free-naming task, resulting in 122 terms.

PERCEPTUAL COLOR SPACE Following previous work (Chaabouni et al., 2021; Regier et al., 2007; Zaslavsky et al., 2018), we map colors to their corresponding points in the 3D CIELAB space, where the first dimension L expresses lightness, the second A expresses position between red and green, and the third B expresses the position between blue and yellow. Distances between colors in the space correspond to their perceptual difference.

LANGUAGE MODELS Our analysis is conducted on three widely used language models: BERT (Devlin et al., 2019a) and RoBERTa (Liu et al., 2019b), both of which employ a masked language modelling objective, and ELECTRA (Clark et al., 2020), which is trained instead with a discriminative token replacement detection objective.[5]

BASELINES In addition to the aforementioned language models, we consider two different baselines:

- PMI statistics, which are computed[6] for the color terms in common crawl, using window sizes of 1 (pmi-1), 2 (pmi-2), and 3 (pmi-3). The result is a vocabulary length vector quantifying the likelihood of co-occurrence of the color term with every other vocabulary item in within that window.
- Word-type FastText embeddings trained on Common Crawl (Bojanowski et al., 2017).

REPRESENTATION EXTRACTION We follow Bommasani et al. (2020a) and Vulić et al. (2020) in defining configurations for the extraction of

4 http://www1.icsi.berkeley.edu/wcs/images/jrus-20100531/wcs-chart-4x.png

5 bert-large-uncased; roberta-large; electra-large-discriminator

6 Using Hyperwords: https://bitbucket.org/omerlevy/hyperwords

word-type representations from LM hidden states. In the first configuration (NC), a color term is encoded without context, with the appropriate delimiter tokens attached (e.g. `[CLS] red [SEP]` for BERT). In the second, S sentential contexts that include the color term are encoded and the hidden states representing these contexts are mean pooled. These S contexts are either randomly sampled from common crawl (RC), or deterministically generated to allow for control over contextual variation (CC). If a color term is split by an LM's tokenizer into more than one token, subword token encodings are averaged over. For each color term and configuration, an embedding vector of hidden state dimension d_{LM} is extracted per layer, per model.

CONTROLLED CONTEXT To control for the effect of variation in the sentence contexts used to construct color term representations, we employ a templative approach to generate a set of identical contexts for all color terms. When generating controlled contexts, we create three frames in which the terms can appear:

- COPULA: the `<obj>` is `<col>`
- POSSESSION: i have a `<col> <obj>`
- SPATIAL: the `<col> <obj>` is there

We use these frames in order to limit the contextual variation across colors (`<col>`) and to isolate their representations amidst as little semantic interference as possible, all while retaining a naturalistic quality to the input. We also aggregate over numerous object nouns (`<obj>`), which the color terms are used to describe. We select objects from the McRae et al. (2005) data which are labelled in the latter as plausibly occurring in many colors and which are stratified across 13 category sets, e.g. *fan* $\in$ `APPLIANCES`, *skirt* $\in$ `CLOTHING`, etc. Collapsing over categories, we generate sentences combinatorially across frames, objects and color terms, resulting in $3 \times 122 \times 18 = 6588$ sentences, 366 per term.

4.4 EVALUATION

We employ two complimentary evaluation methods to gauge the correspondence of the color term text-derived representations to the perceptual color space. The first, Representation Similarity Analysis, is non-parametric and uses pairwise comparisons of stimuli to provide a measure of the global topological alignment between two spaces. The second employs a learned linear mapping, evaluating the extent to which two spaces can be aligned via transformation (rotation, scaling, etc.).

RSA (Kriegeskorte et al., 2008) is a method of relating different representational modalities, which was first employed in neuroscientific

studies. RSA abstracts away from activity patterns themselves (e.g. neuron values in representational vectors) and instead computes representational (dis)-similarity matrices (RSMs), which characterize the information carried by a given representation method through global (dis)-similarity structure. Kendall's rank correlation coefficient (τ) is computed between RSMs derived from the two spaces, providing a summary statistic indicative of the overall representational alignment between them. RSA is non-parametric and therefore circumvents many of the various methodological weaknesses associated with the probing paradigm (Belinkov, 2021).

For each color term, we compute a centroid in the CIELAB space following the approach described in Lindsey and Brown (2014). Each centroid is defined as the average CIELAB coordinate of the samples (i.e. color chips) that were named with the corresponding term (across the 51 subjects). This results in N parallel points in the color term embedding and perceptual color spaces, where N is the number of color terms considered. For our analysis, we exclude color terms used less frequently than a cutoff $f = 100$ in the color lexicon, leaving us with the 18 most commonly used color terms.[7] We then separately construct an $N \times N$ RSM for each of the LM spaces and for CIELAB . Each cell in the RSM corresponds to the similarity between the activity patterns associated with pairs of experimental conditions $n_i, n_j \in N$.

For the color term embedding space, we employ Pearson's correlation coefficient (r) as a similarity measure between each pair of embeddings $n_i, n_j \in N$. For the CIELAB space, we elect to use the following method, per Regier et al. (2007) suggestion: $\mathrm{sim}(n_i, n_j) = \exp(-c \times [\mathrm{dist}(n_i, n_j)]^2)$, where c is a scaling factor (set to 0.001 in all experiments reported here) and $\mathrm{dist}(n_i, n_j)$ is the CIELAB distance (ΔE^*_{CMC})[8] between chips n_i and n_j. This similarity measure is derived from the psychological literature on categorization and is meant to model the assumption that beyond a certain distance colors appear entirely different, so that increasing the distance has no further effect on dissimilarity. Finally, we report the mean Kendall's τ between the color term embedding and color space RSMs. We also report τ per color term (i.e. per row in the RSM), which corresponds to how well-aligned each individual color term is.

LINEAR MAPPING We train regularised linear regression models to map from color term embedding space $X \in \mathbb{R}^{n \times d_{LM}}$ to CIELAB space $Y \in \mathbb{R}^{n \times 3}$, minimising $\mathcal{L}(W; \alpha) = \|XW - Y\|_2^2 + \alpha \|W\|_1$, where

7 This includes all color terms which are considered "basic" (*red*, *blue*, etc.), and commonly used "derived" terms (*pink*, *gray*, *turquoise*, *maroon*, etc.), but excludes the rest which are only infrequently used as color terms (*forest*, *puke*, *dew*, *seafoam*, etc.). See appendix A.2.1 for full list of colors included.

8 We use the `colormath` Python package, setting illuminant to C, and assuming 2 degree standard observer.

$W \in \mathbb{R}^{3 \times d_{LM}}$ is a linear map and α is the lasso regularization hyper parameter. We vary α across a wide range of settings to examine the effect of probe complexity, which we measure using the nuclear norm of the linear projection matrix $W \in \mathbb{R}^{\phi \times \iota}$; $\|W\|_* = \sum_{i=1}^{min(\phi,\iota)} \sigma_i(W)$, where $\sigma_i(W)$ is the ith singular value of W (Pimentel et al., 2020b). The fitness of the regressors, evaluated using n-fold cross-validation ($n = 6$) indicates the alignability of the two spaces, given a linear transformation. Centroids corresponding to each Munsell color chip are computed in the color term embedding space via the weighted mean of the embeddings of the 51 terms used to label it. As in the RSA experiments, terms occurring less frequently than the cutoff ($f = 100$) are excluded. For evaluation, we compute the average (across splits and datapoints) proportion of explained variance as well as the ranking of a predicted color term embedding according to the Pearson distance ($1 - r$) to gold.

	NC				RC				CC			
Model	RSA		lin. map		RSA		lin. map		RSA		lin. map	
	max	mean	max	mean	max	mean	max	mean	max	mean	max	mean
BERT	0.16^{*}	$0.01_{\pm 0.09}$	0.75	$0.73_{\pm 0.01}$	$0.26^{\dagger}$	$0.20_{\pm 0.03}$	0.74	$0.73_{\pm 0.08}$	$0.24^{\dagger}$	$0.19_{\pm 0.03}$	0.76	$0.75_{\pm 0.05}$
RoBERTa	$0.33^{\S}$	$0.02_{\pm 0.11}$	0.75	$0.73_{\pm 0.01}$	0.20^{*}	$0.14_{\pm 0.04}$	0.74	$0.73_{\pm 0.01}$	0.19^{*}	$0.14_{\pm 0.04}$	0.77	$0.76_{\pm 0.09}$
ELECTRA	0.13	$0.01_{\pm 0.08}$	0.75	$0.64_{\pm 0.13}$	$0.25^{\dagger}$	$0.19_{\pm 0.05}$	0.75	$0.73_{\pm 0.01}$	$0.23^{\dagger}$	$0.16_{\pm 0.04}$	0.78	$0.76_{\pm 0.01}$

Table 4.1: Results for the RSA experiments show max and mean (across layers) Kendall's τ; correlations that are significantly non-zero are marked with *, † and § for $p < 0.05$, < 0.01 and < 0.001 respectively. Results for the linear mapping experiments show max and mean selectivity.

CONTROL TASK As proposed by Hewitt and Liang (2019b), we construct a random control task for the linear mapping experiments, wherein we randomly swap each color chip's CIELAB code for another. This is meant to break the mapping between the color chips and their corresponding terms. Control task results are reported as the mean of 10 different random re-mappings. We report probe *selectivity*, which is defined as the difference between proportion of explained variance in the standard experimental condition and in the control task (Hewitt and Liang, 2019b). We run similar control for the RSA experiments, where the CIELAB space centroids are randomly shuffled.

4.5 RESULTS

Table 4.1 shows the max, mean, and standard deviation (across layers) of alignment scores for each of the LMs, per alignment method and setting. For RSA, we observe significant correlations across all configurations: most LM layers show a topological alignment with color space. Notably, this is also true for the static embeddings and for one

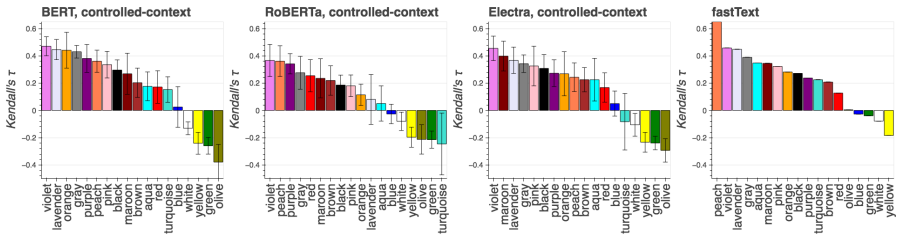


Figure 4.3: RSA results (Kendal's τ) broken down by color term for each of the LMs under the CC configuration and for the fastText baseline.

of the PMI baselines (Table 4.2). Although some variance is observed,[9] the presence of significant correlations is telling, given the small sample size (18). Furthermore, randomly permuting the color space centroids leads to RSA correlations that are non-significant for all setups ($p > 0.05$), which lends further credence to models' alignment with CIELAB structure.

Figure 4.3 shows the breakdown of correlations per color term for the three LMs under CC, as well as for fastText. We find that this ranking of color terms is largely stable across models and layer. Full RSMs for all models and CIELAB are in Appendix A.2.3. The RSMs show evidence of the higher correlations for colors like violet, orange, and purple, being driven by general clusterings of similarity/dissimilarity. For instance, for both the CIELAB and CC BERT RSMs, violet's top *nearest* neighbors include purple, lavender, pink, and orange, and its *furthest* neighbors include aqua, olive, black, and gray. Correlations do not, however, appear to be driven by consistently aligned partial orderings within the clusters. In addition, we compute RSA correlations between the different models. Results show that NC embeddings have low alignment to all others (details in appendix A.2.2).

For the linear mapping experiments, we observe the highest selectivity scores for CC (Table 4.1, right) compared to NC and RC (Table 4.1, left, middle) and baselines (Table 4.2). This validates our intuition that controlling for variation in sentence context would reveal increased alignment to color space. Furthermore, we observe that, over the full range of probe complexities for the experimental condition and the control task (described as in §4.4), all models demonstrate high selectivity (see A.2.7 for full results). It is, therefore, safe to attribute the fitness of the probes to information encoded in the color term representations, rather than to memorization. In terms of individual colors, Figure 4.4a depicts the ranking of predicted CIELAB codes per Munsell color chip for BERT (CC). We find that these results are largely stable across models and layers (see appendix A.2.6 for full set of results and for reference chart). Also, we observe that

9 In particular, results for NC show large variances across layers. The mean correlation across layers in this setup is near zero, even though max correlations for BERT and RoBERTa are significant; this is unsurprising, however, as the LM has likely never encountered single color term tokens in isolation (cf. Bommasani et al. (2020b))

Model	RSA	lin. map
pmi-1	0.14	0.72
pmi-2	0.11	0.70
pmi-3	0.17*	0.71
fastText	0.23*	0.72

Table 4.2: Baseline results. RSA results show Kendall's τ; results with * are significantly non-zero ($p < 0.05$). Linear mapping results show selectivity.

clusterings of chips with certain modal color terms (*green, blue*) show worse rankings than the rest.

4.6 ANALYSIS AND DISCUSSION

Having demonstrated the existence of models' alignment to CIELAB across various configurations, we now present an analysis and discussion of these results.

DIMENSIONALITY OF COLOR SUBSPACE Previous work has shown that linguistic information such as part-of-speech category, dependency relation type, and word sense, is expressed in low-dimensional subspaces of language model representations (Durrani et al., 2020; Hernandez and Andreas, 2021; Reif et al., 2019b). We investigate the dimensionality of the subspace required to predict the CIELAB chip codes from the term embeddings, following the methodology of Durrani et al. (2020). Averaging over the three predicted CIELAB dimensions, we rank the linear mapping coefficients (from the experiments described in §4.3), sorting the weights by their absolute values in descending order. Results (appendix A.2.8) show that across models and layers, ~0.4 of the variance in the CIELAB chip codes can be explained by assigning 95% of the weights to ~10 dimensions. 3040 dimensions are sufficient to explain ~0.7 of the variance, nearly the proportion of variance explained by the full representations (Table 4.1).

EFFECT OF MODEL SIZE We also evaluate the effect of model size on alignment by testing four smaller BERT (CC) models using the same setup described above. The results (see appendix A.2.9 for details) show that alignment as measured by both RSA and linear mapping progressively increases with model size, meaning that that with growing complexity, model representational geometry of color terms moves towards isomorphism to CIELAB.

COLOR TEMPERATURE In Figures 4.3 & 4.4a we observe that on average, warmer colors (*yellow, orange, red, etc.*) show a closer align-

ment than cooler ones (*blue*, *green*, *etc.*). In recent work, Gibson et al. (2017) reported that the former are on average communicated more efficiently (see next paragraph) than the latter, across languages. This is attributed to warmer colors being more prevalent as colors of behaviorally relevant items in the environment — salient objects — compared to cooler ones, which occur more often as background colors. To verify this observation, we partition the space of chips into two (see appendix A.2.4 for details) and compute the average explained variance across warm and cool colors. The results (see appendix A.2.4 for plots) show that, term embeddings of warm colors are better aligned to CIELAB than those of cool ones, across models and configurations. This is consistent with the bias described in Gibson et al. (2017), which we conjecture might be filtering through into the distributional statistics of (color terms in) textual corpora, influencing the representations learned by various methods which leverage these statistics.

CONNECTION TO LISTENER SURPRISAL Gibson et al. (2017)'s findings are based on the application of an information theoretic analysis to color naming, framing it as a communication game where a speaker has a particular color chip c in mind and uses a word w to indicate it then a listener has to correctly guess c, given w. Communication efficiency is measured through surprisal, S, which in this setting corresponds to the average number of guesses an optimal listener takes to arrive at the correct color chip. We calculate $S(c)$ for each chip in the color lexicon. Surprisal is defined as $S(c) = \sum_w P(w|c) \cdot \log\left(\frac{1}{P(c|w)}\right)$, where $P(w|c)$ is the probability that a color c gets labeled as w and $P(c|w)$ is computed using Bayes Theorem. Here, $P(w)$ represents how often a particular word gets used across the color space (and participants), and $P(c)$ is a uniform prior. Figure 4.4b shows surprisal per chip. High surprisal chips correspond to a lower color naming consensus among speakers, meaning that a more variable range of terms is used for these (color) contexts. We hypothesize that this could be reflected in the representations of color terms corresponding to high surprisal chips. To test this, we compute Spearman's correlation (ρ) between a chip's regression score (predicted color chip code ranking) and its surprisal. We find significant Spearman's rank correlation between lower ranking and higher surprisal for all LMs under all configurations ($0.12 \leqslant \rho \leqslant 0.17$, $p < 0.05$).

WHAT FACTORS PREDICT COLOR SPACE ALIGNMENT? Given that LMs are trained exclusively on text corpora, we hypothesize that alignment between their embeddings and CIELAB is influenced by corpus usage statistics. To determine exactly which factors could predict alignment score, we extract color term log frequency, part-of-speech tag (POS), dependency relation (DREL), and dependency tree head (HEAD) statistics for all color terms from a dependency-parsed

(Straka et al., 2016) common crawl corpus. In addition to this, we compute, per color term, the entropy of its normalised PMI distribution (`pmi-col`, see §4.3) as a measure of collocation. [10]

We then fit a Linear Mixed Effects Model (Gałecki and Burzykowski, 2013) to the features enumerated above, with RSA score (Table 4.1) as the response variable, and model type as a random effect. Here, we follow a multi-level step-wise model building sequence, where a baseline model is first fit with color term log frequency as a single fixed effect. A model which includes `pmi-col` as an additional fixed effect is then fit following this, where these two terms are included as control predictors in all later models. Following this, we compute POS, DREL, and HEAD lemma distribution entropies per color term (`pos-ent, deprel-ent, head-ent`). Higher entropies indicate that the term is employed in more diverse contexts with respect to those categories. Following entropy computation, we separately fit models to each of the three entropy statistics. Finally, we calculate the proportion of: POS tags that are adjectives `adj-prop`, DRELs that are adjectival modifiers `amod-prop`, and those that are copulas `cop-prop`. The first two evaluate the effect a color term occurring more or less often as an adjectival modifier, while the latter tests the hypothesis that assertions such as *The banana is yellow* could provide indirect grounding (Merrill et al., 2021), thereby leading to higher alignment. Including the entropy term which led to the best fit (`deprel-ent`) in the previous level, models are fit including terms for each of the proportion statistics. Model comparison is carried out by computing the log likelihood ratio between models that differ in a single term. Results show: (a) `pmi-col` significantly improves fit above log frequency and has a negative coefficient, meaning that terms that occur in more fixed collocations are less aligned to the perceptual space, (b) `deprel-ent` and `head-ent` but not `pos-ent` lead to a significantly improved fit compared to the control predictors; we observe positive coefficients for both, indicating RSA score is higher for terms that occur in more varied syntactic dependency relations and modify a more diverse set of heads, (c) out of the proportion statistics, only the `amod-prop` term improves fit; it has a positive coefficient, thus, color terms occurring more frequently as adjectival modifiers show higher scores. See appendix A.2.10 for model details.

VISION-AND-LANGUAGE MODELS In a preliminary set of experiments, we evaluated multi-modal Vision-and-Language models (VisualBERT (Li et al., 2019) and VideoBERT (Sun et al., 2019)), finding no major differences in results from the text-only models presented in this study.

10 Low entropy reflects frequent co-occurrence with a small subset of the vocabulary and high entropy the converse.

4.7 RELATED WORK

Distributional word representations have long been theorized to capture various types of information about the world (Schütze, 1992). Early work in this regard employed semantic similarity and relatedness datasets to measure alignment to human judgements (Agirre et al., 2009; Bruni et al., 2012; Hill et al., 2015). Rubinstein et al. (2015), however, question whether the distributional hypothesis is equally applicable to all types of semantic information, finding that taxonomic properties (such as animacy) are better modelled than attributive ones (color, size, etc.). To a similar end, Lucy and Gauthier (2017) analyze how well distributional representations encode various aspects of grounded meaning.

They investigate whether language models would *"be worse off for not having physically bumped into walls before they hold discussions on wall-collisions?"*, finding that perceptual features are poorly modelled compared to encyclopedic and taxonomic ones.

More recently, several studies have asked related questions in the context of language models. For example, Feldman et al. (2019) and Petroni et al. (2019) mine LMs for factual and commonsense knowledge by converting knowledge base triplets into cloze statements that are used to query the models. In a similar vein, Forbes et al. (2019) investigate LM representations' encoding of object properties (e.g., *oranges are round*), and affordances (e.g. *oranges can be eaten*), as well as the interplay between the two. Weir et al. (2020) demonstrate that LMs can capture *stereotypic tacit assumptions* about generic concepts, showing that they are adept at retrieving concepts given their associated properties (e.g., **bear** given *A ___ has fur, is big, and has claws.*). Similar to other work, they find that LMs better model encyclopedic and functional properties than they do perceptual ones. In an investigation of whether or not LMs are able to overcome reporting bias, Schwartz et al. (2017) extract all sentences in Wikipedia where one of 11 color terms modifies a noun and test how well predicted the color term is when it is masked. They find that LMs are able to model this relationship between concepts and associated colors to a certain extent, but are prone to over-generalization. Finally, Ilharco et al. (2020) train a probe to map LM representations of textual captions to paired visual representations of image patches, in order to evaluate how useful the former are for discerning between different visual representations. They find that many recent LMs yield representations that are effective at retrieving semantically-aligned image patches, but still far under-perform humans.

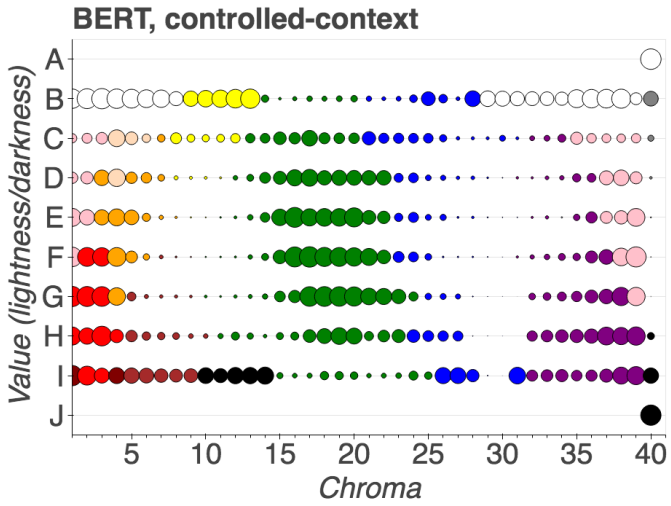
4.8 OUTLOOK

It is commonly held that the learning of phenomena which rely on sensory perception is only possible through direct experience. Indeed, the view that people born blind could not be expected to acquire coherent knowledge about colors has been prevalent since at least the empiricist philosophers (Hume, 1938; Locke, 1847) and still holds currency (Jackson, 1982). Nevertheless, recent research highlighting the contribution of language and of semantic associations between concepts towards learning has demonstrated that the congenitally blind do in fact show a striking understanding of both color similarity (Saysani et al., 2018) and object colors (Kim et al., 2020).

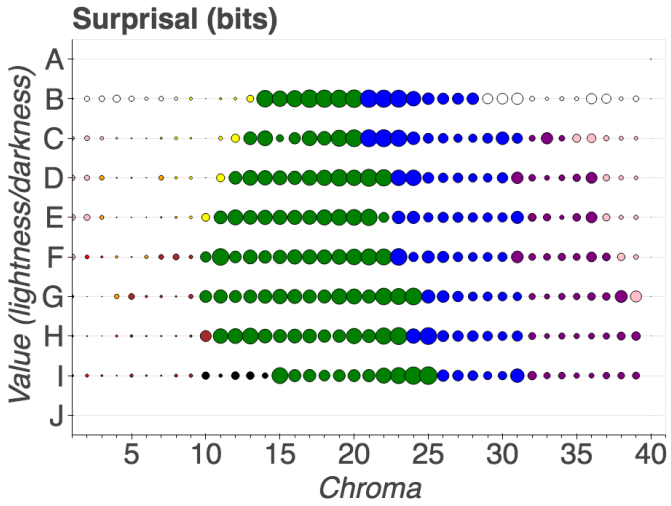
This paper investigated whether representations of color terms that are derived from text only express a degree of isomorphism to the structure of humans' perceptual color space.[11] Results from our experiments evidenced that such a topological correspondence exists. Notably, color term representations based on simple co-occurance statistics already demonstrated correspondence; those extracted from language models aligned more closely. We observed that warm colors, on average, show more alignment than cooler ones, linking to recent findings on communication efficiency in color naming (Gibson et al., 2017). Further analysis based on surprisal — an information theoretic measure, used to evaluate how efficiently a color is communicated between a speaker and a listener — revealed a correlation between lower topological alignment and higher color chip surprisal, suggesting that the kind of contexts a color occurs in play a role in determining alignment. Exploring this, we tested a set of color term corpus-derived statistics for how well they predict alignment, finding that a measure of a color term's collocationality corresponds to lower alignment, while the entropy of its dependency relation distribution and it occurring more frequently as and adjectival modifier correspond to closer alignment.

Our results and analyses provide motivation for future work on understanding the types of information we expect language models to learn with and without grounded or embodied language learning approaches.

11 Clearly, complete isomorphism is rather unlikely: language in general, and color terms by extension, are far from being simply denotational, and language interacts with and is influenced by a myriad of factors besides perception.



(a) Each circle on the chart represents the ranking of the predicted color chip when ranked according to Pearson distance from gold (larger circle ≅ higher/better ranking).



(b) Each circle on the chart represents a color chip's suprisal score (larger circle ≅ higher score).

Figure 4.4: (a) shows linear mapping results for BERT, under the CC configuration, broken down by Munsell color chip; (b) shows suprisal per chip. Circle colors reflect the modal color term assigned to the chips.

Part III

LINGUISTIC THEORY

5

THE SENSITIVITY OF LANGUAGE MODELS AND HUMANS TO PERTURBATIONS

5.1 ABSTRACT

Large-scale pretrained language models are the major driving force behind recent improvements in performance on the Winograd Schema Challenge, a widely employed test of commonsense reasoning ability. We show, however, with a new diagnostic dataset, that these models are sensitive to linguistic perturbations of the Winograd examples that minimally affect human understanding. Our results highlight interesting differences between humans and language models: language models are more sensitive to number or gender alternations and synonym replacements than humans, and humans are more stable and consistent in their predictions, maintain a much higher absolute performance, and perform better on non-associative instances than associative ones. Overall, humans are correct more often than out-of-the-box models, and the models are sometimes right for the wrong reasons. Finally, we show that fine-tuning on a large, task-specific dataset can offer a solution to these issues.

5.2 INTRODUCTION

Large-scale pre-trained language models have recently led to improvements across a range of natural language understanding tasks (Devlin et al., 2019b; Radford et al., 2019; Yang et al., 2019), but there is some scepticism that benchmark leaderboards do not represent the full picture (Jumelet and Hupkes, 2018; Kaushik and Lipton, 2018; Poliak et al., 2018b). An open question is whether these models generalize beyond their training data samples.

In this paper, we examine how pre-trained language models generalize on the Winograd Schema Challenge .

Named after Terry Winograd, the WSC, in its current form, was proposed by Levesque et al. (2012) as an alternative to the Turing Test. The task takes the form of a binary reading comprehension test where a statement with two referents and a pronoun (or a possessive adjective) is given, and the correct antecedent of the pronoun must be chosen. Examples are chosen carefully to have a preferred reading, based on semantic plausibility rather than co-occurrence statistics. WSC examples come in pairs that are distinguished only by a discriminatory segment that *flips* the correct referent, as shown in Figure 1a. Levesque et al. (2012) define a set of qualifying criteria

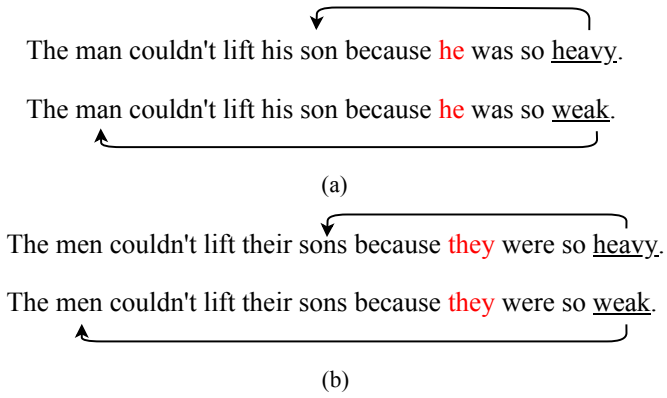


Figure 5.1: An example pair from the Winograd Schema Challange (a) and its perturbation (b). The pronoun resolves to one of the two referents, depending on the choice of the discriminatory segment. The perturbation in (b) pluralizes the referents and the antecedents.

for instances and the pitfalls to be avoided when constructing examples (see §5.4.2). These combine to ensure an instance functions as a test of what they refer to as 'thinking' (or common sense reasoning).

Recent work has reported significant improvements on the WSC (Kocijan et al., 2019; Sakaguchi et al., 2019). As with many other NLU tasks, this improvement is primarily due to large-scale language model pre-training, followed by fine-tuning for the target task. We believe that further examination is warranted to determine whether these impressive results reflect a fundamental advance in reasoning ability, or whether our models have learned to simulate this ability in ways that do not generalize. In other words, do models learn accidental correlations in our datasets, or do they extract patterns that generalize in robust ways beyond the dataset samples?

In this paper, we conduct experiments to investigate this question. We define a set of lexical and syntactic variations and perturbations for the WSC examples and use altered examples (Figure 1b) to test models that have recently reported improved results. These variations and perturbations are designed to highlight the robustness of human linguistic and reasoning abilities and to test models under these conditions.

CONTRIBUTIONS We introduce a new Winograd Schema dataset for evaluating generalization across seven controlled linguistic perturbations.[1] We use this dataset to compare human and language model sensitivity to those perturbations, finding marked differences in model performance. We present a detailed analysis of the behaviour of the language models and how they are affected by the perturba-

1 Code and dataset can be found at: https://github.com/mhany90/enhanced_wsc/

tions. Finally, we investigate the effect of fine-tuning with large task-specific datasets, and present an error analysis for all models.

5.3 RELATED WORK

PROBING DATASETS Previous studies have explored the robustness of ML models towards different linguistic phenomena (Belinkov and Glass, 2019b), e.g., by creating challenge datasets such as the one introduced here. When predicting subject-verb agreement, Linzen et al. (2016) found that inserting a relative clause hurt the performance of recurrent networks.[2]

A large body of research has since emerged on probing pre-trained (masked) language models for linguistic structure (Clark et al., 2019b; Goldberg, 2019; Hewitt and Manning, 2019; Lin et al., 2019) and analysing them via comparison to psycholinguistic and brain imaging data (Abdou et al., 2019; Abnar et al., 2019b; Ettinger, 2019; Gauthier and Levy, 2019). Other recent work has attempted to probe these models for what is referred to as *common sense* or factual knowledge (Feldman et al., 2019; Petroni et al., 2019). Their findings show that these models do indeed encode such knowledge and can be used for knowledge base completion or common sense mining from Wikipedia.

CLEVER HANS A considerable amount of work has also been devoted to what might be described as the Clever Hans effect. This work has aimed to quantify the extent to which models are learning what we expect them to as opposed to leveraging statistical artifacts. This line of work has to date revealed significant problems (and some possible solutions to those problem) with reading comprehension datasets (Chen et al., 2016b; Kaushik and Lipton, 2018), natural language inference datasets (Belinkov et al., 2019b; Gururangan et al., 2018b; McCoy et al., 2019; Poliak et al., 2018b; Tsuchiya, 2018), and the story cloze challenge (Schwartz et al., 2017), among others.

WINOGRAD SCHEMA CHALLENGE Trinh and Le (2018) first proposed using neural language models for the WSC, achieving an accuracy of 63.7% using an ensemble of 14 language models. Ruan et al. (2019) and Kocijan et al. (2019) fine-tune BERT (Devlin et al., 2019b) on the PDP (Rahman and Ng, 2012) and an automatically generated MaskedWiki dataset, reaching an accuracy of 71.1% and 72.5% respectively. Meanwhile, Radford et al. (2019) report an accuracy of 70.7% without fine-tuning using the GPT-2 language model.

2 This contrasts with our results with Transformer-based architecture and is probably explained by memory loss in recurrent networks trained on short sequences. Similarly, Gulordava et al. (2018) tested whether a Recurrent Neural Network can predict long-distance number agreement in various constructions comparing natural and nonsensical sentences where RNNs cannot rely on semantic or lexical cues.

	Instance / Perturbed Instance	Count
Original	Sid explained his theory to Mark but he couldn't convince him.	285
Tense	Sid is explaining his theory to Mark but he can't convince him.	281
Number	**Sid and Johnny** explained their theory to **Mark and Andrew** but they couldn't convince them.	253
Gender	**Lucy** explained her theory to **Emma** but she couldn't convince her.	155
Voice	The theory was explained by Sid to Mark but he couldn't convince him.	220
Relative clause	Sid, **which we had seen on the discussion panel with Chris,** explained his theory to Mark but he couldn't convince him.	283
Adverb	Sid **diligently** explained his theory to Mark but he couldn't convince him.	283
Synonyms/Names	**John** explained his theory to **Jad** but he couldn't convince him.	285

Table 5.1: Examples from our dataset of the different perturbations applied to a WSC instance.

Most recently, Sakaguchi et al. (2019) present an adversarial filtering algorithm which they use for crowd-sourcing a large corpus of WSC-like examples. Fine-tuning RoBERTa (Liu et al., 2019b) on this, they achieve an accuracy of 90.1%.

In an orthogonal direction, Trichelair et al. (2018) presented a timely critical treatment of the WSC. They classified the dataset examples into associative and non-associative subsets, showing that the success of the LM ensemble of Trinh and Le (2018) mainly resulted from improvements on the associative subset. Moreover, they suggested switching the candidate referents (where possible) to test whether systems make predictions by reasoning about the "entirety of a schema" or by exploiting "statistical quirks of individual entities".

In a similar spirit, our work is a controlled study of robustness along different axes of linguistic variation. This type of study is rarely possible in NLP due to the large size of datasets used and the focus on obtaining improved results on said datasets. Like a carefully constructed dataset which is thought to require true natural language understanding, the WSC presents an ideal testbed for this investigation.

5.4 PERTURBATIONS

We define a suite of seven perturbations that can be applied to the 285 WSC examples, which we refer to as the original examples. These perturbations are designed to test the robustness of an answer to semantic, syntactic, and lexical variation. Each of the perturbations is applied to every example in the WSC (where possible), resulting in a dataset of 2330 examples, an example of each type is shown in Table 5.1. Crucially, the correct referent in each of the perturbed examples is

not altered by the perturbation. The perturbations are manually constructed, except for the sampling of names and synonyms. Further details can be found in Appendix A.3.5.

TENSE SWITCH (TEN) Most WSC instances are written in the past tense and thus are changed to the present continuous tense (247 examples). The remaining 34 examples are changed from the present to the past tense.

NUMBER SWITCH (NUM) Referents have their numbers altered: singular referents (and the relevant pronouns) are pluralised (223 examples), and plural referents are modified to the singular (30 examples). Sentences with names have an extra name added via conjunction; eg. "Carol" is replaced with "Carol and Susan". Possessives only mark possession on the second conjunct ("John and Steve's uncle" rather than "John's and Steve's uncle").

GENDER SWITCH (GEN) Each of the referents in the sentence has their gender switched by replacing their names with other randomly drawn frequent English names of the opposite gender.[3] 92% of the generated data involved a gender switch for a name. Though humans may be biased towards gender (Collins, 2011; Desmond and Danilewicz, 2010; Hoyle et al., 2019), the perturbations do not introduce ambiguity concerning gender, only the entity. 101 examples were switched from male to female, and 55 examples the other way around.

VOICE SWITCH (VC) All WSC examples, except for 210 and 211, are originally in the active voice and are therefore passivized. 210 and 211 are changed to the active voice. 65 examples could not be changed. Passive voice is known to be more difficult to process for humans (Feng et al., 2015; Olson and Filby, 1972).

RELATIVE CLAUSE INSERTION (RC) A relative clause is inserted after the first referent. For each example, an appropriate clause was constructed by first choosing a template such as "who we had discussed" or "that is known for" from a pre-selected set of 19 such templates. An appropriate ending, such as "who we had discussed *with the politicians*" is then appended to the template depending on the semantics of the particular instance. Relative clauses impose an increased demand on working memory capacity, thereby making processing more difficult for humans (Gibson, 1998; Just and Carpenter, 1992).

3 Names sourced from https://github.com/AlessandroMinoccheri/human-names/tree/master/data

ADVERBIAL QUALIFICATION (ADV) An adverb is inserted to qualify the main verb of each instance. When a conjunction is present both verbs are modified. For instances with multiple sentences, all main verbs are modified.

SYNONYM/NAME SUBSTITUTION (SYN/NA) Each of the two referents in an example is substituted with an appropriate synonym, or if it is a name, is replaced with a random name of the same gender from the same list of names used for the gender perturbation.

5.4.1 *Human Judgments*

We expect that humans are robust to these perturbations because they represent naturally occurring phenomena in language; we test this hypothesis by collecting human judgements for the perturbed examples. We collect the judgments for the perturbed examples using Amazon Mechanical Turk. The annotators are presented with each instance where the pronoun of interest is boldfaced and in red font. They are also presented with two options, one for each of the possible referents. They are then instructed to choose the most likely option, in exchange for $0.12. Following Sakaguchi et al. (2019), each instance is annotated by three annotators and majority vote results are reported. Results are reported later in §5.6. All three annotators agreed on the most likely option in 82-83% of the instances, except for gender, where a full agreement was obtained for only 78% of the instances. See Appendix A.3.2 for further annotation statistics, a sample of the template presented to annotators, and restrictions applied to pool of annotators. We did not require an initial qualification task to select participants.

5.4.2 *Confounds and Pitfalls*

Constructing WSC problems is known to be difficult. Indeed, the original dataset was carefully crafted by domain experts and subsequent attempts at creating WSC-like datasets by non-experts such as in Rahman and Ng (2012) have produced examples which were found to be less challenging than the original dataset. Two likely pitfalls listed in Levesque et al. (2012) concern **A**) statistical preferences which make one answer more readily associated with the special discriminatory segment or other components of an example[4] (this is termed as *Associativity*, and it is described as *non-Google-proofness* in Levesque et al. (2012)); and **B**) inherent ambiguity which makes the examples open to other plausible interpretations. In what follows, we discuss these pitfalls, demonstrating that the perturbed examples remain resilient to both.

4 Trichelair et al. (2018) find that 13.5% of examples from the original WSC might still be considered to be *associative*.

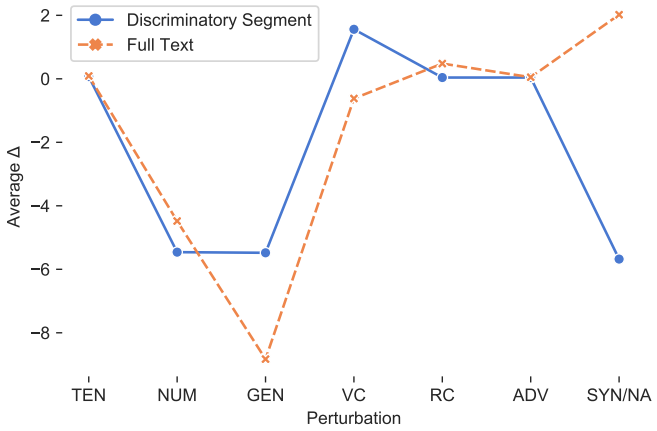


Figure 5.2: PMI divergence from the original WSC examples in average Δ for each perturbation. Values below 0 indicate that the difference in PMI between the correct candidate and the incorrect one decreased.

QUANTIFYING ASSOCIATIVITY To verify that the perturbations have not affected the correctness of the original problems with regards to pitfall **A**, we employ pointwise mutual information to test the associativity of both the original and perturbed examples. PMI is known to be a reasonable measure of associativity (Church and Hanks, 1990) and, among a variety of measures, has been shown to correlate best with association scores from human judgements of contextual word association (Frassinelli, 2015). We compute unigram PMI on the two corpora used to train BERT (see Appendix A.3.3 for details). Figure 5.2 shows the *divergence* of the perturbed examples from the original WSC dataset. We estimate divergence as the average difference in PMI between the correct ($\mathcal{C}$) and incorrect ($\mathcal{I}$) candidates: $\Delta = \texttt{pmi}(c_j, x_j) - \texttt{pmi}(i_j, x_j)$ where $\mathcal{X}$ is either: i) the discriminatory segments or ii) the full text of the example, and $\texttt{pmi}(\cdot,\cdot)$ is average unigram PMI. Δ can be seen as a measure of whether the correct or incorrect candidate is a better 'associative fit' for either the discriminatory segment or the full context, making the examples trivial to resolve. Observe that this difference in PMI declines for the perturbed examples, showing that these the perturbed example do not increase in associativity.

CONFIRMING SOLVABILITY Three expert annotators[5] are asked to solve the small subset of examples (99 in total across perturbations) which were annotated incorrectly by the majority vote of Mechanical Turk workers. To address pitfall **B**, the expert annotators are asked to both attempt to solve the instances and indicate if they believe them to be *too ambiguous* to be solved. The majority vote of the annotators

5 Graduate students of linguistics.

determines the preferred referent or whether an instance is ambiguous. Out of a total of 99 examples, 10 were found to be ambiguous. Of the remaining 89 examples, 67 were answered correctly by the majority vote. See Appendix A.3.4 for more details.

5.5 EXPERIMENTAL PROTOCOL

Our experiments are designed to test the robustness of language models to the Winograd Schema perturbations described in the previous section.

EVALUATION Models are evaluated using two types of measures. The first is *accuracy*. For each of the perturbations, we report (a) the accuracy on the perturbed set (**Perturbation accuracy**), (b) the difference in accuracy on the perturbed set and on the *equivalent subset* of original dataset:[6]

$$\Delta_{\textbf{Acc.}} = \textbf{Perturbation accuracy} - \textbf{Original subset accuracy}$$

and (c) **Pair accuracy**, defined as the number of pairs for which both examples in the pair are correctly answered divided by the total number of pairs.

The second measure is *stability*, S. This is the proportion of perturbed examples $\mathcal{P}'$ for which the predicted referent is the same as the original prediction $\mathcal{P}$:

$$S = \frac{|\{(p'_i, p_i) \mid p'_i \in \mathcal{P}' \wedge p_i \in \mathcal{P} \wedge p'_i = p_i\}|}{|\mathcal{P}|}$$

Since the perturbations do not alter the correct referent, this provides a strong indication of robustness towards them.

BASELINE We take the unigram PMI between candidates and discriminatory segments (see §5.4.2) as a baseline. We expect that this simple baseline will perform well for instances with a high level of associativity but not otherwise.

LANGUAGE MODELS Our analysis is applied to three out-of-the-box language models (LMs): BERT (Devlin et al., 2019b), RoBERTa (Liu et al., 2019b), and XLNet (Yang et al., 2019). These models are considered to be the state-of-the-art for the wide variety of natural language understanding tasks found in the GLUE (Wang et al., 2018a) and SuperGLUE (Wang et al., 2019a) benchmarks. We use the *large* pre-trained publicly available models (Wolf et al., 2019).[7]

6 Recall that is was not possible to perturb all examples.

7 https://github.com/huggingface/pytorch-transformers

	ORIG	TEN	NUM	GEN	VC	RC	ADV	SYN/NA	Avg	Avg $\Delta_{Acc.}$
PMI	54.38	54.09	52.96	57.42	54.09	54.41	54.41	51.92	54.24	−2.13
BERT	61.75	61.92	57.31	57.42	63.64	62.19	61.48	58.59	60.41	−1.26
XLNET	64.56	60.14	62.45	62.58	57.73	62.9	64.31	61.05	61.59	−2.78
ROBERTA	69.82	69.40	64.43	53.55	66.82	68.55	69.61	57.54	64.27	−5.16
BERT+WW	72.28	70.46	71.15	74.84	65.91	64.31	72.44	70.88	70.00	−2.82
ROBERTA+WG	88.42	89.32	88.53	86.45	83.63	86.93	88.7	89.05	87.62	−1.06
HUMANS	97.89	96.79	94.46	92.25	92.27	91.16	95.40	96.14	94.41	−3.83

Table 5.2: Original dataset accuracy (ORIG) and **Perturbation accuracy** results for all models and humans. The penultimate column shows the average **Perturbation accuracy** results. The rightmost column shows the $\Delta_{\textbf{Acc.}}$ results, averaged over all perturbations.

FINE-TUNED LANGUAGE MODELS We also examine the effect of fine-tuning language models. BERT+WW uses BERT fine-tuned on the MaskedWiki and WscR datasets which consist of 2.4M and 1322 examples (Kocijan et al., 2019), and RoBERTa+WG is fine-tuned on WinoGrande **XL**, which consists of 40,938 adversarially filtered examples (Sakaguchi et al., 2019). Both fine-tuned models have been reported by recent work to achieve significant improvements on the WSC.

SCORING To score the two candidate referents in each WSC instance we employ one of two mechanisms. The first, proposed in Trinh and Le (2018) and adapted to masked LMs by Kocijan et al. (2019) involves computing the probability of the two candidates c1 and c2, given the rest of the text in the instance s. To accomplish this, the pronoun of interest is replaced with a number of `MASK` tokens corresponding to the number of tokens in each of c1 and c2. The probability of a candidate, $p(c|s)$ is then computed as the average of the probabilities assigned by the model to the candidate's tokens and the maximum probability candidate is taken as the answer. This scoring method is used for all models, except RoBERTa+WG. For that, we follow the scoring strategy employed in Sakaguchi et al. (2019) where an instance is split into context and option using the candidate answer as a delimiter.[8]

5.6 RESULTS AND ANALYSIS

Following the experimental protocol, we evaluate the three out-of-the-box language models and the two fine-tuned models on the original WSC and each of the perturbed sets. Table 5.2 shows **Perturbation ac-**

8 `[CLS] context [SEP] option [SEP]`, e.g. *[CLS] The sculpture rolled off the shelf because ____ [SEP] wasn't anchored [SEP].* The blank is filled with either option 1 (*the sculpture*) or 2 (*the trophy*).

curacy results for all models[9] and contrasts them with human judgements and the PMI baseline.

5.6.1 *Language Models*

Humans maintain a much higher performance compared to out-of-the-box LMs across perturbations. The difference in accuracy between the perturbed and original examples, $\Delta_{\text{Acc.}}$, as defined in Section 5.5 is shown in Figure 5.4. A general trend of decrease can be observed for both models and humans across the perturbations. This decline in accuracy is on average comparable between models and humans — with a handful of exceptions. Taking the large gap in absolute accuracy into account, this result might be interpreted in two ways. If a comparison is made relative to the upper bound of performance, human performance has suffered from a larger error increase. Alternately, if we compare relative to the lower bound of performance, then the decline in the already low performance of language models is more meaningful, since 'there is not much more to lose'.

A more transparent view can be gleaned from the stability results shown in Table 5.3. Here it can be seen that the three out-of-the-box LMs are *substantially more likely* to switch predictions due to the perturbations than humans. Furthermore, we observe that the LMs are least stable for word-level perturbations like gender (GEN), number (NUM), and synonym or name replacement (SYN/NA), while humans appear to be most affected by sentence-level ones, such as relative clause insertion (RC) and voice perturbation (VC).

Understanding Language Model Performance

To better understand the biases acquired through pre-training which are pertinent to this task, we consider a) a case of essential feature omission and b) the marginal cases where LMs answer very correctly or incorrectly, in both the original and perturbed datasets. We present analysis for BERT, but similar findings hold for the other LMs.

MASKING DISCRIMINATORY SEGMENTS result in identical sentence pairs because these segments are the only part of a sentence that sets WSC pairs apart (see Figure 1a). To determine whether there is a bias in the selectional preference for one of the candidates over the other, we test BERT on examples where these discriminatory segments have been replaced with the `MASK` token. An unbiased model should be close to random selection but BERT consistently prefers (by a margin of ~25-30%) the candidate which appears second in the text to the one appearing first, for all perturbations except voice, where it prefers the

9 It is interesting to note that XLNet is trained on CommonCrawl which indexes an online version of the original WSC found here: https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html.

first. This observation holds even when the two referents are inverted, which is possible for the 'switchable' subset of the examples as shown in Trichelair et al. (2018). This indicates that the selections are not purely semantic but also syntactic or structural and it points towards BERT having a preference referents in the object role. Detailed results are presented in Appendix A.3.6.

	TEN	NUM	GEN	VC	RC	ADV	SYN/NA	Avg
PMI	100	100	73.91	100	100	100	100	96.27
BERT	89.32	69.17	88.39	79.55	83.75	91.87	68.42	81.40
XLNET	82.21	69.17	66.45	69.55	78.45	84.81	70.53	75.02
ROBERTA	91.46	77.47	61.29	79.09	83.75	89.75	68.77	79.26
BERT+WW	90.04	83.00	89.68	80.45	81.98	92.93	85.96	85.14
ROBERTA+WG	96.08	94.07	97.41	91.36	92.22	94.69	96.11	95.24
HUMANS	96.70	94.9	92.9	91.18	91.11	96.11	96.1	94.31

Table 5.3: Stability results for all models and humans.

MARGINAL EXAMPLES are found where the model assigns a much higher probability to one referent over the other. We extract the top 15% examples where the correct candidate is preferred by the largest margin ($P_{correct} \gg P_{incorrect}$) and the bottom 15% where the incorrect one is preferred ($P_{incorrect} \gg P_{correct}$). Surprisingly, we find that there is a large overlap (50%–60%) between these two sets of examples, both in the original and the perturbed datasets.[10] For the examples which are both the most correct and incorrect, BERT strongly prefers one of the candidates without considering the special discriminatory segment which *flips* the correct referent. Indeed we find that the correlation between the probability assigned by BERT to a referent when it is the correct referent and when it is not is very strong and significant, with Spearman's $\rho \approx 0.75$ across perturbations (see Appendix A.3.8 for details).

10 To clarify, consider the following original WSC pair:

(i) Alice looked for her friend Jade in the crowd. Since she always has good luck, Alice spotted her quickly.

(ii) Alice looked for her friend Jade in the crowd. Since she always wears a red turban, Alice spotted her quickly.

The first example gives $P_{correct} \gg P_{incorrect}$ by the largest margin, and its counterpart gives $P_{incorrect} \gg P_{correct}$ by the largest margin. In other words, the model assigns a *much higher probability* for *Alice* in both cases.

5.6.2 *The effect of fine-tuning*

The accuracy and stability results (Tables 5.2 and 5.3) indicate that fine-tuning makes language models more robust to the perturbations. RoBERTa+WG, in particular, is the most accurate and most stable model. While impressive, this is not entirely surprising: fine-tuning on task-specific datasets is a well-tested recipe for bias correction (Belinkov et al., 2019a). Indeed, these results provide evidence that it is possible to construct larger fine-tuning datasets whose distribution is correct for the WSC. We note that both fine-tuned models perform worst on the VC and RC perturbations, which may not frequently occur in the crowd-sourced datasets used for fine-tuning. To test this intuition, we apply a dependency parser (UDPipe (Straka et al., 2016)) to the WinoGrande XL examples, finding that only ~ 5% of the examples are in the passive voice and ~ 6.5% contain relative clauses.

HOW MUCH FINE-TUNING DATA IS NEEDED? To quantify the amount of fine-tuning data needed to achieve robustness, we fine-tune RoBERTA on the five WinoGrande training set splits defined by Sakaguchi et al. (2019): **XS** (160)[11], **S** (640), **M** (2558), **L** (10234), and **XL** (40398). Figure 5.3 shows the average accuracy and stability scores for the models fine-tuned on each of the training splits[12]. We observe that the two smallest splits do not have a sufficient number of examples to adequately bias the classification head, leading to near-random performance. The model fine-tuned on the **M** split—with just 2558 examples—is, however, already able to vastly outperform the non-fine-tuned RoBERTA. Increasing the number of examples five-fold and twenty-fold leads to significant but fast diminishing improvements.

HOW DO PERTURBATIONS AFFECT TOKEN PROBABILITY DISTRIBUTIONS? To obtain a holistic view of the effect the perturbations have on LMs and fine-tuned LMs, we analyze of the shift in the probability distribution (over the entire vocabulary) which a model assigns to a `MASK` token inserted in place of the pronoun of interest. We apply probability distribution truncation with a threshold of $p = 0.9$ as proposed in Holtzman et al. (2019) to filter out the uninformative tail of the distribution. Following this, we compute the Jensen–Shannon distance between this dynamically truncated distribution for an original example and each of its perturbed counterparts. Figure 5.5 shows the

11 No. of examples in set.

12 Note that the stability score for the model fine-tuned on **XL** in Figure 5.3 is different from that reported in Table 5.3. In the latter we reported results from the model provided by Sakaguchi et al. (2019), rather than the model we fine-tuned ourselves. Since we utilise identical hyperparameters to theirs for fine-tuning, this anomalous difference in score may perhaps be explained by a difference in initialization as suggested in Dodge et al. (2020).

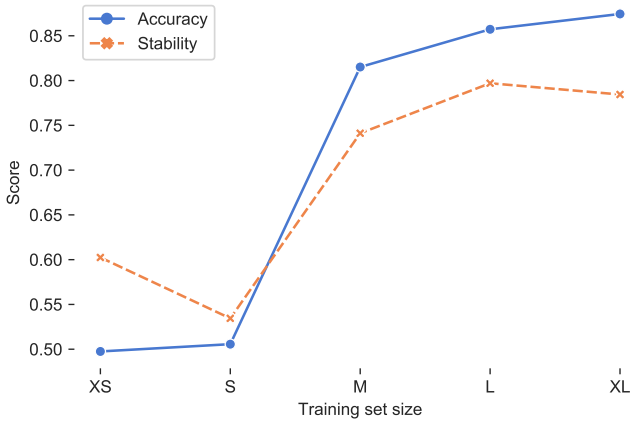


Figure 5.3: Accuracy and stability scores (averaged across perturbations) for RoBERTA when fine-tuned on five increasing training split sizes.

average of this measure over the subset of the 128 examples which are common to all perturbations. Overall, we observe that large shifts in the distribution correspond to lower stability and accuracy scores and that fine-tuned models exhibit lower shifts than their non-fine-tuned counterparts. The difference in shifts between out-of-the-box models and their fine-tuned counterparts is lower for the VC, RC and ADV perturbations, meaning that when fine-tuned, the models' probability distributions are roughly just as divergent for these perturbations as they were before fine-tuning. We hypothesize the same reasons we did in 5.6.2, which is that these examples are just under-represented in our fine-tuning corpus; indeed, these results roughly correspond to the differences in $\Delta_{\mathbf{Acc.}}$ from Figure 5.4.

Further details about the number of examples excluded via the probability distribution truncation and other measures of the perturbations' effect can be found in Appendix A.3.7.

5.6.3 *Error Analysis*

PAIR ACCURACY Here we consider a more challenging evaluation setting where each WSC pair is treated as a single instance. Since the WSC examples are constructed as minimally contrastive pairs (Levesque et al., 2012), we argue that this is an appropriate standard of evaluation. Consider again the example in Figure 1a. It is reasonable to suppose that for an answerer which truly 'understands' (Levesque et al., 2012), being able to link the concepts *heavy* and *son* in one of the resolutions is closely related and complementary to linking the concepts *weak* and *man* in the other.[13]

13 As a sanity check, consider random pairings of WSC examples. There is no such complement.

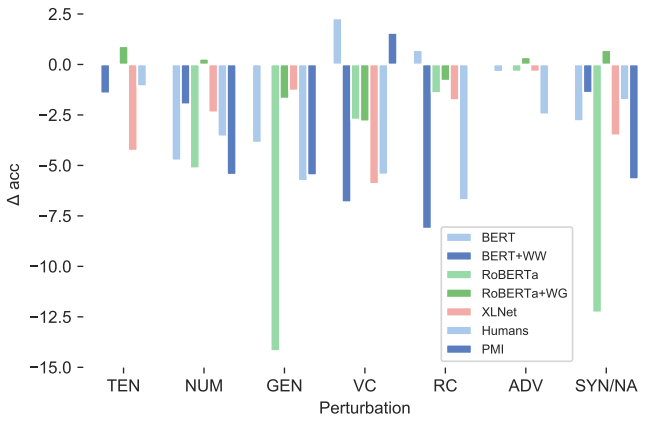


Figure 5.4: $\Delta_{\text{Acc.}}$ results for all models across perturbations. Values below the x-axis indicate a decline in accuracy compared to the original dataset.

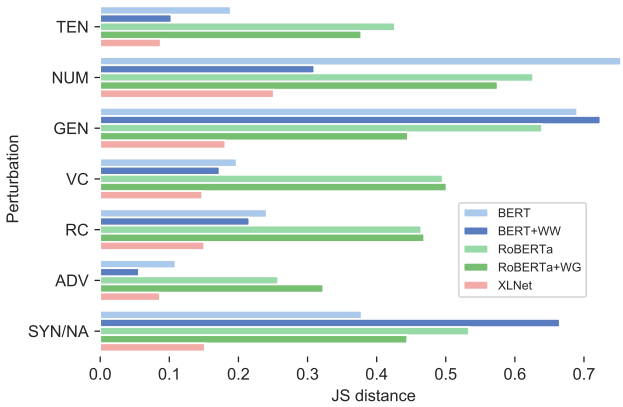


Figure 5.5: Jensen-Shannon distance between the original and perturbed examples when masking the pronoun of interest.

The results for this evaluation are shown in Figure 5.6. They show that human resolution of the problems exhibits greater complementarity compared to the language models; human pair accuracy (pair) is closer to perturbation accuracy (single) than is the case for the LMs. Furthermore, human performance on pair accuracy is more robust to perturbations when compared to the models. Indeed, the large gap between pair accuracy and perturbation accuracy raises some doubts about the performance of these models. However, RoBERTa-WG is a notable exception, showing near-human robustness to pair complementarity.

ASSOCIATIVITY Next, we examine the effect of associativity on performance. Figure 5.7 shows accuracy results[14] for all perturbations

14 Note that the large variance in results on the associative subset of gender is due to it consisting of only two examples.

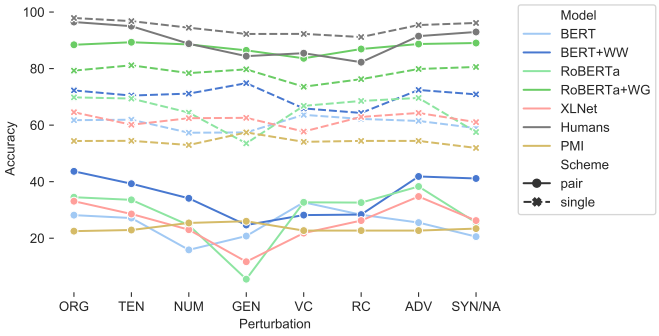


Figure 5.6: **Pair accuracy** and **Perturbation accuracy** results. The latter are labeled as *single*.

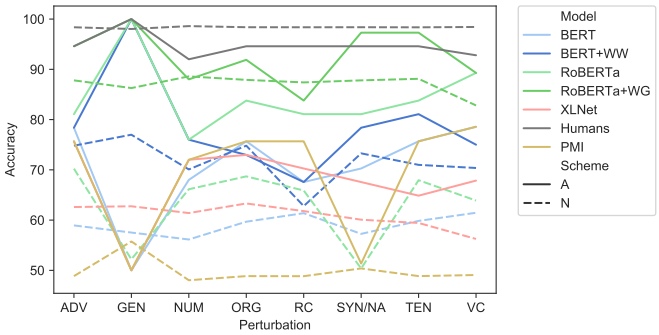


Figure 5.7: Perturbation accuracy on the Associative (A) and Non-Associative (N) subsets of the data.

on the associative and non-associative subsets of the WSC as labelled by Trichelair et al. (2018). We observe that the difference between associative and non-associative is much smaller for humans and that unlike all language models, humans do better on the former than the latter. As expected, the PMI baseline does almost as well as the LMs on the associative subset but it performs at chance level for the non-associative subset.

5.7 CONCLUSION

We presented a detailed investigation of the effect of linguistic perturbations on how language models and humans perform on the Winograd Schema Challenge. We found that compared to out-of-the-box models, humans are significantly more stable to the perturbations and that they answer non-associative examples with higher accuracy than associative ones, show sensitivity to WSC pair complementarity, and are more sensitive to sentence-level (as opposed to word-level) perturbations. In an analysis of the behaviour of language models, we observe that there is a preference for referents in the object role and that the models do not always consider the discriminatory segments of examples. Finally, we find that fine-tuning language models can

lead to much-improved accuracy and stability. It remains an open question whether this task-specific approach to generalisation constitutes a true advancement in "reasoning". Fine-tuning a model on a rather large number of examples similar to the WSC leads to increased robustness, but this stands in stark contrast to humans, who are robust to the perturbations without having been exposed to similar examples in the past.

6

ATTENTION CAN REFLECT SYNTACTIC STRUCTURE *(IF YOU LET IT)*

6.1 ABSTRACT

Since the popularization of the Transformer as a general-purpose feature encoder for NLP, many studies have attempted to decode linguistic structure from its novel multi-head attention mechanism. However, much of such work focused almost exclusively on English — a language with rigid word order and a lack of inflectional morphology. In this study, we present decoding experiments for multilingual BERT across 18 languages in order to test the generalizability of the claim that dependency syntax is reflected in attention patterns. We show that full trees can be decoded above baseline accuracy from single attention heads, and that individual relations are often tracked by the same heads across languages. Furthermore, in an attempt to address recent debates about the status of attention as an explanatory mechanism, we experiment with fine-tuning mBERT on a supervised parsing objective while freezing different series of parameters. Interestingly, in steering the objective to learn explicit linguistic structure, we find much of the same structure represented in the resulting attention patterns, with interesting differences with respect to which parameters are frozen.

6.2 INTRODUCTION

In recent years, the attention mechanism proposed by Bahdanau et al. (2014) has become an indispensable component of many NLP systems. Its widespread adoption was, in part, heralded by the introduction of the Transformer architecture (Vaswani et al., 2017), which constrains a soft alignment to be learned across discrete states in the input (self-attention), rather than across input and output (e.g., Rocktäschel et al., 2015; Xu et al., 2015). The Transformer has, by now, supplanted the popular LSTM (Hochreiter and Schmidhuber, 1997) as NLP's feature-encoder-of-choice, largely due to its compatibility with parallelized training regimes and ability to handle long-distance dependencies.

Certainly, the nature of attention as a distribution over tokens lends itself to a straightforward interpretation of a model's inner workings. Bahdanau et al. (2014) illustrate this nicely in the context of `seq2seq` machine translation, showing that the attention learned by their models reflects expected cross-lingual idiosyncrasies between English and French, e.g., concerning word order. With self-attentive

Transformers, interpretation becomes slightly more difficult, as attention is distributed across words within the input itself. This is further compounded by the use of multiple layers and heads, each combination of which yields its own alignment, representing a different (possibly redundant) view of the data. Given the similarity of such attention matrices to the score matrices employed in arc-factored dependency parsing (McDonald et al., 2005a,b), a salient question concerning interpretability becomes: Can we expect some combination of these parameters to capture linguistic structure in the form of a dependency tree, especially if the model performs well on NLP tasks? If not, can we relax the expectation and examine the extent to which subcomponents of the linguistic structure, such as subject-verb relations, are represented? This prospect was first posed by Raganato, Tiedemann, et al. (2018) for MT encoders, and later explored by Clark et al. (2019c) for BERT. Ultimately, the consensus of these and other studies (Htut et al., 2019; Limisiewicz et al., 2020; Voita et al., 2019b) was that, while there appears to exist no "generalist" head responsible for extracting full dependency structures, standalone heads often specialize in capturing individual grammatical relations.

Unfortunately, most of such studies focused their experiments entirely on English, which is typologically favored to succeed in such scenarios due to its rigid word order and lack of inflectional morphology. It remains to be seen whether the attention patterns of such models can capture structural features across typologically diverse languages, or if the reported experiments on English are a misrepresentation of local positional heuristics as such. Furthermore, though previous work has investigated how attention patterns might change after fine-tuning on different tasks (Htut et al., 2019), a recent debate about attention as an explanatory mechanism (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019) has cast the entire enterprise in doubt. Indeed, it remains to be seen whether fine-tuning on an explicit structured prediction task, e.g. dependency parsing, can force attention to represent the structure being learned, or if the patterns observed in pretrained models are not altered in any meaningful way.

To address these issues, we investigate the prospect of extracting linguistic structure from the attention weights of multilingual Transformer-based language models. In light of the surveyed literature, our research questions are as follows:

1. Can we decode dependency trees for some languages better than others?

2. Do the same layer–head combinations track the same relations across languages?

3. How do attention patterns change after fine-tuning with explicit syntactic annotation?

4. Which components of the model are involved in these changes?

In answering these questions, we believe we can shed further light on the (cross-)linguistic properties of Transformer-based language models, as well as address the question of attention patterns being a reliable representation of linguistic structure.

6.3 ATTENTION AS STRUCTURE

TRANSFORMERS The focus of the present study is mBERT, a multilingual variant of the exceedingly popular language model (Devlin et al., 2019b). BERT is built upon the Transformer architecture (Vaswani et al., 2017), which is a self-attention-based encoder-decoder model (though only the encoder is relevant to our purposes). A Transformer takes a sequence of vectors $\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_n}]$ as input and applies a positional encoding to them, in order to retain the order of words in a sentence. These inputs are then transformed into query (Q), key (K), and value (V) vectors via three separate linear transformations and passed to an attention mechanism. A single attention head computes scaled dot-product attention between K and Q, outputting a weighted sum of V:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{6.1}$$

For multihead attention (MHA), the same process is repeated for k heads, allowing the model to jointly attend to information from different representation subspaces at different positions (Vaswani et al., 2017). Ultimately, the output of all heads is concatenated and passed through a linear projection W^O:

$$H_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{6.2}$$

$$\text{MHA}(Q, K, V) = \text{concat}(H_1, H_2, ..., H_k)W^O \tag{6.3}$$

Every layer also consists of a feed-forward network (FFN), consisting of two Dense layers with ReLU activation functions. For each layer, therefore, the output of MHA is passed through a LayerNorm with residual connections, passed through FFN, and then through another LayerNorm with residual connections.

SEARCHING FOR STRUCTURE Often, the line of inquiry regarding interpretability in NLP has been concerned with extracting and analyzing linguistic information from neural network models of language (Belinkov and Glass, 2019b). Recently, such investigations have targeted Transformer models (Hewitt and Manning, 2019; Rosa and Mareček, 2019; Tenney et al., 2019), at least in part because the self-attention mechanism employed by these models offers a possible window into their inner workings. With large-scale machine translation

and language models being openly distributed for experimentation, several researchers have wondered if self-attention is capable of representing syntactic structure, despite not being trained with any overt parsing objective.

In pursuit of this question, Raganato, Tiedemann, et al. (2018) applied a maximum-spanning-tree algorithm over the attention weights of several trained MT models, comparing them with gold trees from Universal Dependencies (Nivre et al., 2020b, 2016). They found that, while the accuracy was not comparable to that of a supervised parser, it was nonetheless higher than several strong baselines, implying that some structure was consistently represented. Clark et al. (2019c) corroborated the same findings for BERT when decoding full trees, but observed that individual dependency relations were often tracked by specialized heads and were decodable with much higher accuracy than some fixed-offset baselines. Concurrently, Voita et al. (2019b) made a similar observation about heads specializing in specific dependency relations, proposing a coarse taxonomy of head attention functions: *positional*, where heads attend to adjacent tokens; *syntactic*, where heads attend to specific syntactic relations; and *rare words*, where heads point to the least frequent tokens in the sentence. Htut et al. (2019) followed Raganato, Tiedemann, et al. (2018) in decoding dependency trees from BERT-based models, finding that fine-tuning on two classification tasks did not produce syntactically plausible attention patterns. Lastly, Limisiewicz et al. (2020) modified UD annotation to better represent attention patterns and introduced a supervised head-ensembling method for consolidating shared syntactic information across heads.

DOES ATTENTION HAVE EXPLANATORY VALUE? Though many studies have yielded insight about how attention behaves in a variety of models, the question of whether it can be seen as a "faithful" explanation of model predictions has been subject to much recent debate. For example, Jain and Wallace (2019) present compelling arguments that attention does not offer a faithful explanation of predictions. Primarily, they demonstrate that there is little correlation between standard feature importance measures and attention weights. Furthermore, they contend that there exist *counterfactual* attention distributions, which are substantially different from learned attention weights but that do not alter a model's predictions. Using a similar methodology, Serrano and Smith (2019) corroborate that attention does not provide an adequate account of an input component's importance.

In response to these findings, Wiegreffe and Pinter (2019) question the assumptions underlying such claims. Attention, they argue, is not a *primitive*, i.e., it cannot be detached from the rest of a model's components as is done in the experiments of Jain and Wallace (2019). They propose a set of four analyses to test whether a given model's at-

tention mechanism can provide meaningful explanation and demonstrate that the alternative attention distributions found via adversarial training methods do, in fact, perform poorly compared to standard attention mechanisms. On a theoretical level, they argue that, although attention weights do not give an *exclusive* "faithful" explanation, they do provide a meaningful *plausible* explanation.

This discussion is relevant to our study because it remains unclear whether or not attending to syntactic structure serves, in practice, as plausible explanation for model behavior, or whether or not it is even capable of serving as such. Indeed, the studies of Raganato, Tiedemann, et al. (2018) and Clark et al. (2019c) relate a convincing but incomplete picture — tree decoding accuracy just marginally exceeds baselines and various relations tend to be tracked across varying heads and layers. Thus, our fine-tuning experiments (detailed in the following section) serve to enable an "easy" setting wherein we explicitly inform our models of the same structure that we are trying to extract. We posit that, if, after fine-tuning, syntactic structures were still *not* decodable from the attention weights, one could safely conclude that these structures are being stored via a non-transparent mechanism that may not even involve attention weights. Such an insight would allow us to conclude that attention weights cannot provide even a plausible explanation for models relying on syntax.

6.4 EXPERIMENTAL DESIGN

To examine the extent to which we can decode dependency trees from attention patterns, we run a tree decoding algorithm over mBERT's attention heads — before and after fine-tuning via a parsing objective. We surmise that doing so will enable us to determine if attention can be interpreted as a reliable mechanism for capturing linguistic structure.

6.4.1 *Model*

We employ mBERT[1] in our experiments, which has been shown to perform well across a variety of NLP tasks (Hu et al., 2020b; Kondratyuk and Straka, 2019a) and capture aspects of syntactic structure cross-lingually (Chi et al., 2020; Pires et al., 2019). mBERT features 12 layers with 768 hidden units and 12 attention heads, with a joint WordPiece sub-word vocabulary across languages. The model was trained on the concatenation of WikiDumps for the top 104 languages with the largest Wikipedias,where principled sampling was employed to enforce a balance between high- and low-resource languages.

1 https://github.com/google-research/bert

6.4.2 *Decoding Algorithm*

For decoding dependency trees, we follow Raganato, Tiedemann, et al. (2018) in applying the Chu-Liu-Edmonds maximum spanning tree algorithm (Chu, 1965) to every layer/head combination available in mBERT ($12 \times 12 = 144$ in total). In order for the matrices to correspond to gold treebank tokenization, we remove the cells corresponding to the BERT delimiter tokens (`[CLS]` and `[SEP]`). In addition to this, we sum the columns and average the rows corresponding to the constituent subwords of gold tokens, respectively (Clark et al., 2019c). Lastly, since attention patterns across heads may differ in whether they represent heads attending to their dependents or vice versa, we take our input to be the element-wise product of a given attention matrix and its transpose ($A \circ A^\top$). We liken this to the joint probability of a head attending to its dependent and a dependent attending to its head, similarly to Limisiewicz et al. (2020). Per this point, we also follow Htut et al. (2019) in evaluating the decoded trees via Undirected Unlabeled Attachment Score (UUAS) — the percentage of undirected edges recovered correctly. Since we discount directionality, this is effectively a less strict measure than UAS, but one that has a long tradition in unsupervised dependency parsing since Klein and Manning (2004).

6.4.3 *Data*

For our data, we employ the Parallel Universal Dependencies (PUD) treebanks, as collected in UD v2.4 (Nivre et al., 2019). PUD was first released as part of the CONLL 2017 shared task (Zeman et al., 2018), containing 1000 parallel sentences, which were (professionally) translated from English, German, French, Italian, and Spanish to 14 other languages. The sentences are taken from two domains, **news** and **wikipedia**, the latter implying some overlap with mBERT's training data (though we did not investigate this). We include all PUD treebanks except Thai.[2]

6.4.4 *Fine-Tuning Details*

In addition to exploring pretrained mBERT's attention weights, we are also interested in how attention might be guided by a training objective that learns the exact tree structure we aim to decode. To this end, we employ the graph-based decoding algorithm of the biaffine parser introduced by Dozat and Manning (2016). We replace the standard BiLSTM encoder for this parser with the entire mBERT network, which we fine-tune with the parsing loss. The full parser decoder

2 Thai is the only treebank that does not have a non-PUD treebank available in UD, which we need for our fine-tuning experiments.

	ar	cs	de	en	es	fi	fr	hi	id	it	ja	ko	pl	pt	ru	sv	tr	zh
BASELINE	50	40	36	36	40	42	40	46	47	40	43	55	45	41	42	39	52	41
PRE	53 7-6	53 10-8	49 10-8	47 10-8	50 9-5	48 10-8	41 2-3	48 2-3	50 9-5	41 6-4	**45** 2-3	64 9-2	52 10-8	50 9-5	51 10-8	51 10-8	55 3-8	42 2-3
NONE	**76** 11-10	**78** 11-10	**76** 11-10	**71** 10-11	**77** 10-11	**66** 10-11	**45** 11-10	**72** 11-10	**75** 11-10	**58** 11-10	42 11-10	64 11-10	**75** 11-10	**76** 11-10	**75** 10-8	**74** 10-8	55 3-8	38 2-3
KEY	62 10-8	64 10-8	58 11-12	53 10-8	59 11-12	56 10-8	41 7-12	54 10-8	59 10-8	47 9-2	44 2-3	62 10-8	64 10-8	58 11-12	61 10-8	59 12-10	55 3-12	41 2-3
QUERY	69 11-4	74 10-8	70 11-4	66 11-4	73 11-4	63 10-8	42 11-4	62 11-4	67 11-4	54 11-4	**45** 2-3	65 10-8	72 11-4	70 11-4	70 10-8	68 11-4	56 10-8	42 2-3
KQ	71 11-4	76 11-4	70 11-4	65 11-4	74 11-4	62 11-4	43 10-11	64 11-4	69 11-4	55 11-4	44 2-3	64 11-4	73 11-4	73 11-4	69 11-4	69 11-4	55 11-4	41 2-3
VALUE	75 12-5	72 12-5	72 12-5	64 12-5	76 12-5	59 12-5	**45** 12-5	63 12-5	73 12-5	55 12-5	**45** 2-3	**66** 10-8	73 12-5	74 12-5	69 12-5	65 12-5	**57** 12-5	**42** 3-8
DENSE	68 11-10	71 11-10	65 11-10	60 10-8	67 12-10	61 11-10	42 10-8	65 11-10	66 11-10	49 9-5	44 3-12	64 11-10	70 11-10	64 12-5	67 11-10	64 11-10	55 11-10	40 3-12

Table 6.1: Adjacent-branching baseline and maximum UUAS decoding accuracy per PUD treebank, expressed as best score and best layer/-head combination for UUAS decoding. PRE refers to basic mBERT model before fine-tuning, while all cells below correspond different fine-tuned models described in Section 3.4. Best score indicated in **bold**.

consists of four dense layers, two for head/child representations for dependency arcs (dim. 500) and two for head/child representations for dependency labels (dim. 100). These are transformed into the label space via a bilinear transform.

After training the parser, we can decode the fine-tuned mBERT parameters in the same fashion as described in Section 6.4.2. We surmise that, if attention heads are capable of tracking hierarchical relations between words in any capacity, it is precisely in this setting that this ability would be attested. In addition to this, we are interested in what individual *components* of the mBERT network are capable of steering attention patterns towards syntactic structure. We believe that addressing this question will help us not only in interpreting decisions made by BERT-based neural parsers, but also in aiding us developing syntax-aware models in general (Strubell et al., 2018a; Swayamdipta et al., 2018). As such — beyond fine-tuning all parameters of the mBERT network (our basic setting) — we perform a series of ablation experiments wherein we update only one set of parameters per training cycle, e.g. the Query weights W_i^Q, and leave everything else frozen. This gives us a set of 6 models, which are described below. For each model, all non-BERT parser components are always left unfrozen.

- KEY: only the K components of the transformer are unfrozen; these are the representations of tokens that are paying attention *to* other tokens.
- QUERY: only the Q components are unfrozen; these, conversely, are the representations of tokens being paid attention to.
- KQ: both keys and queries are unfrozen.

- VALUE: semantic value vectors per token (V) are unfrozen; they are composed after being weighted with attention scores obtained from the K/Q matrices.
- DENSE: the dense feed-forward networks in the attention mechanism; all three per layer are unfrozen.
- NONE: The basic setting with nothing frozen; all parameters are updated with the parsing loss.

We fine-tune each of these models on a concatentation of all PUD treebanks for 20 epochs, which effectively makes our model multilingual. We do so in order to 1) control for domain and annotation confounds, since all PUD sentences are parallel and are natively annotated (unlike converted UD treebanks, for instance); 2) increase the number of training samples for fine-tuning, as each PUD treebank features only 1000 sentences; and 3) induce a better parser through multilinguality, as in Kondratyuk and Straka (2019b). Furthermore, in order to gauge the overall performance of our parser across all ablated settings, we evaluate on the test set of the largest non-PUD treebank available for each language, since PUD only features test partitions. When training, we employ a combined dense/sparse Adam optimiser, at a learning rate of $3 * 10^{-5}$. We rescale gradients to have a maximum norm of 5.

6.5 DECODING MBERT ATTENTION

The second row of Table 6.1 (PRE) depicts the UUAS after running our decoding algorithm over mBERT attention matrices, per language. We see a familiar pattern to that in Clark et al. (2019c) among others — namely that attention patterns extracted directly from mBERT appear to be incapable of decoding dependency trees beyond a threshold of 50–60% UUAS accuracy. However, we also note that, in all languages, the attention-decoding algorithm outperforms a BASELINE (row 1) that draws an (undirected) edge between any two adjacent words in linear order, which implies that some non-linear structures are captured with regularity. Indeed, head 8 in layer 10 appears to be particularly strong in this regard, returning the highest UUAS for 7 languages. Interestingly, the accuracy patterns across layers depicted in Figure 6.1 tend to follow an identical trend for all languages, with nearly all heads in layer 7 returning high within-language accuracies.

It appears that attention for some languages (Arabic, Czech, Korean, Turkish) is comparatively easier to decode than others (French, Italian, Japanese, Chinese). A possible explanation for this result is that dependency relations between content words, which are favored by the UD annotation, are more likely to be adjacent in the morphologically rich languages of the first group (without intervening function words). This assumption seems to be corroborated by the

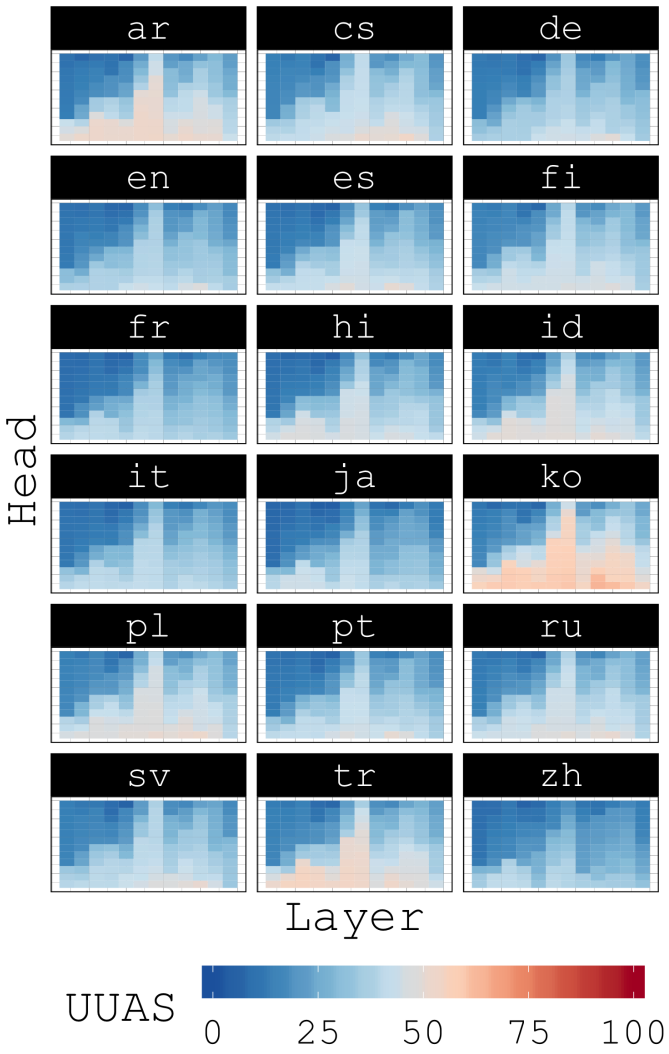


Figure 6.1: UUAS of MST decoding per layer and head, across languages. Heads (y-axis) are sorted by accuracy for easier visualization.

high baseline scores for Arabic, Korean and Turkish (but not Czech). Conversely, the low baselines scores and the likewise low decoding accuracies for the latter four languages are difficult to characterize. Indeed, we could not identify what factors — typological, annotation, tokenization or otherwise — would set French and Italian apart from the remaining languages in terms of score. However, we hypothesize that the tokenization and our treatment of subword tokens plays a part in attempting to decode attention from Chinese and Japanese representations. Per the mBERT documentation,[3] Chinese and Japanese Kanji character spans within the CJK Unicode range are character-tokenized. This lies in contrast with all other languages (Korean Hangul and Japanese Hiragana and Katakana included), which

3 https://github.com/google-research/bert/blob/master/multilingual.md

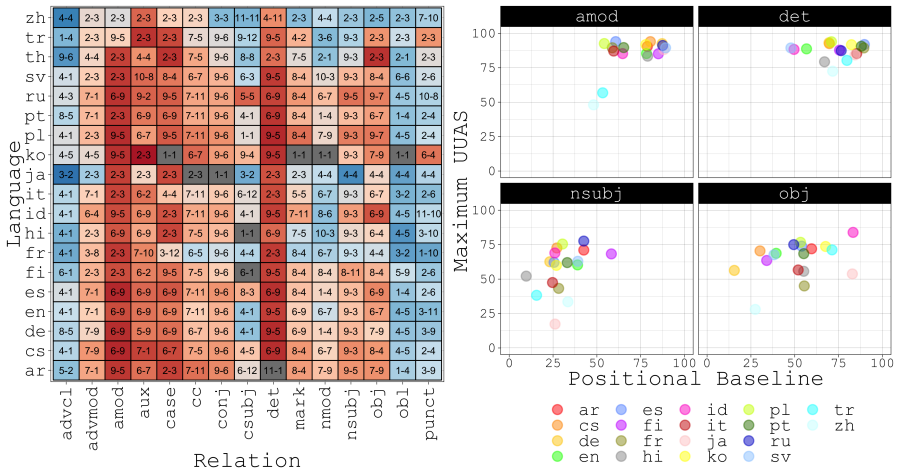


Figure 6.2: Left: UUAS per relation across languages (best layer/head combination indicated in cell). Right: Best UUAS as a function of best positional baseline (derived from the treebank), selected relations.

rely on whitespace and WordPiece (Wu et al., 2016). It is thus possible that the attention distributions for these two languages (at least where CJK characters are relevant) are devoted to composing words, rather than structural relations, which will distort the attention matrices that we compute to correspond with gold tokenization (e.g. by maxing rows and averaging columns).

RELATION ANALYSIS We can disambiguate what sort of structures are captured with regularity by looking at the UUAS returned per dependency relation. Figure 6.2 (left) shows that adjectival modifiers (`amod`, mean UUAS = 85 $\pm$12) and determiners (`det`, 88 $\pm$ 6) are among the easiest relations to decode across languages. Indeed, words that are connected by these relations are often adjacent to each other and may be simple to decode if a head is primarily concerned with tracking linear order. To verify the extent to which this might be happening, we plot the aforementioned decoding accuracy as a function of select relations' positional baselines in Figure 6.2 (right). The positional baselines, in this case, are calculated by picking the most frequent offset at which a dependent occurs with respect to its head, e.g., -1 for `det` in English, meaning one position to the left of the head. Interestingly, while we observe significant variation across the positional baselines for `amod` and `det`, the decoding accuracy remains quite high.

In slight contrast to this, the core subject (`nsubj`, 58 $\pm$ 16 SD) and object (`obj`, 64 $\pm$ 13) relations prove to be more difficult to decode. Unlike the aforementioned relations, `nsubj` and `obj` are much more sensitive to the word order properties of the language at hand. For example, while a language like English, with Subject-Verb-Object (SVO)

order, might have the subject frequently appear to the left of the verb, an SOV language like Hindi might have it several positions further away, with an object and its potential modifiers intervening. Indeed, the best positional baseline for English `nsubj` is 39 UUAS, while it is only 10 for Hindi. Despite this variation, the relation seems to be tracked with some regularity by the same head (layer 3, head 9), returning 60 UUAS for English and 52 for Hindi. The same can largely be said for `obj`, where the positional baselines return 51 ± 18. In this latter case, however, the heads tend to be much differently distributed across languages. Finally, he results for the `obj` relation provides some support for our earlier explanation concerning morphologically rich languages, as Arabic, Czech, Korean and Turkish all have among the highest accuracies (as well as positional baselines).

6.6 FINE-TUNING EXPERIMENTS

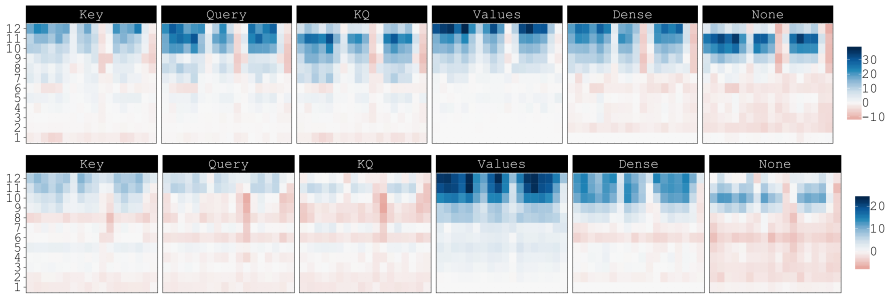


Figure 6.3: (Top) best scores across all heads, per language; (bottom) mean scores across all heads, per language. The languages (hidden from the X-axis for brevity) are, in order, *ar, cs, de, en, es, fi, fr, hi, id, it, ja, ko, pl, pt, ru, sv, tr, zh*

Next, we investigate the effect fine-tuning has on UUAS decoding. Row 3 in Table 6.1 (NONE) indicates that fine-tuning does result in large improvements to UUAS decoding across most languages, often by margins as high as ~ 30%. This shows that with an explicit parsing objective, attention heads are capable of serving as explanatory mechanisms for syntax; syntactic structure can be made to be transparently stored in the heads, in a manner that does not require additional probe fitting or parameterized transformation to extract.

Given that we do manage to decode reasonable syntactic trees, we can then refine our question — what components are capable of learning these trees? One obvious candidate is the key/query component pair, given that attention weights are a scaled softmax of a composition of the two. Figure 6.3 (top) shows the difference between pretrained UUAS and fine-tuned UUAS per layer, across models and languages. Interestingly, the best parsing accuracies do not appear to vary much depending on what component is frozen. We do see a clear trend, however, in that decoding the attention patterns of the fine-

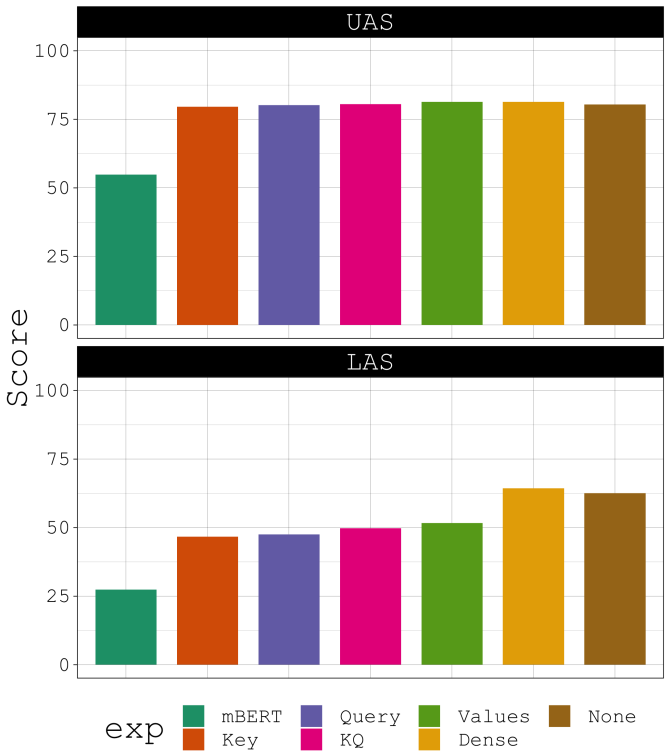


Figure 6.4: Mean UAS and LAS when evaluating different models on language-specific treebanks (Korean excluded due to annotation differences). MBERT refers to models where the entire mBERT network is frozen as input to the parser.

tuned model typically yields better UUAS than the pretrained model, particularly in the highest layers. Indeed, the lowest layer at which fine-tuning appears to improve decoding is layer 7. This implies that, regardless of which component remains frozen, the parameters facing any sort of significant and positive update tend to be those appearing towards the higher-end of the network, closer to the output.

For the frozen components, the best improvements in UUAS are seen at the final layer in VALUE, which is also the only model that shows consistent improvement, as well as the highest average improvement in mean scores[4] for the last few layers. Perhaps most interestingly, the mean UUAS (Figure 6.3 (bottom)) for our "attentive" components – keys, queries, and their combination – does not appear to have improved by much after fine-tuning. In contrast, the maximum does show considerable improvement; this seems to imply that although all components appear to be more or less equally capable of learning decodable heads, the attentive components, when fine-tuned, appear to sharpen fewer heads.

Note that the only difference between keys and queries in an attention mechanism is that keys are transposed to index attention from/to

4 The inner average is over all heads; the outer is over all languages.

appropriately. Surprisingly, KEY and QUERY appear to act somewhat differently, with QUERY being almost uniformly better than KEY with the best heads, whilst KEY is slightly better with averages, implying distinctions in how both store information. Furthermore, allowing both keys and queries seems to result in an interesting contradiction – the ultimate layer, which has reasonable maximums and averages for both KEY and QUERY, now seems to show a UUAS drop almost uniformly. This is also true for the completely unfrozen encoder.

SUPERVISED PARSING In addition to decoding trees from attention matrices, we also measure supervised UAS/LAS on a held-out test set.[5] Based on Figure 6.4, it is apparent that all settings result in generally the same UAS. This is somewhat expected; Lauscher et al. (2020) see better results on parsing with the entire encoder frozen, implying that the task is easy enough for a biaffine parser to learn, given frozen mBERT representations.[6] The LAS distinction is, however, rather interesting: there is a marked difference between how important the dense layers are, as opposed to the attentive components. This is likely not reflected in our UUAS probe as, strictly speaking, labelling arcs is not equivalent to searching for structure in sentences, but more akin to classifying pre-identified structures. We also note that DENSE appears to be better than NONE on average, implying that non-dense components might actually be hurting labelling capacity.

In brief, consolidating the two sets of results above, we can draw three interesting conclusions about the components:

1. **Value** vectors are best aligned with syntactic dependencies; this is reflected both in the best head at the upper layers, and the average score across all heads.
2. **Dense** layers appear to have moderate informative capacity, but appear to have the best learning capacity for the task of arc labelling.
3. Perhaps most surprisingly, **Key** and **Query** vectors do not appear to make any outstanding contributions, save for sharpening a smaller subset of heads.

Our last result is especially surprising for UUAS decoding. Keys and queries, fundamentally, combine to form the attention weight matrix, which is precisely what we use to decode trees. One would expect that allowing these components to learn from labelled syntax would result in the best improvements to decoding, but all three have surprisingly negligible mean improvements. This indicates that we need to further improve our understanding of how attentive structure and weighting really works.

5 Note that the test set in our scenario is from the actual, non-parallel language treebank; as such, we left Korean out of this comparison due to annotation differences.

6 Due to training on concatenated PUD sets, however, our results are not directly comparable/

CROSS-LINGUISTIC OBSERVATIONS We notice no clear cross-linguistic trends here across different component sets; however, certain languages do stand out as being particularly hard to decode from the fine-tuned parser. These include Japanese, Korean, Chinese, French and Turkish. For the first three, we hypothesise that tokenization clashes with mBERT's internal representations may play a role. Indeed, as we hypothesized in Section 6.4.2, it could be the case that the composition of CJK characters into gold tokens for Chinese and Japanese may degrade the representations (and their corresponding attention) therein. Furthermore, for Japanese and Korean specifically, it has been observed that tokenization strategies employed by different treebanks could drastically influence the conclusions one may draw about their inherent hierarchical structure (Kulmizev et al., 2020). Turkish and French are admittedly more difficult to diagnose. Note, however, that we fine-tuned our model on a concatenation of all PUD treebanks. As such, any deviation from PUD's annotation norms is therefore likely to be heavily penalised, by virtue of signal from other languages drowning out these differences.

6.7 CONCLUSION

In this study, we revisited the prospect of decoding dependency trees from the self-attention patterns of Transformer-based language models. We elected to extend our experiments to 18 languages in order to gain better insight about how tree decoding accuracy might be affected in the face of (modest) typological diversity. Surprisingly, across all languages, we were able to decode dependency trees from attention patterns more accurately than an adjacent-linking baseline, implying that some structure was indeed being tracked by the mechanism. In looking at specific relation types, we corroborated previous studies in showing that particular layer-head combinations tracked the same relation with regularity across languages, despite typological differences concerning word order, etc.

In investigating the extent to which attention can be guided to properly capture structural relations between input words, we fine-tuned mBERT as input to a dependency parser. This, we found, yielded large improvements over the pretrained attention patterns in terms of decoding accuracy, demonstrating that the attention mechanism was learning to represent the structural objective of the parser. In addition to fine-tuning the entire mBERT network, we conducted a series of experiments, wherein we updated only select components of model and left the remainder frozen. Most surprisingly, we observed that the Transformer parameters designed for composing the attention matrix, K and Q, were only modestly capable of guiding the attention towards resembling the dependency structure. In contrast, it was the Value (V) parameters, which are used for computing a

weighted sum over the KQ-produced attention, that yielded the most faithful representations of the linguistic structure via attention.

Though prior work (Kovaleva et al., 2019; Zhao and Bethard, 2020) seems to indicate that there is a lack of a substantial change in attention patterns after fine-tuning on syntax- and semantics-oriented classification tasks, the opposite effect has been observed with fine-tuning on negation scope resolution, where a more explanatory attention mechanism can be induced (Htut et al., 2019). Our results are similar to the latter, and we demonstrate that given explicit syntactic annotation, attention weights do end up storing more transparently decodable structure. It is, however, still unclear which sets of transformer parameters are best suited for learning this information and storing it in the form of attention.

7

WORD ORDER DOES MATTER AND SHUFFLED LANGUAGE MODELS KNOW IT

7.1 ABSTRACT

Recent studies have shown that language models pretrained and/or fine-tuned on randomly permuted sentences exhibit competitive performance on GLUE, putting into question the importance of word order information. Somewhat counter-intuitively, some of these studies also report that position embeddings appear to be crucial for models' good performance with shuffled text. We probe these language models for word order information and investigate what position embeddings learnt from shuffled text encode, showing that these models retain considerable information about the original token order. We show this is due to a subtlety in how shuffling is implemented in previous work – *before* rather than *after* subword segmentation. Language models trained on text shuffled *after* subword segmentation exhibit much lower performance, but even these models retain *some* information about word order because of the statistical dependencies between sentence length and unigram probabilities. Finally, we show that beyond GLUE, a variety of language understanding tasks *do* require word order information, often to an extent that cannot be learnt through fine-tuning.

7.2 INTRODUCTION

Transformers (Vaswani et al., 2017), when used in the context of masked language modelling (Devlin et al., 2019b), consume their inputs concurrently. There is no notion of inherent order, unlike in autoregressive setups, where the input is consumed token by token. To compensate for this absence of linear order, the transformer architecture originally proposed in Vaswani et al. (2017) includes a fixed, sinusoidal position embedding added to each token embedding; each token carries a different position embedding, corresponding to its position in the sentence. The transformer-based BERT (Devlin et al., 2019b) replaces these fixed sinusoidal embeddings with unique, learned embeddings per position; RoBERTa (Liu et al., 2019c), the model investigated in this work, does the same.

Position embeddings are the only source of order information in these models; in their absence, contextual representations generated for tokens are independent of the actual position of the tokens in a sentence, and the models thus resemble heavily overparameterised

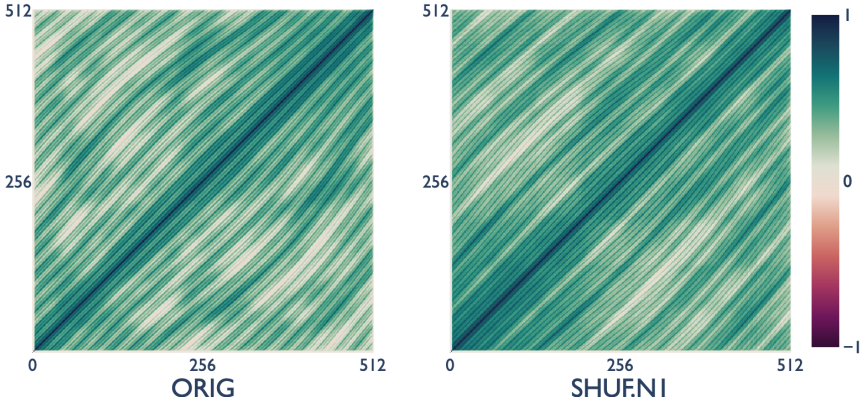


Figure 7.1: Pearson correlations between position embeddings for full-scale models; the patterns are similar to fully learnable absolute embeddings (Wang et al., 2021) and can be said to have learned something about position. We later demonstrate that this is not the case with post-BPE scrambling.

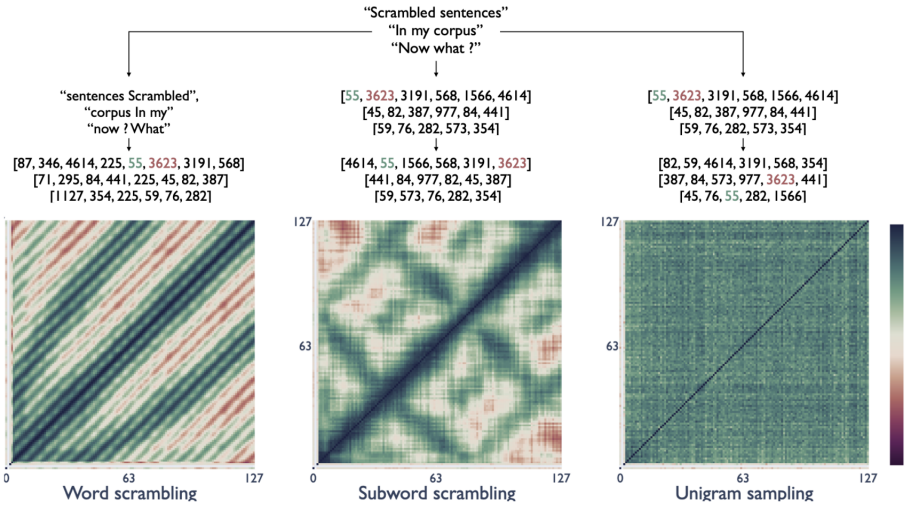


Figure 7.2: Correlations between position embeddings when shuffling training data *before* segmentation (left), i.e, at the word level, and *after* segmentation (middle), i.e., at the subword level, as well as when replacing all subwords with random subwords based on their corpus-level frequencies (right). The latter removes any dependency between subword probability and sentence length. The plots show that shuffling before segmentation retains more order information than shuffling after, and that even when shuffling after segmentation, position embeddings are meaningful because of the dependence between subword probability and sentence length.

bags-of-words. Sinha et al. (2021) pre-trained RoBERTa models on shuffled corpora to demonstrate that the performance gap between these 'shuffled' language models and models trained on unshuffled corpora is minor (when fine-tuned and evaluated downstream on the GLUE (Wang et al., 2018b) benchmark). They further show that this gap is considerably wider when a model is pre-trained without position embeddings. In this paper, we attempt to shed some light on why these models behave the way they do, and in doing so, seek to answer a set of pertinent questions:

- Do shuffled language models still have traces of word order information?
- Why is there a gap in performance between models *without* position embeddings and models trained on shuffled tokens, with the latter performing better?
- Are there NLU benchmarks, other than GLUE, on which shuffled language models perform poorly?

CONTRIBUTIONS We first demonstrate, in Section 7.4, that shuffled language models *do* contain word order information, and are quite responsive to simple tests for word order information, particularly when compared to models trained without position representations. In Section 7.5, we demonstrate that pre-training is sufficient to learn this: position embeddings provide the appropriate inductive bias, and performing BPE segmentation after shuffling results in sensible n-grams appearing in the pre-training corpus; this gives models the capacity to learn word order within smaller local windows. Other minor cues - like correlations between sentence lengths and token distributions - also play a role. We further corroborate our analysis by examining attention patterns across models in Sec. 7.6. In Section 7.7, we show that, while shuffled models might be almost as good as their un-shuffled counterparts on GLUE tasks, there exist NLU benchmarks that do require word order information to an extent that cannot be learned through fine-tuning alone. Finally, in Section 7.8, we describe miscellaneous experiments addressing the utility of positional embeddings when added just prior to fine-tuning.

7.3 MODELS

Sinha et al. (2021) train several full-scale RoBERTa language models on the Toronto Book Corpus (Zhu et al., 2015) and English Wikipedia.[1] Four of their models are trained on shuffled text, i.e., sentences in which n-grams are reordered at random.[2] We dub the original, un-

1 Training reportedly takes 72 hours on 64 GPUs.

2 The shuffling procedure does not reorder tokens *completely* at random, but moves a token in position i to a *new* position selected at random among positions $j \neq i$.

perturbed model ORIG, and the scrambled models SHUF.N1, SHUF.N2, SHUF.N3 and SHUF.N4 depending on the size of the shuffled n-grams: SHUF.N1 reorders the unigrams in a sentence, SHUF.N2 reorders it bigrams, etc. See Sinha et al. (2021) for more details. For comparison, Sinha et al. (2021) also train a RoBERTa language model entirely *without* position embeddings (NOPOS), as well as a RoBERTa language model trained on a corpus drawn solely from unigram distributions of the original Book Corpus, i.e., a reshuffling of the entire corpus (SHUF.CORPUS). We experiment with their models, as well as with smaller models that we can train with a smaller carbon footprint. To this end, we downscale the RoBERTa architecture used in Sinha et al. (2021). Concretely, we train single-headed RoBERTa models, dividing the embedding and feed-forward dimensionality by 12, for 24 hours on a single GPU, on 100k sentences sampled from the Toronto Book Corpus; we trained a custom vocabulary of size 5,000 that we use for indexing in all our subsequent experiments. While these smaller models are in no way meant to be fine-tuned and used downstream, they are useful proofs-of-concept that we later analyse.

7.4 PROBING FOR WORD ORDER

We begin by attempting to ascertain the extent to which shuffled language models are actually capable of encoding word order information. We perform two simple tests on the full-scale models, in line with Wang and Chen (2020): the first of these is a classification task where a logistic regressor is trained to predict whether a randomly sampled token precedes another in a sentence, and the second involves predicting the position of a word in a sentence. The fact that we *do not* fine-tune any of the model parameters is noteworthy; the linear models can only learn word order information if it reflects in the representations the models generate somehow.

PAIRWISE CLASSIFICATION For this experiment, we train a logistic regression classification model given word representations at the final layer of the Transformer encoder, mean pooling over sub-tokens when required. For each word pair x and y, the classifier is given a concatenation of our model m's induced representations $m(x) \oplus m(y)$ and trained to predict a label indicating whether x precedes y or not. Holding out two randomly sampled positions, we use a training sets sized 2k , 5k and 10k, from the Universal Dependencies English-GUM corpus (excluding sentences with more than 30 tokens to increase learnability) and a test set of size $2,000$. We report the mean accuracy from three runs.

REGRESSION Using the same data, we also train a ridge-regularised linear regression model to predict the position of a word $p(x)$ in a

Model	Classification (acc.)			Regression (R^2)
	2k	5k	10k	-
ORIG	81.50	81.74	80.40	0.68
SHUF.N1	65.96	64.98	71.82	0.60
NOPOS	50.41	53.35	50.22	0.03

Table 7.1: Pairwise classification and regression results.

sentence, given that word's model-induced representation $m(x)$. R^2 score is reported per model. To prevent the regressors from memorising word to position mappings, we perform 6-fold cross-validation, where the heldout part of the data contains no vocabulary overlap with the corresponding train set.

RESULTS For both tasks (see Table 7.1), our results indicate that position encodings are particularly important for encoding word order: Classifiers and regressors trained on representations from ORIG and SHUF.N1 achieve high accuracies and R^2 scores, while those for NOPOS are close to random. Both ORIG and SHUF.N1 appear to be better than random given only 2k examples. These results imply that, given positional encodings and a modest training set of 10k examples, a simple linear model is capable of extracting word order information, enabling almost perfect extrapolation to unseen positions. Whether the position encodings come from a model trained on natural or shuffled text does not appear to matter, emphasizing that shuffled language models do indeed contain substantial information about the original word order.

7.5 HIDDEN WORD-ORDER SIGNALS

In §3, we saw that the shuffled language models of Sinha et al. (2021) surprisingly exhibit information about pre-shuffling word order. That these models contain positional information can also be seen by visualizing position embedding similarity. Figure 7.1 displays Pearson correlations[3] for position embeddings with themselves, across positions; the shuffled models satisfy the idealised criteria for position embeddings described by Wang et al. (2021), in that they appear to be a) monotonous within smaller context windows, and b) invariant to translation. If position embedding correlations are consistent across offsets over the entire space of embeddings, the model has 'learnt' distances between tokens. Since transformers process all positions in parallel, and since language models without position embeddings do not exhibit such information, position embeddings have to be the

3 We see similar patterns with dot products for all our plots; we use Pearson correlations to constrain our range to $[-1, 1]$.

source of this information. In this section, we discuss this apparent paradox.

SUBWORD VS. WORD SHUFFLING An important detail when running experiments on shuffled text, is *when* the shuffling operation takes place. When tokens are shuffled *before* BPE segmentation, this leads to word-level shuffling, in which sequences of subwords that form words remain contiguous. Such sequences become a consistent, meaningful signal for language modelling, allowing models to efficiently utilise the inductive bias provided by position embeddings. Thus, even though our pretrained models have, in theory, not seen consecutive tokens in their pre-training data, they have learnt to utilise position to pay attention to adjacent tokens. The influence of this is already somewhat visible in Figure 7.2; while both models trained on text shuffled before segmentation, and models trained on text shuffled after segmentation, have shifts in the *polarity* of their position correlations, only the plots for models trained with word-level shuffling (*before* segmentation) have bands of varying *magnitude*, similar to the full-scale models. Such bands enable position-invariant compositionality (Ravishankar and Søgaard, 2021).

In Section 7.6, we analyse the effect of shuffling the pre-training data on the models' attention mechanisms.

ACCIDENTAL OVERLAP In addition to the n-gram information which results from shuffling before segmentation, we also note that short sentences tend to include original bigrams with high probability, leading to stronger associations for words that are adjacent in the original texts. This effect is obviously much stronger when shuffling before segmentation than after segmentation.

Figure 7.3 shows how frequent overlapping bigrams (of any sort) are, comparing word and subword shuffling over 50k sentences.

SENTENCE LENGTH Finally, we observe some preserved information about the original word order even when shuffling is performed *after* segmentation. We hypothesize that this is a side-effect of the non-random relationship between sentence length and unigram probabilities. That unigram probabilities correlate with sentence length follows from the fact that different genres exhibit different sentence length distributions (Jin and Liu, 2017; Sigurd et al., 2004). Also, some words occur very frequently in formulaic contexts, e.g., *thank* in *thank you*. This potentially means that there is an approximately learnable relationship between the distribution of words and sentence boundary symbols.

To test for this, we train two smaller language models on unigram-sampled corpora: for the first, we use the first 100k BookCorpus sentences as our corpus, shuffling tokens at a corpus level (yet keeping

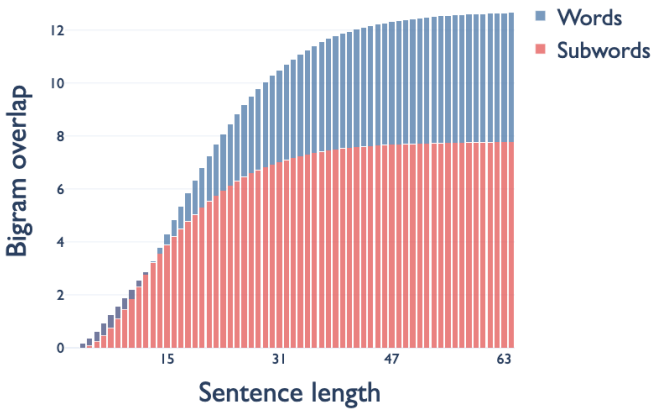


Figure 7.3: (Cumulative) plot showing bigram overlap after shuffling either words or subwords, as a percentage of the total number of seen bigrams. We see the overlap is significant, especially when performing shuffling before segmentation.

the original sentence lengths). The stark difference in position embedding correlations between that and shuffling is seen in Figure 7.2. For the second, we sample from two different unigram distributions: one for short sentences and one for longer sentences (details in Appendix A.5.2). While the first model induces no correlations at all, the second does, as shown in Figure 7.4, implying that sentence length and unigram occurrences is enough to learn *some* order information.

7.6 ATTENTION ANALYSIS

Transformer-based language models commonly have attention heads that attend to neighboring positions (Ravishankar et al., 2021; Voita et al., 2019a). Such attention heads are positional and can only be learnt in the presence of order information. We attempt to visualise the attention mechanism for pre-trained models by calculating, for each head and layer, the offset between a token, and the token that it pays maximimum attention to[4]. We then plot how frequent each offset is, as a percentage, over 100 Book Corpus sentences, in Figure 7.5, where we present results for two full-scale models, and two smaller models (see §2). When compared to NOPOS, SHUF.N1 has a less uniform pattern to its attention mechanism: it is likely, even at layer 0, to prefer to pay attention to adjacent tokens, somewhat mimicking a convolutional window (Cordonnier et al., 2020). We see very similar differences in distribution between our smaller models: Shuffling af-

4 This method of visualisation is somewhat limited, in that it examines only the *maximum* attention paid by each token. We provide more detailed plots over attention *distributions* in the Appendix.

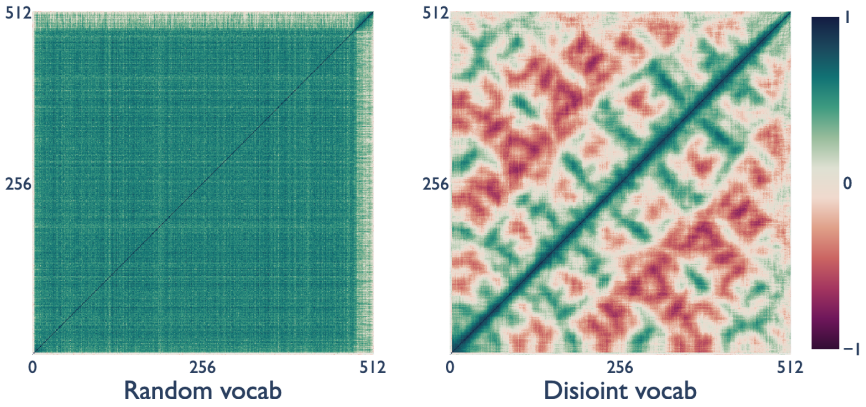


Figure 7.4: Similarity matrix between models with sentences sampled based on unigram corpus statistics; disjoint vocab implies a correlation between token choice and sentence length.

ter segmentation, i.e., at the subword level, influences early attention patterns.

7.7 EVALUATION BEYOND GLUE

SUPERGLUE AND WINOGRANDE Sinha et al. (2021)'s investigation is conducted on GLUE and on the Paraphrase Adversaries from Word shuffling (PAWS) dataset (Zhang et al., 2019b). For these datasets, they find that models pretrained on shuffled text perform only marginally worse than those pretrained on normal text. This result, they argue can be explained in two ways: either a) these tasks do not need word order information to be solved, or b) the required word order information can be acquired during finetuning. While GLUE has been a useful benchmark, several of the tasks which constitute it have been shown to be solvable using various spurious artefacts and heuristics (Gururangan et al., 2018a; Poliak et al., 2018a). If, for instance, through finetuning, models are learning to rely on such heuristics as lexical overlap for MNLI (McCoy et al., 2019), then it is unsurprising that their performance is not greatly impacted by the lack of word order information.

Evaluating on the more rigorous set of SuperGLUE tasks[5] (Wang et al., 2019b) and on the adversarially-filtered Winograd Schema examples (Levesque et al., 2012) of the WinoGrande dataset (Sakaguchi et al., 2020) produces results which paint a more nuanced picture com-

5 Results are reported for an average of 3 runs per task. For SuperGLUE, the `jiant` framework is used for finetuning and evaluation (Wang et al., 2019c). For WinoGrande, the Fairseq implementation (Ott et al., 2019) is used. Default hyperparameters are used for both. The Recognizing Textual Entailment task is excluded from our results as it is also part of GLUE. Results for that can be found in Sinha et al. (2021).

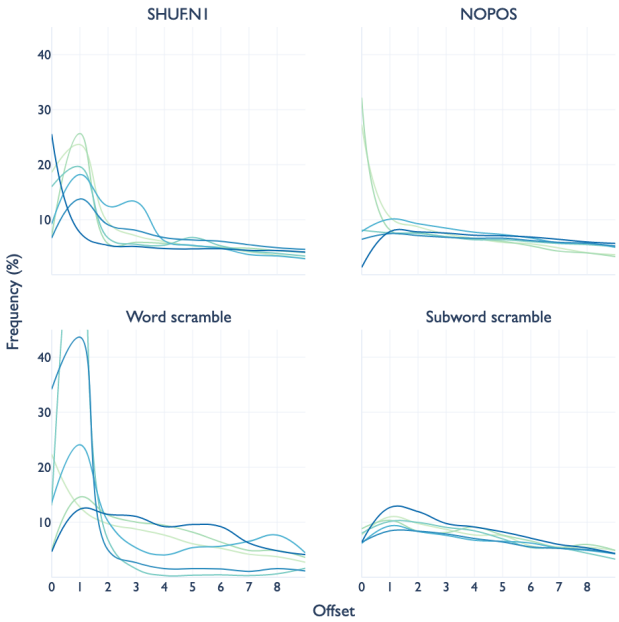


Figure 7.5: Relative frequency of offsets between token pairs in an attention relation; the y-axis denotes the percentage of total attention relations that occur at the offset indicated on the x-axis. We plot layers $l \in \{1, 2, 7, 8, 11, 12\}$ with increasing line darkness.

pared to those of Sinha et al. (2021). The results, presented in Table 7.2, show accuracy or F1 scores for all models. For two of the tasks (MultiRC (Khashabi et al., 2018), COPA (Roemmele et al., 2011)), we observe a pattern in line with that seen in Sinha et al. (2021)'s GLUE and PAWS results: the drop in performance from ORIG to SHUF.N1 is minimal (mean: 1.75 points; mean across GLUE tasks: 3.3 points)[6], while that to NOPOS is more substantial (mean: 10.5 points; mean across GLUE tasks: 18.6 points).

This pattern becomes stronger for the BoolQ Yes/No question answering dataset (Clark et al., 2019a), the CommitmentBank (De Marneffe et al., 2019), the ReCoRD reading comprehension dataset (Zhang et al., 2018), both the Winograd Schema tasks, and to some extent the Words in Context dataset (Pilehvar and Camacho-Collados, 2018). Here we observe a larger gap between ORIG and SHUF.N1 (mean: 8.1 points), and an even larger one between ORIG and NOPOS (mean: 19.78). We note that this latter set of tasks requires inferences which are more context-sensitive, in comparison to the two other tasks or to the GLUE tasks.

Consider the Winograd schema tasks, for example. Each instance takes the form of a binary test with a statement comprising of two

6 CoLA results are excluded from the GLUE calculations here due to the very high variance across random seeds reported by Sinha et al. (2021) which makes the results unreliable.

possible referents (blue) and a pronoun (red) such as: `Sid explained his theory to Mark but he couldn't convince him.` The correct referent of the pronoun must be inferred based on a special discriminatory segment (underlined). In the above example, this depends on a) the identification of "Sid" as the subject of "explained" and b) inferring that the pronoun serving as the subject of "convinced" should refer to the same entity. Since the Winograd schema examples are designed so that the referents are equally associated with their context[7], word order is crucial [8] for establishing the roles of "Sid" and "Mark" as subject and object of "explained" and "he" and "him" as those of "convinced". If these roles cannot be established, making the correct inference becomes impossible.

A similar reasoning can be applied to the the Words in Context dataset and the CommitmentBank. The former task tests the ability of a model to distinguish the senses of a polysemous word based on context. While this might often be feasible via a notion of contextual association that higher-order distributional statistics are sufficient for, some instances will require awareness of the word's role as an argument in the sentence. The latter task investigates the projectivity of finite clausal complements under entailment cancelling operators. This is sensitive to both the scope of the entailment operator and to the identity of the subject of the matrix predicate (De Marneffe et al., 2019), meaning.

A final consideration to take into account is dataset filtering. Two of the tasks where we observe the largest difference between Orig, Shuf.N1, and Nopos — WinoGrande and ReCoRD — apply filtering algorithms to remove cues or biases which would enable models to heuristically solve the tasks. This indicates that by filtering out examples containing cues that make them solvable via higher order statistics, such filtering strategies do succeed at compelling models to (at least partially) rely on word order information.

DEPENDENCY TREE PROBING Beyond GLUE and PAWS, Sinha et al. (2021)'s analysis also includes several probing experiments, wherein they attempt to decode dependency tree structure from model representations. They show, interestingly, that the Shuf.N4, Shuf.N3 and Shuf.N2 models perform only marginally worse than Orig, with Shuf.N1 producing the lowest scores (lower, in fact, than Shuf.Corpus). Given the findings of Section 7.4, we are interested in taking a closer look at this phenomenon. Here, we surmise that *dependency length* plays a crucial role in the probing setup, where permuted models may succeed on par with Orig in capturing local, adjacent dependen-

7 e.g. Sid and Mark are both equally likely subjects/objects here. Not all Winograd schema examples are perfect in this regard, however, which could explain why scrambled models still perform above random. See Trichelair et al. (2018) for a discussion of the latter point.

8 Particularly in a language with limited morphological role marking such as English.

Model	BoolQ	CB	COPA	MultiRC	ReCoRD	WiC	WSC	WinoGrande
ORIG	77.6	88.2 / 87.4	61.6	67.8 / 21.9	73.5 / 72.8	67.4	73.5	62.9
SHUF.N1	72.4	79.7 / 82.5	59.7	66.2 / 15.0	61.1 / 60.4	63.0	62.9	55.7
SHUF.N2	73.1	86.6 / 85.5	60.3	64.8 / 16.1	63.1 / 62.4	63.0	65.3	57.6
SHUF.N4	73.5	87.9 / 87.1	60.8	66.2 / 18.2	64.6 / 63.9	62.4	65.3	59.53
NOPOS	66.0	63.5 / 75.0	55.6	52.8 / 3.8	23.8 / 23.5	55.4	63.09	52.73
SHUF.CORPUS	66.7	65.6 / 73.8	56.1	52.6 / 6.4	31.0 / 30.3	57.3	65.14	51.68

Table 7.2: SuperGLUE and WinoGrande results for all models. Scores displayed are: Avg. F1 / Accuracy for CB; F1a / Exact Match for MultiRC; F1 / Accuracy for ReCoRD ; accuracy for the remaining tasks.

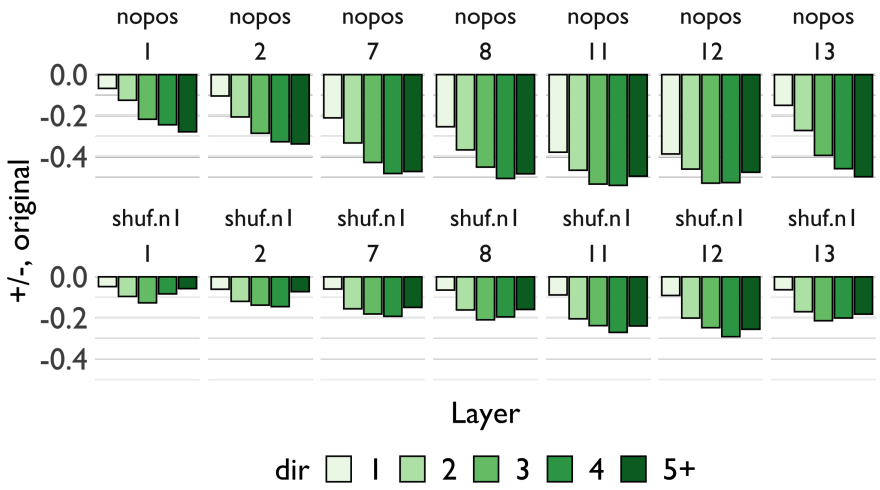


Figure 7.6: Δ, dependency arcs probing accuracy across lengths 1-5+, w.r.t. ORIG.

cies, but increasingly struggle to decode longer ones. To evaluate the extent to which this is true, we train a bilinear probe (used in Hewitt and Liang (2019a)) on top of all model representations and evaluate its accuracy across dependencies binned by length, where length between words w_i and w_j is defined as $|i - j|$. We opt for using the bilinear probe over the Pareto probing framework (Pimentel et al., 2020a), as the former learns a transformation directly over model representations, while the latter adds the parent and child MLP units from Dozat et al. (2017) – acting more like a parser. We train probes on the English Web Treebank (Silveira et al., 2014) and evaluate using UAS, the standard parsing metric.

Figure 7.6 shows Δ probing accuracy across various dependency lengths for NOPOS and SHUF.N1, with respect to ORIG[9]; we include detailed Δs for all models in Appendix A.5.3. For NOPOS, parsing

9 Note that Layer 13 refers to a linear mix of all model layers, as is done for ELMo (Peters et al., 2018a).

difficulty increases almost linearly with distance, often mimicking the actual frequency distribution of dependencies at these distances in the original treebank (Appendix A.5.3); for SHUF.N1, the picture is a lot more nuanced, with dependencies at a distance of 1 consistently being closer in terms of parseability to ORIG, which, we hypothesise, is due to its adjacency bias.

7.8 OTHER FINDINGS

RANDOM POSITION EMBEDDINGS ARE DIFFICULT TO ADD POST–TRAINING We tried to quantify the degree to which the inductive bias imparted by positional embeddings can be utilised, solely via fine-tuning. To do so, for a subset of GLUE tasks (MNLI, QNLI, RTE, SST-2, CoLA), we evaluate NOPOS, and a variant where we randomly initialised learnable position embeddings and add them to the model, with the rest of the model equivalent to NOPOS. We see no improvement in results, except for MNLI, that we hypothesise stems from position embeddings acting as some sort of regularisation parameter. To test this, we repeat the above set of experiments, this time adding random learnable Gaussian noise instead: this led to a slight increase in score for just MNLI, backing up this hypothesis.

MODELS LEARN TO EXPECT SPECIFIC EMBEDDINGS Replacing the positional embeddings in ORIG with fixed, sinusoidal embeddings before fine-tuning significantly hurts scores on the same subset of GLUE tasks, implying that the models expect embeddings that resemble the inductive bias imparted by random embeddings, and that fine-tuning tasks do not have sufficient data to overcome this. The addition of fixed, *sinusoidal* to NOPOS also does not improve model performance on a similar subset of tasks; this implies, given that sinusoidal embeddings are already meaningful, that model weights also need to learn to fit the embeddings they are given, and that they need a substantial amount of data to do so.

7.9 ON WORD ORDER

IN HUMANS It is generally accepted that a majority of languages have "canonical" or "base' word orderings (Comrie, 1989) (e.g. Subject-Verb-Object in English, and Subject-Object-Verb in Hindi). Linguists consider word order to be a *coding property*. Along with other coding properties such such as morphological inflection, function words, etc., word order is used as the surface realisation of the underlying syntactic structure. In English, it is the the most prominent coding property,

playing a crucial role in disambiguating the roles of a sentence's constituents.[10]

Evidence for the role of word order information comes from a variety of studies using acceptability judgements, eye-tracking data, and neural response measurements (Bahlmann et al., 2007; Bever, 1970; Danks and Glucksberg, 1971; Ding et al., 2016; Fedorenko et al., 2016; Friederici et al., 2001, 2000; Just and Carpenter, 1980; Lerner et al., 2011; Pallier et al., 2011). Psycholinguistic research has, however, also highlighted the robustness of our sentence processing mechanisms to a variety of perturbations, including those which violate word order restrictions (Ferreira et al., 2002; Gibson et al., 2013; Traxler, 2014). In recent work, Mollica et al. (2020) tested the hypothesis that composition is the core function of the brain's language-selective network and that it can take place even when grammatical word order constrains are violated. Their findings confirmed this, showing that stimuli with shuffled word order where local dependencies were preserved — as is, roughly speaking, the case for many dependencies in the sentences SHUF.N4 is trained on — elicited a neural response in the language network that is comparable that elicited by normal sentences. When interword dependencies were disrupted so combinable words were so far apart that composition among nearby words was highly unlikely — as in SHUF.N1, neural response fell to a level compared to unconnected word lists.

IN MACHINES A fair deal of recent research has gone into investigating the role of word order information in language models. Using diagnostic classifiers and an attention analyses, Lin et al. (2019) find that earlier, but not later layers of BERT encode order information. Papadimitriou et al. (2021) find that Multilingual BERT is sensitive to morphosyntactic alignment (how each language defines what classifies as a "subject") across 24 languages, many of which — like English — use word order to mark this feature. Alleman et al. (2021) implement a input perturbation framework (n-gram shuffling, phrase swaps, etc.), and use it to test the sensitivity of BERT's representations to several kinds of structure in sentences, finding that there is sensitivity to larger parts of an sentence in deeper layers, and that this is influenced by hierarchical phrase structure. O'Connor and Andreas (2021) examine the contribution of various contextual features to the ability of GPT-2 (Radford et al., 2019) to predict future tokens. Their findings show that several destructive manipulations, including in-sentence word shuffling, applied to mid- and long range contexts lead only to a modest increase in *usable information* as defined according to the V-information framework of Xu et al. (2020).

10 For more morphologically complex languages, on the other hand, (e.g. Finnish and Turkish), word order is primarily used to convey pragmatic information such as topicalisation or focus rather than grammatical information which is conveyed via morphological inflection.

Similarly, word order information has been found not to be essential for various NLU tasks and datasets. Early work showed that Natural Language Inference tasks are largely insensitive to permutations of word order (Parikh et al., 2016; Sinha et al., 2020). Pham et al. (2020) and Gupta (2003) expanded on this, demonstrating that test-time word order perturbations applied to GLUE benchmark tasks have little impact on LM performance. Going a step further, Sinha et al. (2021), which our work builds on, found that even pretraining on scrambled text appears to only marginally affect model performance. Most related to this study, Clouatre et al. (2021) introduce two metrics for gauging the local and global ordering of tokens in scrambled texts, observing that only the latter is altered by the perturbation functions found in prior literature. In experiments with GLUE, they find that local (sub-word) perturbations show a substantially stronger performance decay compared to global ones.

7.10 CONCLUSION

Much discussion has resulted from recent work showing that scrambling text at different stages of testing or training does not drastically alter the performance of language models on NLU tasks. In this work, we presented analyses painting a more nuanced picture of such findings; we demonstrate that a) as far as altered pre-training is concerned, models still do retain a semblance of word order knowledge; b) this knowledge stems from cues in the altered data, such as adjacent BPE symbols and correlations between sentence length and content; and c) that there exist NLU tasks that are far more sensitive to sentence structure as expressed by word order (as a coding property).

Part IV

NEURAL RESPONSE MEASUREMENTS

8

CONNECTING NEURAL RESPONSE MEASUREMENTS & COMPUTATIONAL MODELS OF LANGUAGE: A NON-COMPREHENSIVE SURVEY

8.1 ABSTRACT

Understanding the neural basis of language comprehension in the brain has been a long-standing goal of various scientific research programs. Recent advances in language modelling and in neuroimaging methedology promise potential improvements in both the investigation of language's neurobiology and in the building of better and more human-like language models. This survey traces a line from early research linking Event Related Potentials and complexity measures derived from simple language models to contemporary studies employing Artificial Neural Network models trained on large corpora in combination with neural response recordings from multiple modalities using naturalistic stimuli.

8.2 INTRODUCTION

The mechanisms which underlie language comprehension in humans have been the object of study of a broad range of scientific research programs. Early work in theoretical and psycho- linguistics hypothesized certain mechanisms and structures underlying language processing, often employing behavioural data to confirm or refute them (Chomsky, 1957, 2014a,b; Fodor, 1983; Fodor and Garrett, 1966; Geschwind, 1970; Greenberg, 1963; Katz and Fodor, 1963; Lakoff and Johnson, 2008; Lenneberg, 1967; Luce and Pisoni, 1998; McClelland et al., 1986; Prince and Smolensky, 2008; Rayner, 1998; Taylor, 1953), inter alia. With the advancement of neuroimaging technologies, it became possible to begin to localise some of the computations responsible for language in the brain — both in time and space — investigating a) the timeline of such computations and b) the brain regions (or networks) which carry them out.

Orthogonally, advances in the field of artificial intelligence have enabled the training and deployment of large artificial neural network models (ANNs). These models, which are loosely-based on upon the structure of biological brains (Haykin, 1994), have demonstrated remarkable adeptness at a wide variety of tasks (Bengio, 2009; Goodfellow et al., 2016; Graves et al., 2013; Krizhevsky et al., 2012; Schmidhuber, 2015; Silver et al., 2016). In the field of natural language processing, ANNs have all but replaced other types of statistical methods

previously employed, showing far superior performance on an assortment of natural language understanding tasks (Brown et al., 2020; Devlin et al., 2019b; Radford et al., 2019).

Although they fundamentally differ from the neural architecture of the human brain, the success of these models in approximating human behaviour on tasks such as object recognition and speech recognition and in various language understanding tasks led to the suggestion that they could be adopted as potential models of the representations and structures which underpin human cognition. In seminal work, researchers found that convolutional neural networks (LeCun, Bengio, et al., 1995) trained on large image classification datasets could predict image-evoked neural activations in the ventral visual stream with a higher accuracy than all previous models — even those directly optimised to fit neural activations (Yamins and DiCarlo, 2016; Yamins et al., 2014). Similar research followed for a variety of perceptual domains, showing that ANNs exhibit similarities to both human behaviour and neural responses and that those similarities arise simply as a consequence of learning to perform a task such as image classification (Eickenberg et al., 2017; Kell et al., 2018; Kriegeskorte, 2015). Analogously, for language, ANN models trained to predict future or masked words from context have recently been found to show considerable representational alignment to neural response the human brain (Caucheteux and King, 2020; Goldstein et al., 2021; Schrimpf et al., 2020c).

SCOPE This paper makes a survey of research which links computational models of and neural responses to language. Particular attention is given to more recent work which makes use of ANNs, with the goal of tracing a line between this trend and previous work leveraging simpler computational models such as syntactic parsing models or context-free grammars. Studies which combine those, or any type of computational model, with neuroimaging data are considered to be within the scope of the survey, while those e.g. primarily focusing on the stimuli themselves are not. For other relevant surveys with different foci, readers are referred to Murphy et al. (2018) and Hale et al. (2021).

OVERVIEW The rest of this paper is structured as follows: Section 8.3 outlines a set of preliminaries, establishing terminology that will be used throughout the paper; Section 8.4 offers historical context; Section 8.5 describes early work which focused on representations of and neural responses to words, treated as isolated units; Section 8.6 surveys research that employs computational models in investigations of syntactic structure in the brain; Section 8.7 moves forward to studies which apply more complex analyses that account for multiple levels of perceptual and linguistic abstraction; Section 8.8 looks at

work that applies large-scale "integrative benchmarking" to establish patterns of performance across many language models and neural response datasets; Section 8.9 presents recent work on using language models to apply controls at the computational level rather than at the level of stimuli; Section 8.10 shifts to work that aims to evaluate and improve language models using insights and data from neurolinguistics, and finally, Section 8.11 presents a discussion and outlook.

8.3 PRELIMINARIES

LANGUAGE MODELS We take "language model" to mean any computational model that aims to explain and make predictions about some aspect of language. This includes early models of syntactic structure (Bresnan et al., 2015; Chomsky, 1956; Pollard and Sag, 1988), lexical distributional models (Mikolov et al., 2013; Pennington et al., 2014; Schütze, 1993), and more recent ANN models (Hochreiter and Schmidhuber, 1997; Vaswani et al., 2017).

NEUROIMAGING METHODS Work surveyed in this paper makes use of the following neuroimaging modalities:

- **Electroencephalogram (EEG)** involves recording the electrical activity occurring in the cortex over a period of time using multiple electrodes placed on the scalp (Henry, 2006). EEG is generally considered to have a high temporal but a relatively poor spatial resolution (centimeters) and is often used for deriving event-related potentials (ERPs) which average over the EEG signals resulting from of a specific event, e.g. reading a word (Luck, 2014).
- **Magnetoencephalography (MEG)** involves measurement of the magnetic field generated by the electrical activity of neurons in the cortex. Like EEG, MEG offers an accurate resolution of the timing of neuronal activity. Unlike EEG, it also offers a relatively good spatial resolution (millimeters) (Baars and Gage, 2013).
- **Functional Magnetic Resonance Imaging** measures neuronal activity in the brain via blood oxygenation level-dependent (BOLD) contrast. fMRI offers high spatial resolution and signal reliability, but poor temporal resolution (3 to 6 seconds) due to slow nature of the hemodynamic response (Soares et al., 2016).
- **Electrocorticography (ECoG)** records electrical activity in the brain through electrodes placed in direct contact with the surface of the brain (Baars and Gage, 2013). ECoG data has a fine spatial and temporal resolution and a high signal-to-noise ratio. As the procedure is invasive, however, ECoG data can only be gathered as part of a clinical procedure.

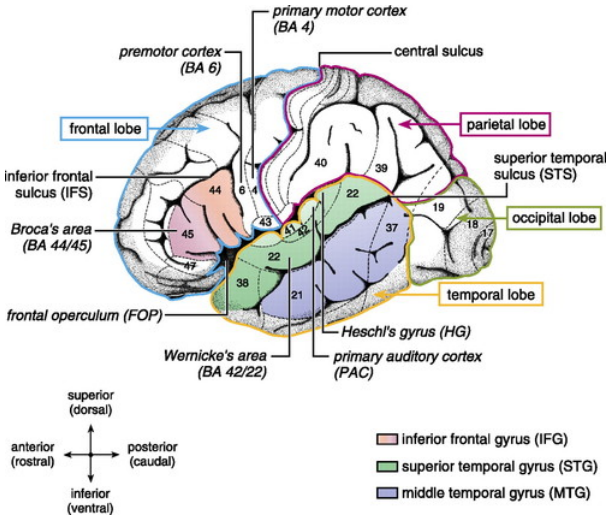


Figure 8.1: Major language relevant gyri and Brodmann areas in the Left Hemisphere. Figure from Friederici (2011).

THE LANGUAGE NETWORK A range of brain regions — often refered to as regions of interest (ROIs) — in the left and right hemisphere have been implicated in facilitating language. Although there is no consensus on the exact functional neuroanatomy, the network is known to, broadly speaking, include parts of the inferior frontal, the superior temporal, and the middle temporal gyra in the frontal and temporal lobes as well as the inferior parietal and angular gyrus in the parietal lobe (Friederici, 2011) — see Figure 8.1. For more details see Blank et al. (2016), Fedorenko et al. (2020), Fedorenko and Thompson-Schill (2014), Friederici and Gierhan (2013), Poeppel (2014), and Pylkkänen (2019, 2020).

LINKING HYPOTHESES To relate the computational and neural recording paradigms, a linking hypothesis is assumed — a commonly employed hypothesis, for instance, is that brain activation magnitude should correspond to measures of processing difficulty or complexity derived from the stimuli. Another common linking hypothesis is that of linear mapping, where a linear model is trained to map between e.g. features extracted from a computational model and neural recordings of a set of stimuli, under the assumption that a linear transformation should suffice to model the relationship between the two spaces. Goodness of fit or predictive accuracy are used to evaluate correspondence between the two.

8.4 BACKGROUND

Investigations into the neural basis of language processing aim to characterize the processes which occur once an utterance is perceived that enable a listener to arrive at a contextualised meaning from the

sensory input she perceives. Such investigations date back at least to Broca et al. (1861) and Wernicke (1874)'s description of two brain regions linked to the language faculty[1]. In the 1970s, the introduction of non-invasive brain-monitoring techniques allowed scientists to characterize activity in the brain using methods such as EEG and MEG. Using these methods, researchers initially identified Event-related potential components such as N400, P600, and ELAN[2] which corresponded to levels of syntactic or semantic processing (Friederici et al., 1993; Hagoort et al., 1993; Kutas and Hillyard, 1980, 1984).

While these early studies were successful at identifying particular patterns of activation which occurred during linguistic processing, they did not offer computational models which could make directly testable predictions of activation patterns. Mitchell et al. (2008) did this, demonstrating that word representations based on co-occurrence statistics could in fact be used to predict the activations associated with concrete nouns as measured via fMRI recordings. To accomplish this, they encode each stimulus word as a vector of semantic features computed from the its occurrences in a large text corpus. An linear encoding model is then trained to predicts fMRI activation per voxel in the brain, as a weighted sum of the semantic features. Evaluation was carried out through "leave-two-out" cross-validation, in which a model is repeatedly trained while holding out two word stimuli from the full set, then tested on whether its predicted fMRI image for these two stimuli can select the correct one via cosine similarity.

8.5 THE MEANINGS OF WORDS AND PHRASES

Following a setup similar to Mitchell et al. (2008): Murphy et al. (2009) applied the same methodology using EEG instead of fMRI; Devereux et al. (2010) used four automatic feature extraction methods leveraging different sources of information from corpora; Pereira et al. (2013) employed low-dimensional representations constructed by applying topic modelling to a small wikipedia corpus, showing that this feature space can outperform the one used by Mitchell et al. (2008) in a classification task based on decoding the values of semantic features present in concepts from the fMRI data; Palatucci et al. (2009) also reversed the original task, decoding word representations based on both co-occurance statistics and human annotations from their corre-

1 Although the term is still commonly employed, it is now understood that Broca's area 'is not a natural kind' and instead consists of consists of multiple functionally distinct components (Fedorenko and Blank, 2020) only part of which are directly linked to language.

2 N400, related to semantic processing, is a negative-going potential, which peaks around 400ms after stimulus onset; P600, associated with syntactic processing, is positive-going potential peaking around 600ms after stimulus onset. ELAN is an early left anterior negativity, characterized by a negative-going wave that peaks around 200 ms or less after onset.

sponding fMRI recordings in a zero-shot setting; Sudre et al. (2012) studied the temporal sequence of language processing, showing that perceptual and semantic features could be decoded at different times from MEG data; Anderson et al. (2013) used image-based distributional semantic representations of concepts instead of text-based ones; Anderson et al. (2017a) built on this, showing that although both perform equally well for concrete concepts, text-based word representations better predict the brain activations of abstract concepts compared to visually-grounded ones; Bulat et al. (2017) employed multiple evaluation methods to carry out a systematic appraisal of how well a wide range of text-based and grounded semantic models — including more recent skip-gram and bag-of-words word embeddings (Mikolov et al., 2013) — can predict fMRI measuremets; Pereira et al. (2018) presented a new fMRI dataset of subjects reading words and passages, showing that a decoder trained to predict word embeddings from imaging data for individual concepts that were selected using a novel sampling procedure designed to cover the entire semantic space, can generalize to new concepts and can, further, decode sentence stimuli represented as a simple average of (content) word embeddings; finally, Abnar et al. (2017) evaluate eight types of word embeddings on how well they predict the fMRI activation recordings from Mitchell et al. (2008), finding that a word embedding method which incorporates syntactic information fares best compared to multiple other word representations base on the skip-gram approach, matrix factorization, or crowd-sourced association features.

In a notable critique, Bullinaria and Levy (2013) present evidence that it is the lack of representational distinctiveness of the fMRI voxel activation vectors that is the major limiting factor in the kind of work described above. Gauthier and Ivanova (2018) take aim at the evaluation methods employed in decoding studies where 'semantic representations' derived from stimuli are decoded from brain activations. They show that the evaluation techniques used in these studies are underspecified and are therefore not able to distinguish between sentence representations drawn from models optimized for very different tasks. For using these models to make meaningful conclusions about the way linguistic processing is realized in brain activity, they recommend: a) clear specification of the task and mechanisms in the brain hypothesized to be generating or consuming a given representation, b) breaking down the feature space into interpretable subspaces (see Section 8.9 for examples of work where this is applied), and c) using encoding models to ablate the extent to which different model components can explain variance in neural response. Finally, Beinborn et al. (2019) presented a standardized framework for brain-encoding studies, demonstrating the effect which choice of evaluation measure has on the interpretation of model predictive power. Based

on this, they offered a set of recommendations regarding choice of metrics and reporting of results.

8.6 SEARCHING FOR SYNTAX

Concurrently, researchers have also sought to understand how the brain computes and represents syntactic structure during language comprehension. By varying the type of stimuli presented to subjects and carrying out comparisons between e.g. a) sentences of different complexities or b) sentences and lists of words, such studies were able to make conclusions about the brain regions which show sensitivity to sentence structure and about the temporal profile of brain activity (Brennan and Pylkkänen, 2012; Dapretto and Bookheimer, 1999; Fedorenko et al., 2012; Hagoort, 2005; Humphries et al., 2006; Just et al., 1996; Pallier et al., 2011).

The use of computational models in such work can be traced back to pyscholinguistic studies of syntactic processing difficulty where probablistic language models (such as probabilistic context-free grammars (PCFG)) were used to provide predictions about human reading times or grammaticality judgements (Christiansen and Chater, 1999; Hale, 2001; Levy, 2008; Reitter et al., 2011; Tabor and Tanenhaus, 1999). Parviz et al. (2011) found that measures derived from an incremental syntactic parser and a 4-gram markov chain language model were predictive of the N400 ERP component.

Frank et al. (2013) expanded on these findings, showing that word surprisal estimates from a Recurrent Neural Network language (RNN) model provided better predictions compared to n-gram models and phrase structure grammars. Frank et al. (2015) went a step further, extracting four different word information (word and part-of-speech surprisal and entropy reduction) measures and evaluating their predictivity of six different ERP components which are known to be sensitive to violations. Their results indicated that readers' expectations about upcoming words do not necessarily rely on hierarchical sentence structure (see Frank and Christiansen (2018) for further relevant discussion).

Hale et al. (2015) apply the same approach to neural time courses obtained using fMRI. In contrast with Frank et al. (2015), their findings showed that grammatical predictors were predictive of BOLD (see Section 8.3) over n-gram baselines, indicating the sensitivity of human sentence processing to hierarchical structure, at least in the anterior temporal lobe. They posited that this discrepancy might be due the difficulty of measuring nuanced syntactic processing activity with behavioral and ERP measures. Brennan et al. (2016) collected fMRI recordings of subjects listening to naturalistic storytelling data. Using this, they showed findings similar to those of Hale et al. (2015): abstract syntactic structures (context-free phrase structure grammars

and and mildly context-sensitive X-bar structural descriptions) were predictive of brain activity in the temporal lobes, but not in other areas, whereas predictors derived from n-gram models showed correlations across a broad network of areas. Brennan and Pylkkänen (2017) built on this, using MEG data, they found that anterior temporal lobe (ATL) activity is well-predicted by a parse-step measure derived from a predictive left-corner parser, which is consistent with the hypothesis that the ATL is sensitive to basic combinatoric operations. Henderson et al. (2016) also arrived at similar findings. Using fixation-related fMRI[3] and syntactic surprisal statistic derived from a PFCG, they found that this surprisal measure modulates activity in the left ATL and the left inferior frontal gyrus. Brennan and Hale (2019) further addressed the extent to which hierarchical structure is needed for language comprehension, using a naturalistic EEG dataset of participants listening to a chapter of Alice in Wonderland. They corroborate that, for left-anterior and right-anterior electrodes after around 200 ms from onset, syntactic surprisal measures derived from models which condition on hierarchical structure capture variance beyond models that condition on word sequences alone.

Zooming in on the algorithmic level, Stanojević et al. (2021) employed Combined Categorical Grammar (CCG) (Steedman and Baldridge, 2011), a mildly context-sensitive grammar, to extract node count based complexity metrics. They found that these metrics could better predict fMRI activation time courses in language ROIs compared to Penn Treebank-style context-free phrase structure grammars, confirming Brennan et al. (2016)'s finding that mildly context-sensitive grammars can better capture aspects of human sentence processing. They attribute this to an operation designed to make CCG handle "movement" constructions in a more plausible manner compared to CFGs.

Hale et al. (2018) pioneered the use of Recurrent neural network grammars (Dyer et al., 2016) (RNNGs) as models of human syntactic processing. RNNGs are generative models of both trees and strings (jointly) where neural networks are parametrized to choose parsing transitions. Using the same linking hypotheses described above and a beam search procedure (over derivations), they found that the measures estimated from the RNNG's intermediate states predicted several ERP components including P600. These same components were not significantly predicted using an LSTM (Hochreiter and Schmidhuber, 1997) language language model which operates sequentially, having no access to structural information. Brennan et al. (2020) build on this, using RNNGs in conjunction with fMRI data recorded for the same Alice in Wonderland chapter to localise the information used for predictive processing across six ROIs associated with language pro-

3 A method that combines eyetracking with BOLD to locate brain activity as a function of the currently fixated item.

cessing. Their findings, confirm earlier ones: word-by-word surprisal[4] derived from a sequential LSTM language model correlates with activity in a range of temporal and frontal brain regions; when surprisal is conditioned by explicit hierarchy as in the RNNG language model, fit is improved above the LSTM model in the left posterior temporal lobe and the left inferior parietal lobule. In addition, a measure based on the number of parsing steps explored by the RNNG model between words — which is hypothesized to reflect compositional processing — is found to be predictive of activation in most ROIs, particularly when the RNNG is setup to directly compose phrases, and not only encode hierarchy.

Most recently, Reddy and Wehbe (2021) proposed a shift from effort-based metrics (node count, surprisal, etc.) that allow for the localisation in the brain of a general notion of syntax to a methodology which allows for the studying of specific syntactic features. To accomplish this, they presented a subgraph embeddings-based method that models the constituency tree based structure of sentences. They showed that this method was more predictive of brain activity than a commonly used effort-based metric (node count), and used it to demonstrate that the brain encodes complex, phrase-level syntactic information.

8.7 MODELLING MULTIPLE LEVELS OF ABSTRACTION

Language comprehension from speech or text involves many perceptual and cognitive subprocesses, from perceiving individual words, to parsing sentences, to building semantic representations that are contextualised by world-knowledge and previous utterances in a discourse, etc. In this section, we survey work which has performed simultaneous analyses at multiple levels of perceptual and linguistic abstraction.

In one of the first studies to use fMRI recordings from subjects reading naturalistic data (a chapter from *Harry Potter and the Sorcerer's stone*), Wehbe et al. (2014a) presented one such integrated analysis. Employing a model consisting of a diverse set of visual, orthographic, lexical, syntactic, semantic, and discourse properties, they predicted neural activity per voxel as a linear combination of the features. Using this encoding (linear regression) model, they were able to examine the brain areas which were sensitive to the different types of features, enabling them to distinguish between areas on the basis of the type of information they represent. Using MEG data gathered for the same chapter of Harry Potter, Wehbe et al. (2014b) was one of the earliest works to investigate the alignment between the representations

4 Note that surprisal here and in Hale et al. (2018) is not the same as the syntactic surprisal measure based on part-of-speach category which is used in most previous studies, but is based on the lexical item itself.

used by RNN language models and brain activity as subjects read a story. They train auto-regressive neural language models (Mikolov et al., 2011) on a corpus of Harry potter fan fiction and extract three classes of features per time-step: the embeddings, the hidden state vectors (previous and current), and the predicted output probabilities. In a series of classification experiments involving prediction of the MEG signal from each of these feature classes, they analyse how well each predicts MEG activity along the temporal and spatial dimensions, finding: a) brain activity is well predicted by the hidden state representation of (past) context, b) the embedding features are good predictors of the current word, and c) the activity in different brain regions is predicted with a delay that corresponds with the processing pathway that starts in the visual cortex and moves up. Lopopolo et al. (2017) followed a similar approach with the goal of disentangling phonological, lexical, and syntactic information present in the fMRI recordings of subjects listening to a set of naturalistic (Dutch) literary texts. Instead of a neural network language model, however, they estimate perplexity statistics from trigram markov models (based on lexical forms, part-of-speech tags, and transcribed phonemes). Their results evidence a set of cortical networks that are separately activated for each of the three types of information, with no significant overlap between them.

Huth et al. (2016) also made use of a naturalistic dataset of spoken stories. They represented each word in the stories as a 985-dimensional vector built from co-occurrence statistics, intended to encode semantic information. A linear regression model was estimated per voxel, from the word representations. Controlling for lower-level features, they evaluated correlation between predicted and actual BOLD signal, finding significant predictiveness in various areas associated with the brain's semantic network. They propose a Bayesian algorithm that constructs a generative model of areas tiling the cortex across subjects, resulting in single atlas that describes the distribution of semantically selective functional areas in human cerebral cortex. Jain and Huth (2018) follow Wehbe et al. (2014b) in using an RNN language model to incorporate context into encoding models that predict the neural response (fMRI in this case) of subjects listening to natural speech. They find that the representations from LM hidden states outperform previously used non-contextual word embedding models in predicting brain response and that context length and choice of layer differentially predict the activation across cortical regions. LeBel et al. (2021) followed a similar approach, focusing on the Cerebellum. Using features that span the hierarchy of language processing, they showed that neural response to language in the Cerebellum is best explained by high-level conceptual features — especially those associated with social semantic categories — rather than lower-level acoustic or phonemic ones. Finally, Li et al. (2020) formalised several

linguistic theories about pronoun resolution as symbolic computational models, evaluating them as well as an ANN conference resolution model according to their ability to predict brain activity patterns (both MEG and fMRI) time-locked at each third person pronoun in 'The Little Prince' dataset as English and Chinese subjects listened to an audiobook recording (Stehwien et al., 2020). They find that the memory-based ACT-R model (Van Rij et al., 2013), which chooses the entity in working memory with the highest activation as the antecedent of a pronoun, best explains the neural response associated with pronoun resolution.

In seminal work adjacent to that described above because it studies neural oscillation patterns, Martin and Doumas (2017) showed that a computational model (Doumas et al., 2008), which uses time to encode hierarchy and was originally designed for relational reasoning, can be applied to sentence processing, exhibiting oscillatory patterns of activation closely resembling the human cortical response to the same stimuli. From this model which learns and generates structured, symbolic representations using time-based binding in a layered neural network, they are able to derive an explicit computational mechanism[5]) for how the human brain might convert perceptual features into hierarchical representations, offering a linking hypothesis between linguistic and cortical computations. Martin (2020) built on this, proposing an integrated model of language comprehension across multiple timescales starting from perception and lexicalisation to syntactic composition and comprehension. Drawing on ideas from models of entertainment to speech, structure building mechanisms from systems neuroscience (coordinate transforms via gain modulation), and the neurosymbolic relation learning model referenced above, the proposed model is able to unify the neurphysiological, cognitive, and linguistic computational levels.

8.8 INTEGRATIVE BENCHMARKING AND COMPUTATIONAL CONVERGENCE

While the shift from using several separate models which operate at different levels of abstraction to synergistic neural models has led to both better prediction of neural response and better performance on a range of linguistics tasks, Schrimpf et al. (2020b) posited that in order to understand the relationship between these computational models, neural response, and behaviour, large-scale integrative benchmarking is needed wherein patterns of performance are established across many models and datasets. Using measurements from three datasets[6] of neural response recordings (fMRI and ECoG) as well self-paced

5 See discussion of Gauthier and Ivanova (2018) in Sec. 8.5.

6 With stimuli that varied in length and domain and was either presented to subjects as audio or read.

reading data, they tested a wide range of computational models, from simple word embeddings to larger, recurrent and self-attention based ones, evaluating them based on how well they predicted the neural response recordings and the self-paced reading patterns, with reference to how well they perform tasks such as next word prediction. Their results demonstrated that: a) there is a variance across models in ability to predict neural response and self-paced reading patterns (e.g. GPT-2 (Radford et al., 2019) almost completely explains the variance in neural response, while GloVe performs poorly), b) there is a consistency in how models score across datasets and experiments, c) models that perform better at next word prediction (but not the GLUE suite of natural language understanding tasks (Wang et al., 2018a)) better predict neural response measurements and self-paced reading times, d) models that better predict neural response better predict reading times, and e) for some models , architecture alone, randomly initialised, can reliably predict brain activity and reading times.

Also aiming to establish a systematic ontology of whether and when ANN models representations align with brain activations, Caucheteux and King (2020) trained 7,400 different ANN models with different architectures and objectives. They evaluated the models according to how well a ridge-regularised linear could be trained to map from their internal representations to MEG recordings of 104 subjects reading words sequentially presented randomly word or as sentences. Aiming to arrive at a spatio-temporal decomposition of the reading network, they found: a) as previously shown, the final-layer activations a deep convolutional neural network trained on character recognition are predictive of activation in the early visual cortex, b) word-type embeddings (Word2Vec (Mikolov et al., 2013)) predicted brain response above and beyond the visual representations starting from $\approx$ 200 ms after onset, in the left-lateralized temporal and prefrontal cortices, especially, c) After $\approx$ 1 second from word onset, contextualised word representations from LMs led to significantly better prediction than the previous two feature sets, particularly in the regions associated with high-level sentence processing. Subsequently, Caucheteux et al. (2021b) found that GPT-2's predictiveness of fMRI activation for a subject-story pair in the Narratives dataset correlated to subjects' comprehension scores assessed per story.

Most recently, Antonello et al. (2021) adapted a encoder-decoder method from computer vision that measures transferability between different models to construct a language representation embedding space. Using this, they could describe and visualise the relationships between representations derived from a 100 diverse types of language models, ranging from static word-embeddings (GloVe, etc.) and interpretable tagging models (part-of-speech, named entity recognition, chunking etc.) to machine translation models and autoregressive LMs

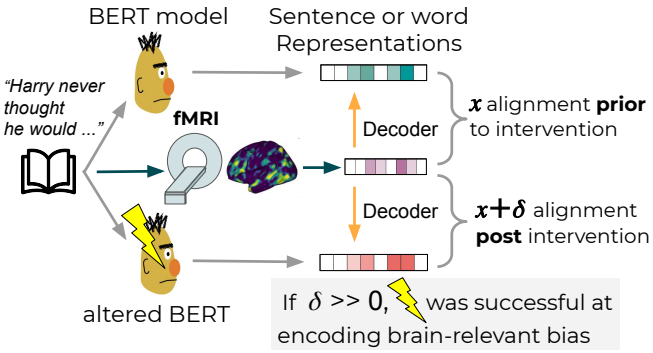


Figure 8.2: An example of the application of computational control. A baseline BERT model and an 'altered' one are used to generate linguistic representation of stimuli. The intervention to alter the LM can be evaluated based on how well the resulting representations can be decoded from brain activation data compared to the baseline, controlling for other factors.

(GPT-2, etc.) or masked LMs (BERT, etc.). They show that this space has a low-dimensional structure and that it intuitively models how different representations relate to one another. Moreover, fitting encoding models to predict fMRI data from each of the 100 language representations, they find the embedding space's structure, when mapped to the brain, reflects well-known language processing hierarchies and predicts which representations map well to which areas in the brain.

8.9 COMPUTATIONAL CONTROLS

To establish a relationship between linguistic function and neurobiology, it is necessary to decompose the various facets of linguistic processing and map them onto the units of neurobiology (Poeppel, 2012). Traditionally, this has been attempted through careful manipulation of stimuli e.g. scrambled sentences vs. natural sentences. Instead of applying strict controls at the level of the stimuli, recent work has explored the possibility of applying computational controls (see Figure 8.2 for a demonstrative example). When successful, this allows for a) testing a wider range of more specific questions concerning conditions which would be significantly more difficult or expensive to control for at the level of the stimuli and b) the use of naturalistic stimuli which more closely resemble real-world contexts.

For instance, by varying the length of context which is fed to a LM or providing it with distorted context, Jain and Huth (2018) were able to ablate and the effect contextual information on alignment to neural response. Abnar et al. (2019a) followed a similar protocol, employing an extension of the commonly used Representation Similarity Analysis paradigm (Kriegeskorte et al., 2008) where different instances of the same model are compared as a single parameter is altered.

Varying the parameter of context length for four classes of language models, they find that increased context length not lead to increased representational alignment to brain recordings.

Gauthier and Levy (2019) took a pretrained masked language model (BERT (Devlin et al., 2019b)) as a baseline sentence representation model, finetuning it on a suite of tasks which are commonly used to evaluate 'Natural Language Understanding'. Training a regularised linear decoder to map from the fMRI sentence-level data of Pereira et al. (2018) to representations extracted using the finetuned models, they find decreased brain decoding performance across all tested NLU tasks, but find improved decoding performance for scrambled language modelling tasks where fine-grained syntactic information is removed. This result might be seen as disagreeing with an emerging consensus in the literature regarding the role of hierarchical structure in sentence processing [7] and once again raised questions regarding the fine-grained-ness of the information which can be expected to be found in fMRI data using linear mapping as a linking hypothesis. Abdou et al. (2021) followed a similar methodology, proposing an approach which enables the evaluation of more targeted hypotheses about linguistic composition and structure by finetuning LMs with an auxilliary attention constraint to inject structural bias derived from three linguistic formalisms into LM representations. They showed that, across the three formalisms, this improves brain decoding performance for the Harry Potter data, while for the Pereira et al. (2018) data the effect is less clear.

The methods outlined so far apply controls by manipulation of either the input or the training objective and task. Toneva et al. (2020) devised an approach for studying the neural substrates associated with the 'supra-word meaning' of a phrase with a computational control that disentangles composed-from individual-word meaning. Using regularised linear regression models, they learn mappings between word embeddings and context representations, which they then subtract from the representations of items in the target space to obtain 'residual' representations for a word (sans contextual information) or for a phrase *beyond* its individual words. For an fMRI dataset, they find that this supra-word representation predicts activity in the anterior and posterior temporal cortices. In MEG recordings, however, they find no clear signal for supra-word meaning. Similarly, Caucheteux et al. (2021a) proposed a method to factorise distributed LM representations according to a taxonomy of: *syntax vs. semantics* and *lexical vs. compositional*. They constructed syntactic representations by averaging over the LM activations for a set of sentences with the same syntactic structure. A LM's word embedding is taken as a lexical representation, and the contextualised representation of higher layers as the

7 e.g. Brennan et al. (2016), Hale et al. (2015), and Henderson et al. (2016), etc. but perhaps agreeing with Frank et al. (2013) and Frank et al. (2015). Refer to Section 8.6.

compositional one. A semantic representation is taken to be the residual of subtracting the syntactic representation, which can be done at the lexical or the compositional level. Extracting each of these representations from the relevant stimuli text, they learn ridge-regularised mappings to fMRI recordings of subjects listening to the stories from the Narratives dataset (Nastase et al., 2021). Their results showed: a) compositional representations recruit a broader cortical network than lexical ones and b) syntax and semantics appear to share a common neural basis, in line with recent findings from the neuroscience literature (Fedorenko et al., 2020).

Examining the question of predictive coding (Rao and Ballard, 1999) in language processing, Goldstein et al. (2021) investigated whether (and how) humans and LMs engage in spontaneous prediction when processing language. To accomplish this, they: a) used a sliding-window method to show humans could effectively predict the next word in a transcribed story across sentence positions and part-of-speech; b) extracted word-by-word prediction probabilities from GPT-2 for the same story, finding them to be correlated to the human predictability scores, with better correlations as the amount of previous context fed to the model increased; c) showed, in a set of experiments with ECoG data recorded while subjects listened to a spoken story, that static word embeddings (GloVe) as well as an arbitrary embeddings baseline designed to ablate the effect of correlations between adjacent word embeddings of bigrams, could predict neural activation (across electrodes in the left hemisphere) starting from 800 ms before word onset; d) at earlier time from word onset, neural responses were better predicted for predictable words than for unpredictable words, e) neural response before onset was better modelled for the words subjects predicted even when they did not match the correct words which they subsequently perceived; f) employing representations which incorporate previous context from GPT-2's activations led to both better and earlier (i.e. pre-onset) prediction of neural activity, particularly in high-order language areas; g) ablated this result, showing that it was due to both better representation of previous context (compared e.g. to a baseline of mean-pooled GloVe embeddings) and to improved next-word prediction. Finally, Jain et al. (2021) presented a novel multi-timescale encoding model for predicting fMRI responses to natural speech. Using a) an LSTM where the memory timescale of each individual unit is fixed according to a power law distribution and b) a Gaussian radial basis function kernel to downsample the stimuli representations, they are able directly estimate the timescale of information represented in a voxel of fMRI data from the encoding model's weights. They find that this method which relies on 'computational control' leads to a more fine-grained map of timescale selectivity compared to previous work which relied on stimulus manipulation (Lerner et al., 2011).

8.10 USING BRAIN ACTIVATION MEASUREMENTS TO IMPROVE/EVALUATE NLP MODELS

The work described in this survey so far has employed computational language models towards the study of human language processing. Working in the other direction, researchers have also attempted to leverage data and insights from neurolinguistics to evaluate and improve language models. To that end, Fyshe et al. (2014) presented an algorithm that integrates brain activations into the construction of word-type vector space models. Their approach, based on Non-Negative Sparse Embeddings (Murphy et al., 2012), constrains the embedding space so that words close in brain activation space also have similar representations in the embedding space. They find the resulting embeddings to better match a behavioral measure of semantics and to better predict corpus data for unseen words. In an opinion paper on word embedding evaluation, Søgaard (2016) presented some preliminary experiments and argued that because evaluation on downstream tasks is expensive and impractical and evaluation on corpus statistics is circular, alignment to human word processing measurements should be used as a an evaluation approach. On that account, Hollenstein et al. (2020a) presented CogniVal, a framework for the evaluation of word embeddings based on their ability to predict data from 15 datasets of eyetracking, EEG, and fMRI signals recorded during language processing.

Testing the possibility of using EEG data to improve NLP models, Hollenstein et al. (2019) extract word-level EEG features from different frequency ranges, using the ZUCO dataset (Hollenstein et al., 2020b) which contains simultaneous eye-tracking and EEG measurements of natural sentence reading. Combining these features with standard word-level and character-level information employed by NLP models, they find (modest) improvements over baselines which do not include the EEG features across three tasks (named entity recognition, relation classification and sentiment analysis).

Pioneering the use of LMs for *in silico* modelling[8], Schwartz et al. (2019) finetuned BERT models to predict (fMRI and MEG) brain activity measurements, biasing them to learn generalizable relationships between text and brain activity. They find that the the fine-tuned models are better at predicting brain activity across subjects and recording modalities than non-finetuned models. Finally, Toneva and Wehbe (2019a) used brain response data to interpret the information encoded in the internal representations of different layers of LMs for various context lengths, based on how well they predicted activation in different groupings of language ROIs. Ablating the role of attention, they found that replacing the learned attention function with a uniform distribution in early BERT layers led to better prediction of

8 Where a computational model is used to simulate (some facet of) brain function.

brain activation. Applying these altered models to a task which evaluates the syntactic capabilities of LMs (Marvin and Linzen, 2018), they found significant improvements compared to the vanilla pretrained model, demonstrating that altering models so that they align better with brain recordings can lead to better performance on NLP benchmarks.

8.11 OUTLOOK

Present-day advances in machine learning have enabled the building of language models which through training on immense volumes of data are capable of simulating human behaviour on various linguistic tasks better than ever before. Concurrently, datasets of neural response recordings that both include more subjects and utilise a larger amount of naturalistic stimuli than previously possible have been made openly accessible to the research community (Bhattasali et al., 2020; Nastase et al., 2021; Stehwien et al., 2020). Exploiting these advances, recent studies a) found evidence that autoregressive ANN language models converge to solutions that reliably align to brain activations (Caucheteux and King, 2020; Schrimpf et al., 2020c) and b) worked towards furthering our understanding of fundamental aspects of naturalistic language comprehension, e.g. the role of hierarchical structure (Brennan and Hale, 2019; Stanojević et al., 2021), the function of predictive coding (Goldstein et al., 2021), and the neural substrates responsible for lexical and combinatorial semantics (Toneva et al., 2020).

While the manner in which current ANN language models learn is manifestly inefficient and un-human-like — leveraging text-only information and requiring orders of magnitude more data than a human child — the solutions to which they converge have been shown to both remarkably simulate aspects of linguistic understanding and align to cognitive measurements. This offers reasonable grounds for optimism that the time is ripe for these models to both contribute to and benefit from work on the *Mapping Problem* (Poeppel, 2012): the problem of mapping the elementary units of linguistic processing to their neurobiological counterparts. As Hale et al. (2021) argues, linguistically-interpretable models are likely to be key for this symbiosis, allowing for a principled decomposition of a model's components into smaller linguistically meaningful units. In view of the approaches to computational control described in Section 8.9, I posit that even models that are not intrinsically interpretable will be useful, given the plethora of interpretability methods recently developed (Belinkov et al., 2020).

By serving as a plausible simulation of human learning, future work that explores more human-like training setups (e.g. multi-modal or embodied learning), objectives, and model architectures can also

help empirically answer long-standing questions in neuro- and cognitive linguistics.

9

DOES INJECTING LINGUISTIC STRUCTURE INTO LANGUAGE MODELS LEAD TO BETTER ALIGNMENT WITH BRAIN RECORDINGS?

9.1 ABSTRACT

Neuroscientists evaluate deep neural networks for natural language processing as possible candidate models for how language is processed in the brain. These models are often trained without explicit linguistic supervision, but have been shown to learn some linguistic structure in the absence of such supervision, potentially questioning the relevance of symbolic linguistic theories in modeling such cognitive processes. We evaluate whether biasing the attention of language models using annotations from syntactic or semantic formalisms leads to better alignment with fMRI brain recordings. Using structure from dependency or minimal recursion semantic annotations, we find alignments improve significantly for a dataset with complex, naturalistic stimuli; another dataset shows mixed results. We present an extensive analysis of the results. Our proposed approach enables the evaluation of more targeted hypotheses about the composition of meaning in the brain, expanding the range of possible scientific inferences a neuroscientist could make, and opens up new opportunities for cross-pollination between computational neuroscience and linguistics.

9.2 INTRODUCTION

Recent advances in deep neural networks for natural language processing have generated excitement among computational neuroscientists, who aim to model how the brain processes language. These models are argued to better capture the complexity of natural language semantics than previous computational models, and are thought to represent meaning in a more similar way to how it is hypothesized to be represented in the human brain. For neuroscientists, these models provide possible hypotheses for *how* word meanings compose in the brain. Previous work has evaluated the plausibility of such candidate models by testing how well representations of text extracted from these models align with brain recordings of humans during language comprehension tasks (Abnar et al., 2019a; Caucheteux and King, 2020; Gauthier and Ivanova, 2018; Gauthier and Levy, 2019; Goldstein et al., 2021; Jain and Huth, 2018; Schrimpf et al., 2020a;

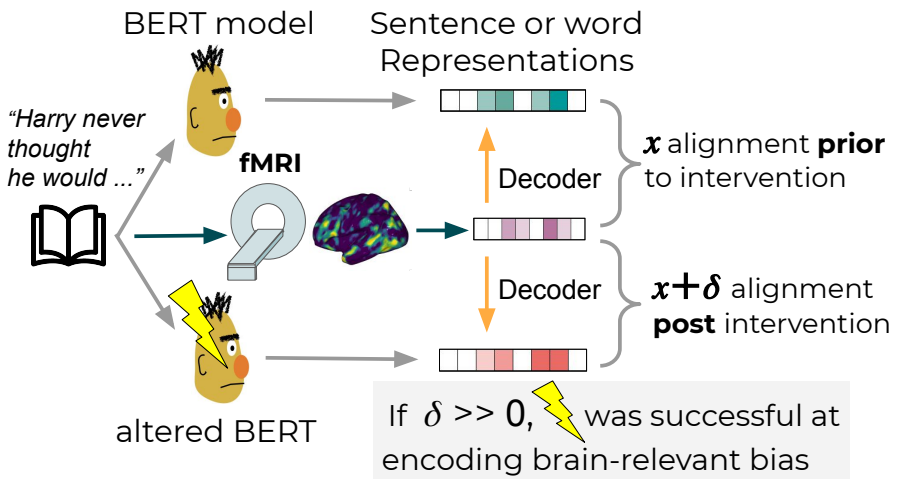


Figure 9.1: Overview of our approach. We use BERT as a baseline and fine-tune to inject structural bias. Through a brain decoding task, we then compare the alignment of the representations of the baseline and the altered models with brain activations.

Toneva et al., 2020; Toneva and Wehbe, 2019b; Wehbe et al., 2014b), and found some correspondences.

However, modern NLP models are often trained without explicit linguistic supervision (Devlin et al., 2019a; Radford et al., 2019), and the observation that they nevertheless learn some linguistic structure (Manning et al., 2020) has been used to question the relevance of symbolic linguistic theories. Whether injecting such symbolic structures into language models would lead to even better alignment with cognitive measurements, however, has not been studied. If this were the case, it would suggest that these theories represent aspects of linguistic structure relevant to human language comprehension which are not currently captured by language models via their pretraining objectives. To investigate this, we train a widely used language model (BERT, see §9.4.1) with an additional structural bias, and evaluate its alignment with brain recordings (§9.4.2). Structure is derived from three formalisms—Univeral Dependencies, Delph-in MRS Bi-Lexical Dependencies, and Universal Cognitive Conceptual Annotation (§9.4.3)—which come from different linguistic traditions, and capture different aspects of syntax and semantics.

Our approach, illustrated in Figure 9.1, allows for quantifying the brain alignment of the structurally-biased NLP models in comparison to the base models, as related to new information about linguistic structure learned by the models that is also potentially relevant to language comprehension in the brain. More specifically, in this paper, we:

(a) Propose a finetuning method utilising structurally guided attention for injecting structural bias into language model representations.

(b) Assess the representational alignment to brain activity measurements of the finetuned and non-finetuned LMs.

(c) Further evaluate the LMs on a range of targeted syntactic probing tasks and a semantic tagging task, which allow us to uncover fine-grained information about their structure-sensitive linguistic capabilities.

(d) Present an analysis of various linguistic factors that may lead to improved or deteriorated brain alignment.

9.3 BACKGROUND: BRAIN ACTIVITY AND NLP

Mitchell et al. (2008) first showed that there is a relationship between the co-occurrence patterns of words in text and brain activation for processing the semantics of words. Specifically, they showed that a computational model trained on co-occurrence patterns for a few verbs was able to predict fMRI activations for novel nouns. Since then, researchers have attempted to isolate other features that enable prediction and interpretation of brain activity (Anderson et al., 2017b; Brennan et al., 2016; Frank et al., 2015; Lopopolo et al., 2017; Pereira et al., 2018; Wang et al., 2020). Gauthier and Ivanova (2018) however, emphasize that directly optimizing for the decoding of neural representation is limiting, as it does not allow for the uncovering of the mechanisms that underlie these representations. The authors suggest that in order for us to better understand linguistic processing in the brain, we should also aim to train models that optimize for a specific linguistic task and explicitly test these against brain activity.

Following this line of work, Toneva and Wehbe (2019b) present experiments both predicting brain activity and evaluating representations on a set of linguistic tasks. They first show that using uniform attention in early layers of BERT (Devlin et al., 2019a) instead of pretrained attention leads to better prediction of brain activity. They then use the representations of this altered model to make predictions on a range of syntactic probe tasks, which isolate different syntactic phenomena (Marvin and Linzen, 2019), finding improvements against the pretrained BERT attention.

Gauthier and Levy (2019) present a series of experiments in which they finetune BERT on a variety of tasks including language modeling as well as some custom tasks such as scrambled language modeling and part-of-speech-language modeling. They then perform brain decoding, where a linear mapping is learnt from fMRI recordings to the finetuned BERT model activations. They find that the best mapping is obtained with scrambled language modelling finetuning, which they

confirm leads to poor performance when reconstructing Universal Dependencies (UD; Nivre et al., 2020a) parse trees with a structural probe. Building on this, in this work we propose a framework for investigating how incorporating particular structural biases — such as those derived from linguistic theory — into language model representations can affect their alignment to cognitive measurements. Unlike previous work, our approach allows for the evaluation of more targeted hypotheses about the structures which underlie the composition of meaning in the brain.

9.4 APPROACH

Figure 9.1 shows a high-level outline of our experimental design, which aims to establish whether injecting structure derived from a variety of syntacto-semantic formalisms into language model representations can lead to better correspondence with human brain activation data. We utilize fMRI recordings of human subjects reading a set of texts. Representations of these texts are then derived from the activations of the language models. Following Gauthier and Levy (2019), we obtain LM representations from BERT[1] for all our experiments. We apply masked language model finetuning with attention guided by the formalisms to incorporate structural bias into BERT's hidden-state representations. Finally, to compute alignment between the BERT-derived representations—with and without structural bias—and the fMRI recordings, we employ the brain decoding framework, where a linear decoder is trained to predict the LM derived representation of a word or a sentence from the corresponding fMRI recordings.

9.4.1 *LM-derived Representations*

BERT uses wordpiece tokenization, dividing the text to sub-word units. For a sentence S made up of P wordpieces , we perform mean-pooling over BERT's final layer hidden-states $[h_1, ..., h_P]$, obtaining a vector representation of the sentence $S_{mean} = \frac{1}{P}\sum_p h_p$ (Wu et al., 2016). In initial experiments, we found that this leads to a closer match with brain activity measurements compared to both max-pooling and the special `[CLS]` token, which is used by Gauthier and Levy (2019). Similarly, for a word W made up of P wordpieces, to derive word representations, we apply mean-pooling over hidden-states $[h_1, ..., h_P]$, which correspond to the wordpieces that make up W: $W_{mean} = \frac{1}{P}\sum_p h_p$. For each dataset, $D_{LM} \in \mathbb{R}^{n \times d_H}$ denotes a ma-

1 Specifically: `bert-large-uncased` trained with whole-word masking. Note, however, that the choice of model is not very consequential as the approach to inducing structural bias and overal methodology are general and can be applied to any class of models which employs attention mechanisms.

trix of n LM-derived word or sentence representations where d_H is BERT's hidden layer dimensionality ($d_H = 1024$ in our experiments).

9.4.2 *Neuroimaging Datasets*

We utilize two fMRI datasets, which differ in the granularity of linguistic cues to which human responses were recorded. The first, collected in Pereira et al. (2018)'s experiment 2, comprises a single brain image per entire sentence. In the second, more fine-grained dataset, recorded by Wehbe et al. (2014b), each brain image corresponds to 4 words. We conduct a **sentence-level** analysis for the former and a **word-level** one for the latter.[2] Stimuli set complexity also varies between the two datasets; the former comprising of simple wikipedia-style sentences, and the latter more complex, naturalistic ones (see Appendix A.6.7 for examples).

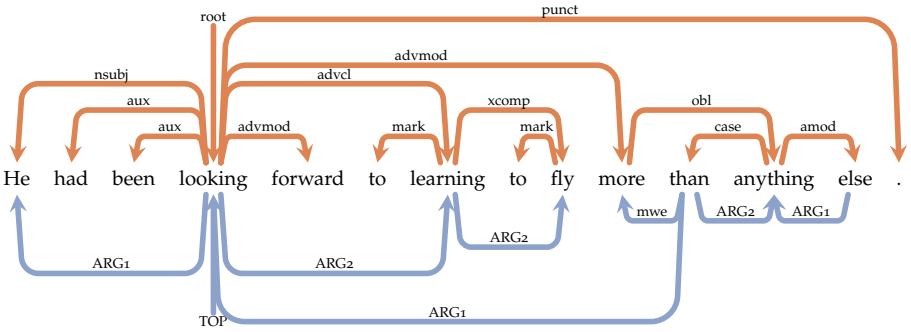
PEREIRA2018 consists of fMRI recordings from 8 subjects. The subjects were presented with stimuli consisting of 96 Wikipedia-style passages written by the authors, consisting of 4 sentences each. The subjects read the sentences one by one and were instructed to think about their meaning. The resulting data for each subject consists of 384 vectors of dimension 200,000; a vector per sentence. These were reduced to 256 dimensions using PCA by Gauthier and Levy (2019). These PCA projections explain more than 95% of the variance among sentence responses within each subject. We use this reduced version in our experiments.

WEHBE2014 consists of fMRI recordings from 8 subjects as they read chapter 9 from *Harry Potter and the Sorcerer's Stone*. For the 5000 word chapter, subjects were presented with words one by one for 0.5 seconds each. An fMRI image was taken every 2 seconds, as a result, each image corresponds to 4 words. The data was further preprocessed (i.e. detrended, smoothed, trimmed) and released by Toneva and Wehbe (2019b). We use this preprocessed version to conduct word-level analysis, for which we use PCA to reduce the dimensions of the fMRI images from 25,000 to 750, explaining at least 95% variance for each participant.

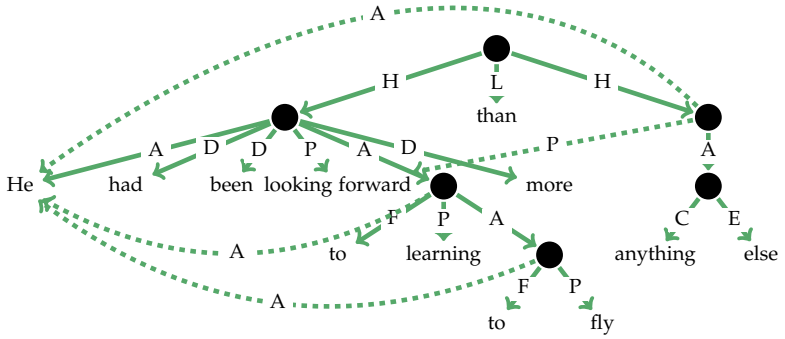
9.4.3 *Formalisms and Data*

To inject linguistic structure into language models, we experiment with three distinct formalisms for representation of syntactic/semantic structure, coming from different linguistic traditions and capturing different aspects of linguistic signal: UD, DM and UCCA. An

2 Even though the images are recorded at the 4-gram level of granularity, a word-level analysis is applied, as in (Schwartz et al., 2019).



(a) UD (above, orange), DM (below, blue)



(b) UCCA

Figure 9.2: Manually annotated example graphs for a sentence from the Wehbe2014 dataset. While UCCA and UD attach all words, DM only connects content words. However, all formalisms capture basic predicate-argument structure, for example, denoting that "more than anything else" modifies "looking forward" rather than "fly".

example graph for each formalism is shown in Figure 9.2. Although there are other important linguistic structured representation frameworks, including meaning representations such as AMR (Banarescu et al., 2013), DRS (Bos et al., 2017; Kamp and Reyle, 1993) and FGD (Hajic et al., 2012; Sgall et al., 1986), we select three relatively different formalisms as a somewhat representative sample. All three have manually annotated datasets, which we use for our experiments.

UD (Universal Dependencies; Nivre et al., 2020a) is a syntactic bi-lexical dependency framework (dependencies are denoted as arcs between words, with one word being the *head* and another the *dependent*), which represents grammatical relations according to a coarse cross-lingual scheme. We use UD 2.0 data for the English Web Treebank corpus (EWT; Silveira et al., 2014), which contains 254,830 words and 16,622 sentences, taken from five genres of web media: weblogs, newsgroups, emails, reviews, and Yahoo! answers.

DM (DELPH-IN MRS Bi-Lexical Dependencies; Ivanova et al., 2012) is derived from the underspecified logical forms computed by the English Resource Grammar (Copestake et al., 2005; Flickinger et al., 2017). We use the English SDP (Semantic Dependency Parsing) data for DM (Oepen et al., 2016), annotated on newspaper text from the Wall Street Journal (WSJ), containing 802,717 words and 35,656 sentences.

UCCA (Universal Cognitive Conceptual Annotation; Abend and Rappoport, 2013) is based on cognitive linguistic and typological theories, primarily Basic Linguistic Theory (Dixon, 2010/2012). We use UCCA annotations over web reviews from the English Web Treebank (Hershcovich et al., 2019), and from English Wikipedia articles on celebrities. In total, they contain 138,268 words and 6,572 sentences. For uniformity with the other formalisms, we use bi-lexical approximation to convert UCCA graphs, which have a hierarchical constituency-like structure, to bi-lexical graphs with edges between words. This conversion keeps about 91% of the information (Hershcovich et al., 2017).

9.4.4 *Injecting Structural Bias into LMs*

Recent work has explored ways of modifying attention in order to incorporate structure into neural models (Bugliarello and Okazaki, 2020; Chen et al., 2016c; Strubell and McCallum, 2018; Strubell et al., 2018b; Zhang et al., 2019a). For instance, Strubell et al. (2018b) incorporate syntactic information by training one attention head to attend to syntactic heads, and find that this leads to improvements in Semantic Role Labeling (SRL). Drawing on these approaches, we modify the BERT Masked Language Model (MLM) objective with an additional structural attention constraint. $\text{BERT}_{\text{LARGE}}$ consists of 24 layers and 16 attention heads. Each attention head $\texttt{head}_i$ takes in as input a

sequence of representations $h = [h_1, ..., h_P]$ corresponding to the P wordpieces in the input sequence. Each representation in h_p is transformed into query, key, and value vectors. The scaled dot product is computed between the query and all keys and a softmax function is applied to obtain the attention weights. The output of $head_i$ is a matrix O_i, corresponding to the weighted sum of the value vectors.

For each formalism and its corresponding corpus, we extract an adjacency matrix from each sentence's parse. For the sequence S, the adjacency matrix A_S is a matrix of size $P \times P$, where the columns correspond to the heads in the parse tree and the rows correspond to the dependents. The matrix elements denote which tokens are connected in the parse tree, taking into account BERT's wordpiece tokenization. All edges are modeled as bi-directional.[3]. We modify BERT to accept as input a matrix A_S as well as S; maintaining the original MLM objective. For each attention head $head_i$, we compute the binary cross-entropy loss between O_i and A_S and add that to our total loss, potentially down-weighted by a factor of α (a hyperparameter). BERT's default MLM finetuning hyperparameters are employed and α is set to 0.1 based on validation set perplexity scores in initial experiments.

Structural information can be injected into BERT in many ways, in many heads, across many layers. Because the appropriate level and extent of supervision is unknown a priori, we run various finetuninig settings with respect to combinations of number of layers $(1, \ldots, 24)$ and attention heads $(1, 3, 5, 7, 9, 11, 12)$ supervised via attention guidance. Layers are excluded from the bottom up (e.g.: when 10 layers are supervised, it is the topmost 10); heads are chosen according to their indices (which are arbitrary).

This results in a total of 168 finetuning settings per formalism. For each finetuning setting, we perform two finetuning runs.[4] For each run r of each finetuning setting f, we derive a set of sentence or word representations $D_{fr} \in \mathbb{R}^{n \times d_H}$ from each finetuned model using the approach described in §9.4.1 for obtaining D_{LM}, the baseline set of representations from BERT before finetuning. We then use development set[5] embedding space hubness—an indicator of the degree of difficulty of indexing and analysing data (Houle, 2015) which has been used to evaluate embedding space quality (Dinu et al., 2014)—as an unsupervised selection criterion for the finetuned models, selecting the model with the lowest degree of hubness (per formalism) ac-

3 By modelling edges as bi-directional we bias both dependents' attention weights to have higher values for heads, as well as those of heads to have higher values for dependents. Although edges are directed in the linguistic formalisms, when framed in the context of attention, it is not clear than either direction is preferable.

4 We find that the mean difference in brain decoding score (Pearson's r) between two runs of the same setting (across all settings) is low (0.003), indicating that random initialization does not play a major part in our results. We, therefore, do not carry out more runs.

5 For **Wehbe2014**: second chapter of Harry Potter. For **Pereira2018**: first 500 sentences of English Wikipedia.

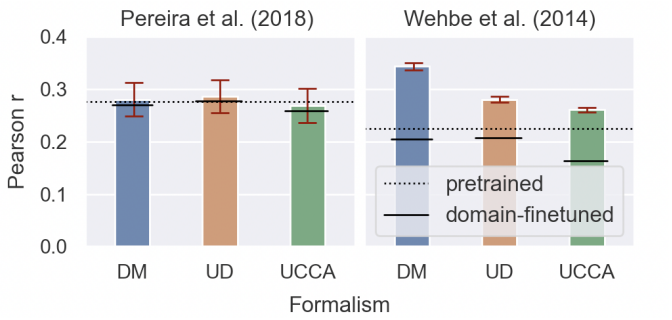


Figure 9.3: Brain decoding score (mean Pearson's r; with 95% confidence intervals shown for subject scores) for models finetuned by MLM with **guided attention** on each of the formalisms, as well the baseline models: *pretrained* BERT (dotted) and *domain-finetuned* BERT (solid).

cording to the Robin Hood Index (Feldbauer et al., 2018). This yields three models for each of the two datasets—one per formalism—for which we present results below.

In addition to the approach described above, we also experiment with directly optimizing for the prediction of the formalism graphs (i.e., parsing) as a way of encoding structural information in LM representations. We find that this leads to a consistent decline in alignment of the LMs' representations to brain recordings.

9.4.5 *Brain Decoding*

To measure the alignment of the different LM-derived representations to the brain activity measurements, brain decoding is performed, following the setup described in Gauthier and Levy (2019).[6] For each subject i's fMRI images corresponding to a set of n sentences or words, a ridge regression model is trained to linearly map from brain activity $B_i \in \mathbb{R}^{n \times d_B}$ ($n = 384$; $d_B = 256$ for **Pereira2018** and $n = 4369$; $d_B = 750$ for **Wehbe2014**) to a LM-derived representation (D_{fr} or D_{LM}), minimizing the following loss:

$$\mathcal{L}(W; \alpha) = \|XW - Y\|_2^2 + \lambda \|W\|_2^2$$

where $W : \mathbb{R}^{d_H \times d_B}$ is a linear map, and λ is a hyperparameter for ridge regularization. Nested 12-fold cross-validation (Cawley and Talbot, 2010) is used for selection of λ, training and evaluation.

EVALUATION To evaluate the regression models, Pearson's correlation coefficient between the predicted and the corresponding heldout

6 Other methods for evaluating representational correspondence such as Representational Similarity Analysis (Kriegeskorte et al., 2008) and the Centered Kernel Alignment similarity index (Kornblith et al., 2019) were also explored but were found to be either less powerful or less consistent across subjects and datasets.

true sentence or word representations is computed. We find that this metric[7] is consistent across subjects and across the two datasets. We run 5000 bootstrap resampling iterations and a) report the mean[8] correlation coefficient (referred to as *brain decoding score/performance*), b) use a paired bootstrap test to establish whether two models' mean (across stimuli) scores were drawn from populations having the same distribution[9], c) apply the Wilcoxon signed rank test (Wilcoxon, 1992) to the by-subject scores to test for evidence of strength of generalization over subjects. Bonferroni correction (for 3 multiple comparisons) is used to adjust for multiple hypothesis testing. See Appendix A.6.2 for details.

9.5 RESULTS

To evaluate the effect of the structurally-guided attention, we compute the brain decoding scores for the guided attention models corresponding to each formalism and fMRI dataset and compare these scores against the brain decoding scores from two baseline models: 1) a *domain-finetuned* BERT (DF), which finetunes BERT using the regular MLM objective on the text of each formalism's training data, and a *pretrained* BERT. We introduce the *domain-finetuned* baseline in order to control for any effect that finetuning using a specific text domain may have on the model representations.

Comparing against this baseline allows us to better isolate the effect of injecting the structural bias from the possible effect of simply finetuning on the text domain. We further compare to a pretrained baseline in order to evaluate how the structurally-guided attention approach performs against an off-the-shelf model that is commonly used in brain-alignment experiments.

9.5.1 *Pereira2018*

Figure 9.3 shows the sentence-level decoding performance on the **Pereira2018** dataset, for the guided attention finetuned models (GA) and both baseline models (domain-finetuned and pretrained).

We find that the DF baseline (shown in Figure 9.3 as solid lines)

leads to brain decoding scores that are either lower than or not significantly different from the pretrained baseline.

7 Appendix A.6.1 shows results for the rank-based metric reported in (Gauthier and Levy, 2019), which we find to strongly correspond to Pearson's correlation. This metric evaluates representations based on their support for contrasts between sentences/words which are relevant to the brain recordings. Other metrics for the evaluation of goodness of fit were found to be less consistent.

8 Across finetuning runs, cross-validation splits, and bootstrap iterations.

9 This is applied per subject to test for strength of evidence of generalization over sentence stimuli.

Specifically, for DM and UCCA, it performs below the pretrained baseline, which suggests that simply finetuning on these corpora results in BERT's representations becoming less aligned with the brain activation measurements from **Pereira2018**.

We find that all GA models outperform their respective DF baselines (for all subjects, $p < 0.05$). We further find that compared to the pretrained baselines, with $p < 0.05$: a) the UD GA model shows significantly better brain decoding scores for 7 out of 8 subjects, b) the DM GA model for 4 out of 8 subjects, c) UCCA GA shows scores not significantly different from or lower, for all subjects. For details see Appendix A.6.2.

9.5.2 *Wehbe2014*

For **Wehbe2014**, where analysis is conducted on the word level, we again find that DF baselines—especially the one finetuned on the UCCA domain text—achieve considerably lower brain decoding scores than the pretrained model, as shown in Figure 9.3.

Furthermore, the guided attention models for all three formalisms outperform both baselines by a large, significant margin (after Bonferroni correction, $p < 0.0001$).

9.6 DISCUSSION AND ANALYSIS

Overall, our results show that structural bias from syntacto-semantic formalisms can improve the ability of a linear decoder to map the BERT representations of stimuli sentences to their brain recordings. This improvement is especially clear for **wehbe2014**, where token representations and not aggregated sentence representations (as in **Pereira2018**) are decoded, indicating that finer-grain recordings and analyses might be necessary for modelling the correlates of linguistic structure in brain imaging data. To arrive at a better understanding of the effect of the structural bias and its relationship to brain alignment, in what follows, we present an analysis of various factors which affect and interact with this relationship.

THE EFFECT OF DOMAIN Our results suggest that the domain of finetuning data and of stimuli might play a significant role, despite having been previously overlooked: simply finetuning on data from different domains leads to varying degrees of alignment to brain data. To quantify this effect, we compute the average word perplexity of the stimuli from both fMRI datasets for the pretrained and DF baselines on each of the three domain datasets.[10] If the domain of the corpora

10 Note that this is not equivalent to the commonly utilised sequence perplexity (which can not be calculated for non-auto-regressive models) but suffices for quantifying the effect of domain shift.

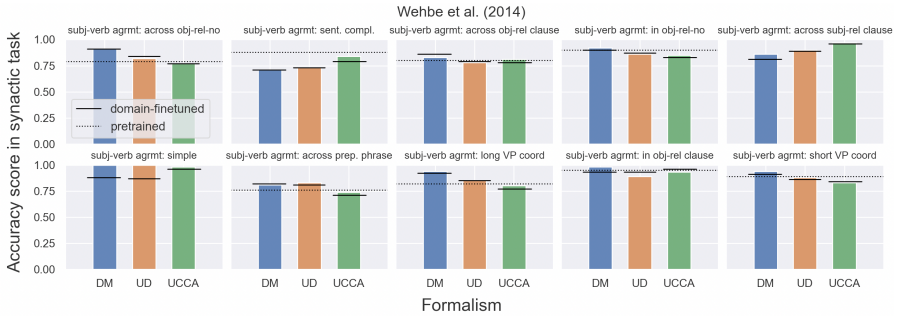


Figure 9.4: Accuracy per subject-verb agreement category of (Marvin and Linzen, 2019) for the three **Wehbe2014** models and each of the four baselines.

used for finetuning influences our results as hypothesized, we expect this score to be higher for the DF baselines. We find that this is indeed the case and that for those baselines (DF), increase in perplexity roughly corresponds to lower brain decoding scores—see details in Appendix A.6.3. This finding calls to attention the necessity of accounting for domain match in work utilizing cognitive measurements and emphasizes the importance of the DF baseline in this study.

TARGETED SYNTACTIC EVALUATION We evaluate all models on a range of syntactic probing tasks proposed by Marvin and Linzen (2019).[11] This dataset tests the ability of models to distinguish minimal pairs of grammatical and ungrammatical sentences across a range of syntactic phenomena. Figure 9.4 shows the results for the three **Wehbe2014** models across all subject-verb agreement (SVA) tasks.[12]

We observe that after GA finetuning: a) the DM guided-attention model, and to a lesser extent the UD guided-attention model have a higher score than the pretrained baseline and the domain-finetuned baselines for most SVA tasks and b) the ranking of the models corresponds to their ranking on the brain decoding task (DM > UD > UCCA).[13] Although all three formalisms annotate the subject-verb-object or predicate-argument structure necessary for solving SVA tasks, it appears that some of them do so more effectively, at least when encoded into a LM by GA.

EFFECT ON SEMANTICS To evaluate the impact of structural bias on encoding of semantic information, we consider Semantic Tagging

11 Using the evaluation script from Goldberg (2019).

12 Results for **Pereira2018** show similar patterns and are included in Appendix A.6.6.

13 For reflexive anaphora tasks, these trends are reversed: the models underperform the pretrained baseline and their ranking is the converse of their brain decoding scores. Reflexive Anaphora, are not explicitly annotated for in any of the three formalisms. We find, however, that they occur in a larger proportion of the sentences comprising the UCCA corpus (1.4%) than those the UD (0.67%) or DM (0.64%) ones, indicating that domain might play a role here too.

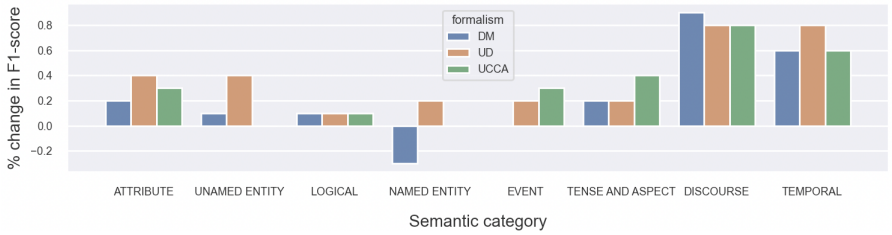


Figure 9.5: Change in F1-score per coarse-grained semantic class compared to the *pretrained* baseline for the three guided attention **Wehbe2014** models.

(Abzianidze and Bos, 2017), commonly used to analyse the semantics encoded in LM representations (Belinkov et al., 2017b; Liu et al., 2019a): tokens are labeled to reflect their semantic role in context. For each of the three guided attention **Wehbe2014** models and the *pretrained* model, a linear probe is trained to predict a word's semantic tag, given the contextual representation induced by the model (see Appendix A.6.4 for details). For each of the three GA models, Figure 9.5 shows the change in test set classification F1-score,[14] relative to the *pretrained* baseline, per coarse-grained grouping of tags.[15]

We find that the structural bias improves the ability to correctly recognize almost all of the semantic phenomena considered, indicating that our method for injecting linguistic structure leads to better encoding of a broad range of semantic distinctions. Furthermore, the improvements are largest for phenomena that have a special treatment in the linguistic formalisms, namely discourse markers and temporal entities. Identifying named entities is negatively impacted by GA with DM, where they are indiscriminately labeled as compounds.

CONTENT WORDS AND FUNCTION WORDS are treated differently by each of the formalisms: UD and UCCA encode all words, where function words have special labels, and DM only attaches content words. Our guided attention ignores edge labels (dependency relations), and so it considers UD and UCCA's attachment of function words just as meaningful as that of content words. Figure A.24 in Appendix A.6.5 shows a breakdown of decoding performance on content and function words for **Wehbe2014**. We find that: a) all GA models and the pretrained model show a higher function than content word decoding score, b) a large part of the decrease in score of two of

14 Note that the test set consists of 263,516 instances, therefore, the margin of change in number of instances here is considerable, e.g. $5652 * 0.6 \approx 40$ instances for the DM and UCCA models on the `temporal` category, which is the least frequent in the test set. See test set category frequencies in the Appendix.

15 The eight most frequent coarse-grained categories from an original set of ten are included—ordered by frequency from left to right; we exclude the `UNKNOWN` category because it is uninformative and the `ANAPHORIC` category because it shows no change from the baseline for all three models.

the three domain-finetuned baselines (UD and DM) compared to the pretrained model is due to content words.

DISPARITY BETWEEN DATASETS While the models finetuned with GA show considerable improvement in brain decoding for **Wehbe2014** (word level analysis), the improvements are much more modest for **Pereira2018** (sentence level analysis). A possible reason for this is the loss of structural information that occurs when aggregating over token representations to construct sentence-level ones. For a more direct comparison, we conduct a sentence-level analysis for the **Wehbe2014** dataset, mean pooling over token hidden-states and their corresponding fMRI time slices[16]. If the advantage of the guided attention models over the baseline drops, this would indicate that mean pooling is at least partially responsible for the lower improvements observed for **Pereira2018**. We find that this is indeed the case: in this setting, decoding scores for the GA models are not significantly different from or lower than the pretrained baseline. Another possible reason for the difference between the two datasets is that the structural bias induced through GA finetuning is more useful when encoding the more diverse naturalistic stimuli used in **Wehbe2014**, compared to **Pereira2018** where the stimuli are short, simple wikipedia-style sentences (see Appendix A.6.7 for examples). Stimuli set perplexity results (Appendix A.6.3) support this hypothesis: compared to the baselines, perplexity after GA finetuning is lower for **Wehbe2014** than is it for **Pereira2018**.

LIMITATIONS When dealing with brain activity, there are many potential confounds, such as the size of fMRI data, the temporal resolution of fMRI, the low signal-to-noise ratio, as well as how the tasks were presented to the subjects, among many other factors. It is, therefore, essential to take sound measures for reporting results, such as cross-validating models, evaluating on unseen test sets, and conducting a thorough statistical analysis. The fMRI data used for both the sentence and word level analyses was recorded while participants read text without performing a specific task.

Although we observe some correlates of linguistic structure, it is possible that uncovering more fine-grained patterns would necessitate brain data recorded while participants perform a targeted task. For future work it would be interesting to investigate if an analysis based on a continuous, naturalistic listening fMRI dataset (Nastase et al., 2020) matches up to the results we have obtained. Regarding the different linguistic formalisms, there are potential confounds such

16 Note that since the **Pereira2018** fMRI recordings are taken and the sentence level and **Wehbe2014** at the 4-gram level, the comparison is still approximate. A possible confound is that averaging over fMRI time slices could also lead to loss of information.

as domain, corpus size[17], and dependency length, (i.e. the distance between words attached by a relation), which depend both on the formalism and on the underlying training set text. Controlling for these confounds using a single corpus annotated for all three formalisms to enable a direct comparison of cognitive plausibility is an avenue for future investigation which could leverage the framework proposed in this work.

CONCLUSIONS We propose a framework to investigate the effect of incorporating specific structural biases in language models for brain decoding.

Using this framework, we present evidence that inducing linguistic structural bias through finetuning using attention guided according to syntacto-semantic formalisms improves brain decoding performance, particularly on a dataset with more complex stimuli. We observe that the models which align most with the brain perform best at a range of syntactic (subject-verb agreement) and semantic tagging tasks, suggesting that language comprehension in the brain, as captured by fMRI recordings, and the tested tasks may rely on common linguistic structure which is partly induced by the added attention constraints. Our results corroborate recent findings on language models not surfacing aspects of semantic structure that are relevant to language comprehension (Hou and Sachan, 2021; Wu et al., 2021), motivating further work on both linguistically-informed training methods and general-purpose training objectives that better capture semantic structure.

Overall, our proposed approach enables the evaluation of more targeted hypotheses about the composition of meaning in the brain, and opens up new opportunities for cross-pollination between computational neuroscience and linguistics.

17 It is interesting to note that decoding score rank for **Wehbe2014** corresponds to finetuning corpus size for the GA models (DM > UD > UCCA), but not the domain-finetuned models. A reasonable conclusion to draw from this is that dataset size might play a role in the effective learning of a structural bias.

Part V

CONCLUSION & OUTLOOK

10 CONCLUSIONS AND OUTLOOK

10.1 CONCLUSIONS

In this dissertation, we present eight studies that each work to further the field's understanding of where ANN language models correspond to or diverge from humans and how they might be used in the study of different aspects of language.

The studies are grouped into three related parts: (i) those that make use of behavioural data, (ii) those that address questions based on linguistic theory, and (iii) those concerned with neural response measurements. Below, we revisit each of the key research questions posed in Chapter 1 and review how the chapters which comprise this thesis contribute to answering them.

BEHAVIOURAL DATA

In Part ii, we describe two studies where representational alignment between LMs and behavioural data was evaluated. Overall, our results indicate that LM representations can reflect some of the patterns of linguistic behaviour.

Are there structural correspondences between eye-tracking fixation patterns and LM representations?

To relate human eye-tracking data patterns with LM representations, in Chapter 3, we leverage the framework of Representational Similarity Analysis, which was developed by neuroscientists for comparing activity patterns from across modalities. In this study, we highlight the utility of RSA for analysing NLP models. Our results show that sentences which are difficult for humans to process have more divergent representations both between the layers of an LM, and between different LMs.

Can language models encode perceptual structure without grounding?

This question relates to an ongoing debate in the field regarding whether it is possible for LMs which are trained on text only to really capture "meaning" (Bender and Koller, 2020a). In Chapter 4, we address it through a case study on color and color terms. The semantics of color terms have long been understood to hold particular linguistic significance, as they are theorised to be subject to universal constraints that arise directly from the neurophysiological mechanisms and properties underlying visual perception. Our results demonstrate

that, even though LMs are trained on textual data only, their representations for color terms do, to a certain extent, show isomorphism to the topology of humans' perceptual color space. Analyzing the differences in alignment across the color spectrum, we also show that warmer colors are, on average, better aligned to the perceptual color space than cooler ones, linking to findings from recent work on efficient communication in color naming. Further analysis based on how efficiently a color is communicated between a speaker and a listener reveals a correlation between lower topological alignment and higher color chip surprisal, suggesting that the kind of contexts a color occurs in play a role in determining alignment.

LINGUISTIC THEORY

Part iii contains three studies that examine LMs through the lens of linguistic theory. Overall, the results presented in this part of the thesis show that LMs conform to some (but not all) of the prognoses made by linguistic theory regarding the syntactic structure, word order, and robustness to perturbations.

How sensitive are humans and LMs to linguistic perturbations of Winograd Schema Challenge examples?

The sentence processing mechanisms of humans are known to be robust towards a variety of perturbations (Ferreira et al., 2002; Gibson et al., 2013). Chapter 5 presents a comparison of the robustness of several LMs to that of humans. A dataset is constructed by applying perturbations that span various linguistic dimensions (tense, gender, voice, etc.) to Winograd Schema Challenge examples. Testing LMs and human annotators using this dataset shows that the latter are overall more robust than the former, but that when finetuned on task-specific data LMs become more robust.

Can the attention patterns learned by LMs reflect linguistic structure (in the form of dependency trees)?

In Chapter 6, we investigate the attention patterns learned by a multilingual LM for whether they encode linguistic structure. The results of experiments carried out across 18 languages show that full dependency trees can indeed be decoded with above baseline accuracy from single attention heads, and that individual relations are often tracked by the same heads across languages. Further experiments show that finetuning the multilingual LM with a supervised dependency parsing objective leads to its attention mechanism resembling dependency tree structure to a considerably larger extent.

Why do LMs trained on sentences with shuffled words still perform well on Natural Language Understanding tasks? Do they still encode some word

order information? Are there tasks where degrading word order information has a stronger effect on performance?

Chapter 7 examines recent findings showing that LM performance on NLU benchmarks (like GLUE) is only marginally affected by shuffling the order of words within their training or finetuning data. Our analysis reveals that these models, through their position encodings, retain a modicum of word order information when trained on shuffled data. We find this to be due to (a) word segmentation being carried out after rather than before shuffling, leading to a locality bias being learned by the position encodings and (b) various other statistical cues like correlations between sentence lengths and token distributions. In experiments with more rigorous benchmarks, we show that there exist NLU tasks that are more sensitive to the degradation of word order information.

NEURAL RESPONSE MEASURMENTS

In two Chapters, Part iv presents an overview of work connecting LMs with neurobiology via recordings of neural response, and then introduces a use-case for them in the study how meaning is composed in the brain.

In Chapter 8, we conduct a literature review of work linking computational models of language and neural response measurements. In this review, we trace a line from early research using simple language models (context-free grammars, etc.) to contemporary studies employing ANN models and neural response recordings from multiple modalities. Overall, the review showcases ANN models' potential to both contribute to and benefit from work investigating the neurobiological underpinnings of language.

Can LMs be used to enable the evaluation of targeted hypotheses about the composition of meaning in the brain?

Finally, in Chapter 9, we introduce a framework where LMs can be used for the evaluation of targeted hypotheses about the composition of meaning in the brain. Leveraging this framework, we demonstrate that when the attention mechanisms of LMs are biased to match structures from three syntaco-semantic formalisms, their representations align better with brain recordings.

10.2 OUTLOOK

The research presented in this dissertation has contributed to furthering our knowledge of where artificial neural network language models agree with or diverge from what we know about language processing in humans. Subsequently, we presented work where these

models are employed in the evaluation of possible hypotheses about how word meanings are composed in the brain.

As the field gradually improves its understanding of ANN LMs, they are likely to feature more prominently within core research in theoretical, psycho-, and neuro- linguistics, functioning as tools for theory-building and hypothesis formulation or testing (Baroni, 2021; Futrell et al., 2019; Marblestone et al., 2016). We argued in Chapter 8 that the prospering domains of interpretability and explainability (Belinkov and Glass, 2019a; Danilevsky et al., 2020) are likely to play an important role in facilitating this, enabling a fine-grained dissection and analysis of network components and behaviour. Additionally, we believe that research pushing towards more human-like training settings and objectives, such as the work on developing multi-modal or embodied models, also has an important part to play in making ANN LMs into more plausible models of language (Bisk et al., 2020b; McClelland et al., 2020).

By enabling the modelling of complex, naturalistic linguistic tasks and providing a framework for the simultaneous linking of computation, behaviour, and brain function (Schrimpf et al., 2020c), we are ultimately optimistic that this class of models can meaningfully contribute to the study of language.

Part VI

APPENDIX

A

APPENDIX

A.1 CHAPTER 3

A.1.1 *Correlation Heatmaps*

Figure A.1 shows correlation heatmaps between disagreement among layers i and j $V_{Corr_{L_i - L_j}}$ and each of $V_{firstpass}$, $V_{wordSense}$ and $V_{logFreq}$.

A.2 CHAPTER 4

A.2.1 *List of included color terms*

Red, green, maroon, brown, black, blue, purple, orange, pink, yellow, peach, white, gray, olive, turquoise, violet, lavender, and aqua.

A.2.2 *RSA between models*

Figure A.2 shows a the result of representation similarity analysis between the representations derived from all models (and configurations) as well as CIELAB, showing Kendall's correlation coefficient between flattened RSMs.

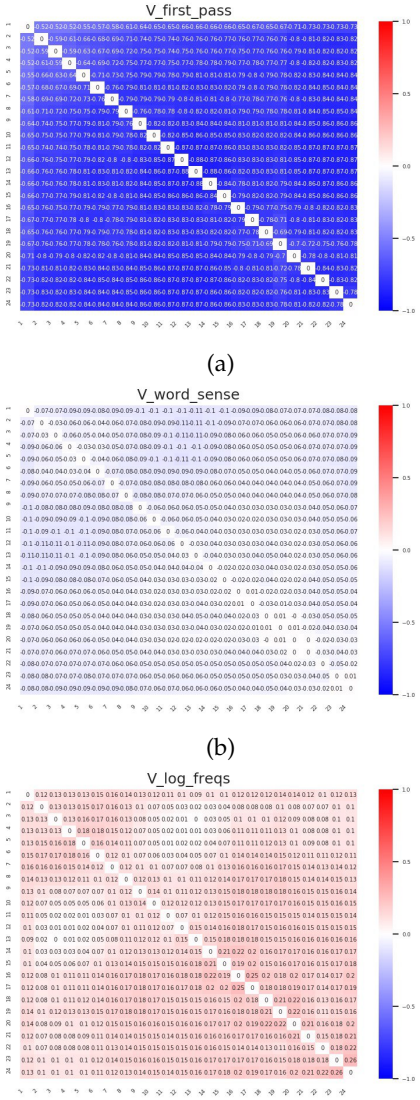
A.2.3 *Representation Similarity Matrices*

Figures A.3 to A.6 show the representation similarity matrices employed for the RSA analyses, for the layer with the highest RSA score from each of the controlled-context (CC) models.

A.2.4 *Warm vs. Cool colors*

Figures A.7 and A.8 show Linear Mapping and RSA results broken down by color temperature. The color space is split according to temperature measured according to the Hue dimension in the Hue-Value-Saturation space[1].

1 https://psychology.wikia.org/wiki/HSV_color_space



(a)

(b)

(c)

Figure A.1: RSM heatmaps showing (Spearman's ρ) correlation between disagreement among layers i and j ($V_{Corr_{L_i - L_j}}$) and **(a)** $V_{firstpass}$ (top), **(b)** $V_{wordSense}$ (middle) and, **(c)** $V_{logFreq}$ (bottom).

Figure A.2: Result of representation similarity analysis between all models (and configurations), showing Kendall's correlation coefficient between flattened RSMs. Results are shown for layers which are maximally correlated with CIELAB, per model. `-rc` indicates **random-context**, `-cc` indicates **controlled-context**, and `-nc` indicates **non-context**.

A.2.5 *Corpus statistics*

Figures A.9 and A.10 show log frequency and entropy of distributions over part-of-speech categories, dependency relations, and lemmas of dependency tree heads of color terms in common crawl.

A.2.6 *Linear mapping results by munsell color chip*

Figure A.11 shows linear mapping results broken down by Munsell chip for all models and configurations.

A.2.7 *Linear mapping control task and probe complexity*

Figure A.12 shows the full results over a range of probe complexities for the standard experimental condition as well the random control task.

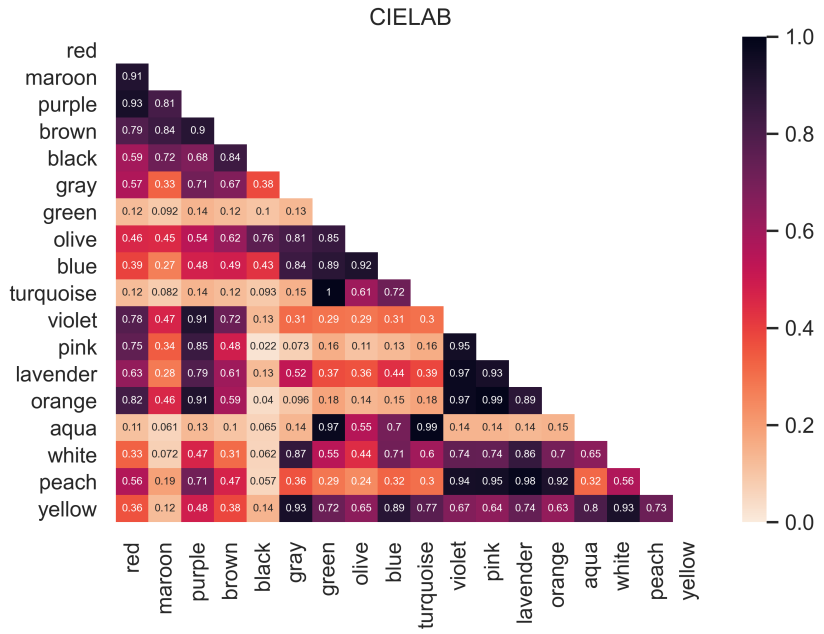


Figure A.3: CIELAB RSM

A.2.8 *Dimensionality of color subspace*

Figure A.13 shows the proportion of explained variance with respect to the number of dimensions which are assigned 95% of the linear regression coefficient weights.

A.2.9 *Effect of model size*

Table A.1 shows the RSA and linear mapping (selectivity) results for four BERT models: BERT-mini (4 layers, hidden size: 256), BERT-small (4 layers, hidden size: 512), BERT-medium (8 layers, hidden size: 512), and BERT-base (12 layers, hidden size: 768). Model specification and training details for the first three can be found in Turc et al. (2019) and for last in Devlin et al. (2019a).

A.2.10 *Linear Mixed Effects Model*

To fit Linear Mixed Effects Models, we use the LME4 package. With model type (BERT-CC, RoBERTa-NC, etc.) as a random effect, we follow a step-wise model construction sequence which proceeds along four levels of nesting: (i) in the first level color log-frequency is the only fixed effect, (ii) in the second `pmi-colloc` is added to that, (iii) in the third, each of `pos-ent, deprel-ent, head-ent` is added separately to the a model with log frequency and `pmi-colloc`, (iv) the term that leads to the best fit from the previous level `deprel-ent` is

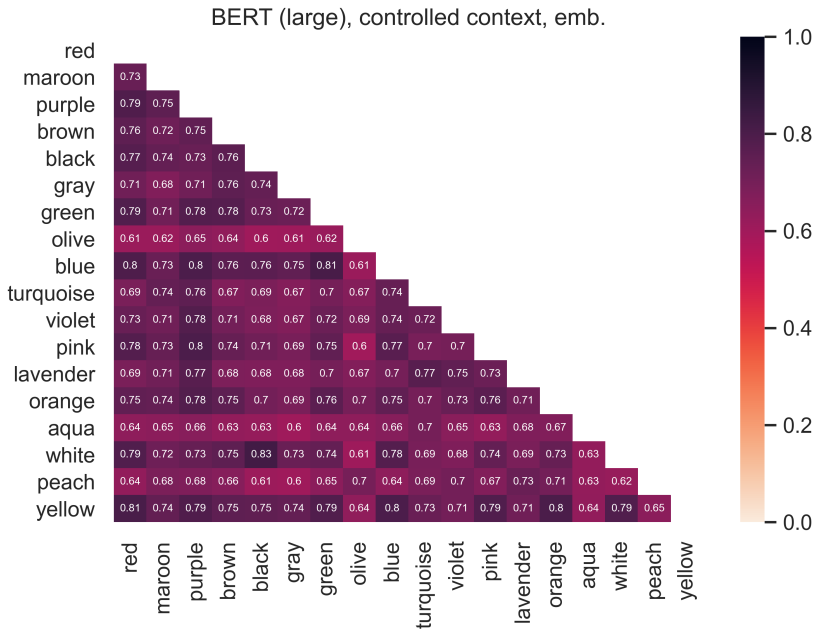


Figure A.4: BERT(CC) RSM

Model	RSA max	RSA mean	lin. map.. max	lin. map. mean
BERT-mini	0.077	0.043 ± 0.340	0.729	0.582 ± 0.291
BERT-small	0.106	0.070 ± 0.191	0.734	0.598 ± 0.294
BERT-medium	0.097	0.057 ± 0.035	0.739	0.654 ± 0.221
BERT-base	0.162*	0.092 ± 0.058	0.740	0.677 ± 0.182

Table A.1: Results for the four smaller BERT models. RSA results (left) show max and mean (across layers) Kendall's correlation coefficient (τ). Correlations that are significantly non-zero are indicated with: * : $p < 0.05$. Results for the Linear Mapping experiments (right) show max and mean selectivity. Standard deviation across layers is included with the mean results.

included, then each of the proportion terms `adj-prop`, `amod-prop`, `cop-prop` is added. The reported regression coefficients are extracted from the minimal model containing each term.

A.3 CHAPTER 5

A.3.1 *Observations on original dataset*

1. A few of the original examples were of unorthodox design: for instance, consider the pair:

   (1) a. Look! There is a minnow swimming right below that duck! It had better get away to safety fast!

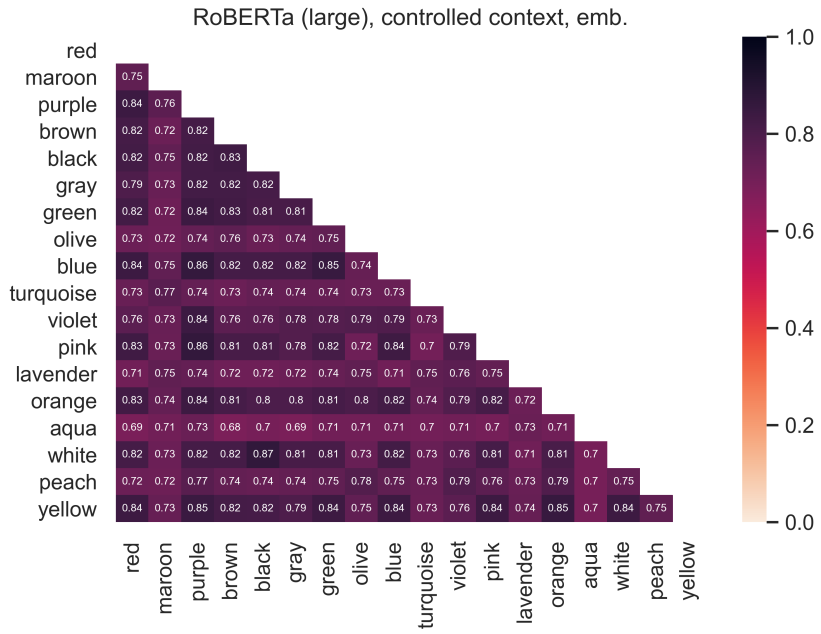


Figure A.5: RoBERTa(CC) RSM

b. Look! There is a shark swimming right below that duck! It had better get away to safety fast!

Here, instead of having a discriminatory segment select which of the two nouns could be the antecedent, one of the nouns is switched out with another.

2. Example 90 has a typo in the question where Kamchatka is spelled as 'Kamtchatka'.

A.3.2 *Human Judgements*

Table A.2 shows the proportion of instances for which all three annotators agreed and the average time required by annotators for the original examples and each of the perturbed datasets. Figure A.14 shows the Amazon Mechanical Turk template used. The annotator pool was restricted to native speakers of English located in the United States who were classified by Mturk as 'masters' and had a HITs approval rate above 99%.

A.3.3 *Pointwise Mutual Information*

We compute unigram Pointwise Mutual Information statistics using the Hyperwords[2] package (Levy et al., 2015). If a corpus is split into a collection D of words W and their contexts C, we can compute co-

2 https://bitbucket.org/omerlevy/hyperwords/

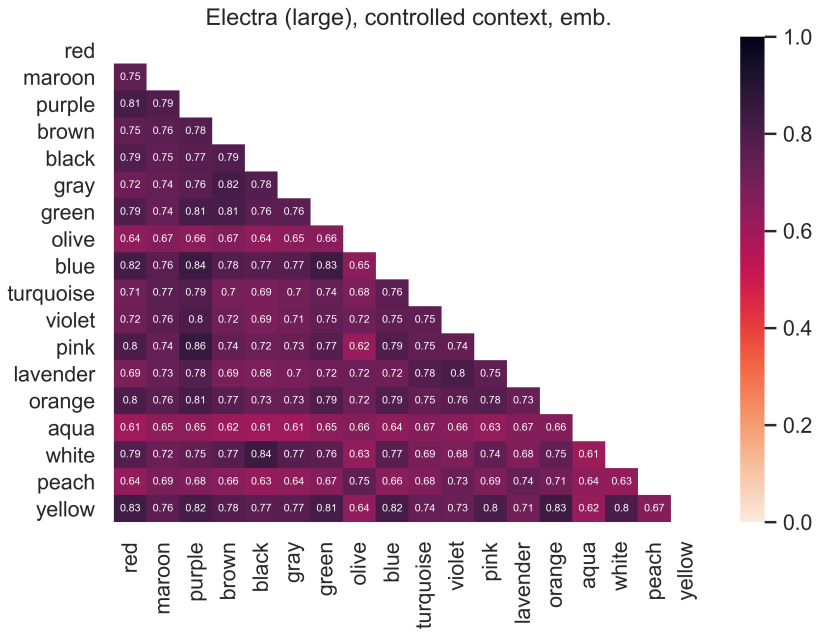


Figure A.6: ELECTRA(CC) RSM

occurrence counts for each pair of $w \in W$ and $c \in C$. PMI is then defined as the log-ratio between the joint probability of w with c and the product of their marginal probabilities. Refer to Levy et al. (2015) for further details. For generating a collection D of word-context pairs, we use the following hyperparameter settings: a minimal word count of 200 for being in the vocabulary, a context window size of 6, dynamic context windows, positional contexts (where each context is a conjunction of a word and its relative position to the target word).

A.3.4 *Confirming Solvability*

Table A.3 shows the breakdown by perturbation type of the expert annotations which were gathered for examples that were annotated incorrectly by the Mechanical Turk workers.

A.3.5 *Notes on construction of perturbed dataset*

TENSE SWITCH (TEN) Examples 168–172 could not be changed while maintaining the semantics of the instance intact.

RELATIVE CLAUSE INSERTION (RC) The pre-selected set of 19 templates is shown below:

- "who we had discussed __"
- "who he had discussed __"

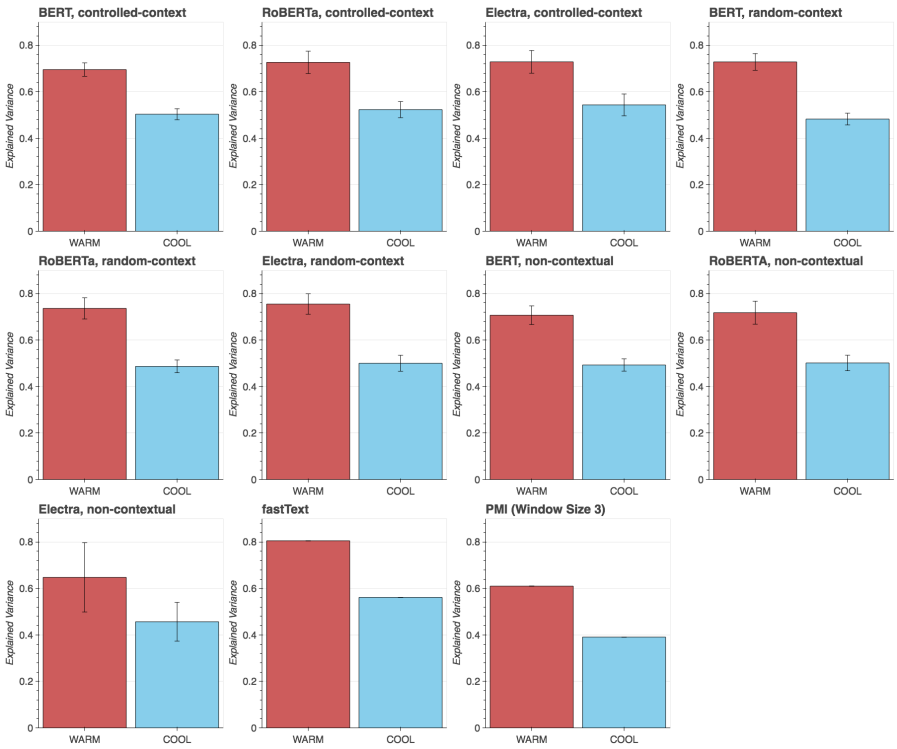


Figure A.7: Linear mapping results (proportion of explained variance) broken down by color chip temperature for each of the baselines and the LMs.

- “who she had discussed __"
- “who you had discussed __"
- “which we had seen __"
- “which he had seen __"
- “which she had seen __"
- “which you had seen __"
- “who we know from __"
- “who he knows from __"
- “who she knows from __"
- “who you know from __"
- “that is mentioned in __"
- “that is located at __"
- “that is close to __"
- “that is known for __"

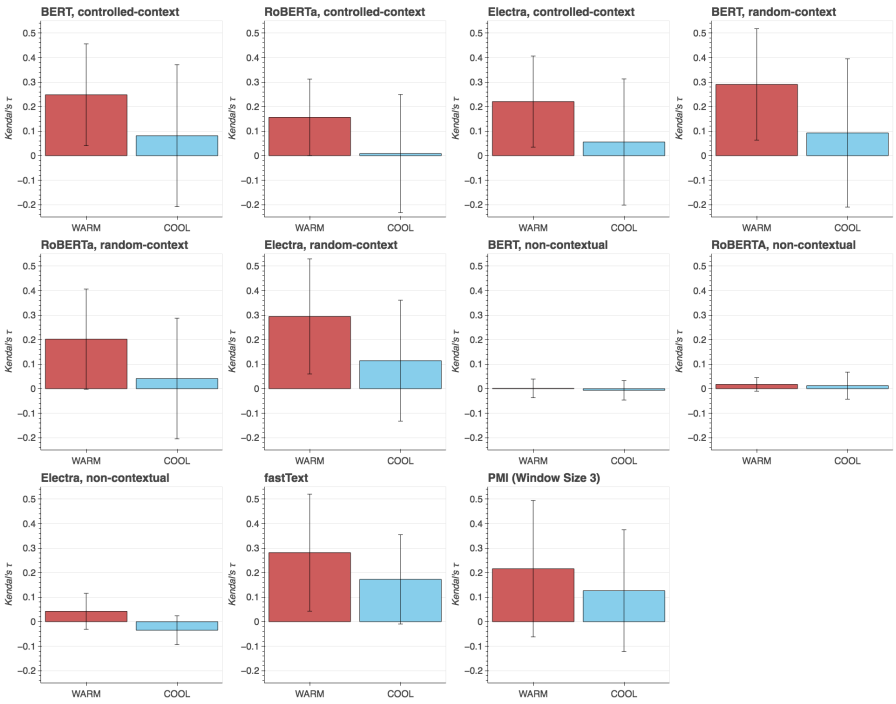


Figure A.8: RSA results (Kendall's τ) broken down by color temperature for each for each of the baselines and the LMs.

- "which had been __",
- "who you met __"
- "that is __"
- "which was put there __"

SYNONYM / NAME SUBSTITUTION (SYN / NA) No appropriate synonyms were found for *tide* and *wind* in examples 130 and 131.

ADVERBIAL QUALIFICATION (ADV) Two instances (95 and 96) in which the main verb was already modified were excluded.

A.3.6 *Referent preferences*

Table A.4 shows the percentage of examples in the switchable subset of the datasets where the second referent in the text was assigned a higher probability than the first, for both the original and reversed referent order.

A.3.7 *Effect of perturbations*

NUCLEUS SAMPLING Table A.5 shows the average number of vocabulary items kept after Nucleus sampling with $p = 0.9$ is applied.

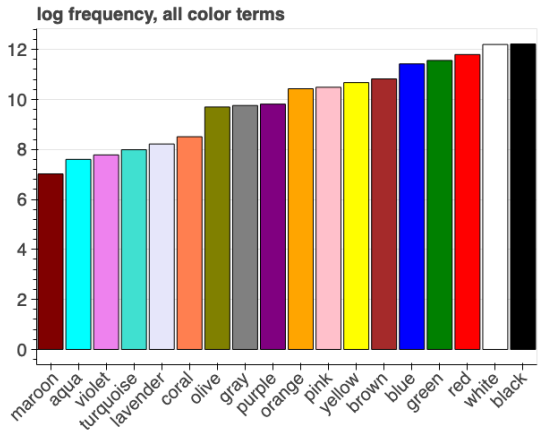


Figure A.9: Log frequency of color terms in common crawl.

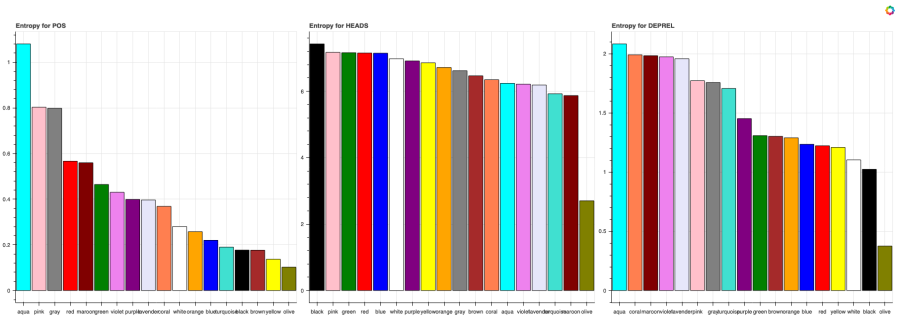


Figure A.10: Entropy of distributions over part-of-speech categories, dependency relations, and lemmas of dependency tree heads of color terms in common crawl.

PROBABILITY SHIFT is defined as the difference in the probability of a candidate before and after a perturbation is applied. Figure A.15 shows the difference in average probability shift between the correct candidates and the incorrect candidates for each of the models per perturbation type. This provides a view that is meaningfully different from accuracy, as the probability of a candidate can shift without exceeding the threshold required to change a model's prediction. We find that there is a general trend of the incorrect candidates becoming more likely relative to the correct ones. This can be seen as confirming that, on average, nearly all perturbations make the problems more difficult for all models.

HIDDEN STATE REPRESENTATION DISTANCE is used to provide a more holistic view of the correspondence between the representations derived for the different perturbations. The analysis is conducted on the 128 examples which are common between all datasets. A representation is derived for each example by taking the max-pool of hidden-state representations of a model's final layer. For each of the

Pert.	Full Agreement	Avg. Time
ORG	82.45	15.32
TEN	82.91	16.39
NUM	83.00	19.56
GEN	78.06	19.24
VC	82.72	17.02
RC	82.68	17.83
ADV	82.68	17.69
SYN/NA	82.45	15.26

Table A.2: Annotation statistics: Proportion of examples with full agreement and average time required for answering in seconds.

Counts	All	Ambig.	Non-Ambig.	Correct
TEN	9	0	9	8
NUM	14	2	12	9
GEN	12	2	10	10
VC	17	3	14	12
RC	25	1	24	13
ADV	13	0	13	11
SYN/NA	9	2	7	4

Table A.3: Breakdown of solvability annotation counts by perturbation. **Ambig.** indicates the count of examples labeled as Ambiguous, **Non-Ambig.** is the number of remaining examples. **Correct** indicates the number of those which is solved correctly.

seven perturbations p, we compute pairwise correlation distance[3] between each pair of original and perturbed example representations yielding a vector $\vec{D_p} \in \mathbb{R}^{128}$. The mean of $\vec{D_p}$ is then computed as an aggregate measure of the distance between the representations derived from a perturbation p and the original o. Figure A.16 shows a plot of this for all perturbations for each of the models.

A.3.8 *Candidate probability correlations*

Figure A.17 shows the average correlation between a candidate's probability when it is the correct referent and when it is not.

3 This is preferable to other distance measures as it normalizes both the mean and variance of activity patterns over experimental conditions.

Pert.	Original	Reversed
ORG	66.90	70.42
TEN	62.38	65.14
NUM	60.16	56.10
GEN	72.17	75.65
VC	38.14	39.83
RC	63.57	68.57
ADV	68.08	70.92
SYN/NA	59.12	64.23

Table A.4: Percentage of examples in switchable subset with probabilities assigned to the second referent in the text rather than the first, for both the original and reversed referent order.

A.4 CHAPTER 6

A.4.1 *Positional Scores Per Offset*

A.4.2 *Decoding UUAS Across Relations*

A.4.3 *Full Parsing Scores*

A.5 CHAPTER 7

A.5.1 *Subword vs. word scrambling*

A.5.2 *On biased sampling*

We first split our vocab of size 5,000 into two halves, both of size 2500, such that the sum total of unigram frequencies of tokens in each half is roughly equivalent. Next, iterating over 100k BookCorpus sentences, we determine the sentence length l, for which there are an equivalent number of tokens in sentences with length $< l$ and sentences with length $>= l$. We then sample tokens from the first vocab half for sentences $< l$, and from the second vocab half for sentences with length $>= l$, 80% of the time; for the other 20%, we sample from the opposite half to introduce some overlap.

Perturbation	BERT	RoBERTa	XLNet	BERT+WW	RoBERTa+WG
ORG	19.81	203	1.26	1.07	1021.44
TEN	23.88	165.84	1.26	1.09	947.53
NUM	90.35	341.05	1.57	1.30	1087.78
GEN	18.11	128.37	1.44	1.19	1039.84
VC	41.88	154.21	1.28	1.09	961.04
RC	21.02	97.35	1.35	1.14	952.09
ADV	17.01	145.35	1.23	1.10	1004.14
SYN/NA	31.50	199.26	1.39	1.11	1055.71
VOCAB. SIZE	30522	50265	32000	30522	50265

Table A.5: Average number of vocabulary items left after probability distribution truncation with $p = 0.9$ is applied.

A.5.3 *Full UD results*

A.6 CHAPTER 8

A.6.1 *Mean/median rank results*

Table A.6 shows results for the Pearson's r metric reported in the main paper, alongside the mean/median rank metrics reported in gauthier2019linking, which give the rank of a ground-truth sentence or word representation in the list of nearest neighbors of a predicted representation, ordered by increasing cosine distance. This metric evaluates representations based on their support for contrasts between sentences/words which are relevant to the brain recordings. The table shows that models with higher Pearson r scores, also have a lower average ground truth word/sentence nearest neighbour rank i.e. induce representations that better support contrasts between sentences/words which are relevant to the brain recordings.

A.6.2 *Significance testing*

BOOTSTRAPPING The bootstrapping procedure is described below. For each of m subjects:

1. There are n stimuli sentences/words, corresponding to n fMRI images. A linear decoder is trained to map a recording to its corresponding LM-extracted representation. This is done using 12-fold cross-validation and yields a 'predicted representation' per stimulus.

Table A.6: Brain decoding scores as measured via three metrics — Pearson's r, Mean rank, and Median Rank — for each of the domain-finetuned baseline (DF-B) models, the guided attention models (GA), and the pretrained (PRE) model.

Model	Pearson's r	Mean rank	Median rank
	Wehbe2014		
DF-B DM	0.204	493.11	89.32
DF-B UD	0.206	497.24	81.69
DF-B UCCA	0.164	689.89	227.30
GA DM	0.343	172.45	10.96
GA UD	0.280	255.127	18.28
GA UCCA	0.261	315.73	25.78
PRE	0.225	436.70	53.13

2. To compensate for the small size of the dataset which might lead to a noisy estimate of the linear decoder's performance, we now randomly resample n datapoints (with replacement) from the full n datapoints.

3. For each resampling, our evaluation metrics (pearson's r, mean rank, etc.) are computed between the sampled predictions and their corresponding 'gold representations', for all sets of LM reps. We store the mean metric value (e.g. pearson r score) across the n 'sampled' datapoints. We run 5000 such iterations. This gives us 5000 such paired mean scores (across the n samples, that is) for all models.

4. When comparing two models, e.g. **GA DM** vs. **PRE**, to test our results for strength of evidence of generalization over stimuli, we compute the proportion of these 5000 paired samples where e.g. **GA DM**'s mean sample score is greater than **PRE**. After Bonferroni correction (*bc*) for multiple hypothesis testing, this is the p-value we report. For **Wehbe 2014**, comparisons between each of the GA models and the pretrained baseline lead to $p = 0.000$ (i.e. The GA model mean score is greater than the pretrained baseline's mean score for all 5000 sets of paired samples), for all subjects. We, therefore, do not include a similar table.

5. We average over the 5000 samples per subject, and use these m subject means for the across-subject significance testing, described below.

STRENGTH OF GENERALIZATION ACROSS SUBJECTS To test our results for strength of generalization across subjects, we apply the Wilcoxon signed rank test (Wilcoxon, 1992) to the m by-subject mean

scores (see above), comparing the GA models to the pretrained baselines. Since $m = 8$ for both datasets, the lowest p-value is 0.0078 (if every subject's difference score consistently favors the GA model over the baseline or vice versa). In the case of **Wehbe 2014**: all comparisons yield a p-value of 0.0078 (0.045 after *bc*), where the GA model > the pretrained baseline.

A.6.3 *The Domain effect*

Table A.7 shows average word perplexity scores for the pretrained model and the domain-finetuned models for each of the three text domains on the stimuli from **Wehbe2014**. Scores are averaged over the words in a sentence and the sentences (stimuli) in the datasets.

Table A.7: Average word perplexity for the domain-finetuned baseline (DF-B) models, the guided attention models (GA), and the pretrained (PRE) model.

Model	**Wehbe2014**
PRE	34.79
DF-B DM	36.11
DF-B UD	38.41
DF-B UCCA	40.45
GA DM	33.24
GA UD	37.16
GA UCCA	33.60

A.6.4 *Semantic Tagging*

PROBING DETAILS Representations for the probing task are derived for each sentence in the dev and test sets from Abzianidze and Bos (2017). The dev set is employed as a training set, because it is mostly manually annotated/corrected (as opposed to the much noisier training set) and because it is already possible to train rather accurate semantic taggers which suffice for our analysis with a training set of that size (131337 instances). We report results for the official test set. Table A.8 shows the frequency of each semantic tag we report scores for in the test set. An L2 regularised logistic regression model is used.

DISCUSSION We observe the largest improvements for the `DISCOURSE` and `TEMPORAL` categories. The former involves identifying subordinate, coordinate, appositional, and contrast relations. These relations are highly influenced by context, and correctly classifying them can often be contingent on longer dependencies, which the structural bias increases 'awareness' of. The `TEMPORAL` category, on the other hand, consists of tags such as `clocktime` or `time of day` which are applied to multi-word expressions, e.g *27th December*. Highlighting these dependencies by assigning more attention weight between their subparts is likely helpful for their accurate identification.

Table A.8: Semantic tag frequency in the test set.

Category / Frequency	
Attribute	63763
Unamed Entity	48654
Logical	32973
Named Entity	29271
Event	25338
Tense and Aspect	15208
Discourse	9948
Temporal	5652

A.6.5 *Content words and function words analysis*

Figure A.24 shows the breakdown of brain decoding score by content and function words for **Wehbe2014**. We consider content words as words whose universal POS according to spaCy is one of the following: {ADJ, ADV, NOUN, PROPN, VERB, X, NUM}. Out of a total of 4369, 2804 are considered as content and 1835 as function words.

A.6.6 *Targeted Syntactic Evaluation Scores*

.

A.6.7 *Stimuli examples*

The **Wehbe2014** stimuli set consists of 384 sentences from chapter 9 of *Harry Potter and the Sorcerer's Stone*. As would be expected from naturalistic text, the sentences show a range of variance in complexity with both simples sentences such as the following:

- *Blood was pounding in his ears.*
- *Harry grabbed his broom.*
- *Harry ignored her.*

- *The same thought seemed to have struck Malfoy.*
- *It was dinnertime.*

And longer, substantially more complex ones such as:

- *He didn't have a clue what was going on, but he didn't seem to be being expelled, and some of the feeling started coming back to his legs.*
- *He leaned forward and pointed his broom handle down — next second he was gathering speed in a steep dive, racing the ball — wind whistled in his ears, mingled with the screams of people watching — he stretched out his hand — a foot from the ground he caught it, just in time to pull his broom straight, and he toppled gently onto the grass with the Remembrall clutched safely in his fist.*
- *Perhaps brooms, like horses, could tell when you were afraid, thought Harry; there was a quaver in Neville's voice that said only too clearly that he wanted to keep his feet on the ground.*
- *Harry had heard Fred and George Weasley complain about the school brooms, saying that some of them started to vibrate if you flew too high, or always flew slightly to the left.*
- *Harry hadn't had a single letter since Hagrid's note, something that Malfoy had been quick to notice, of course.*

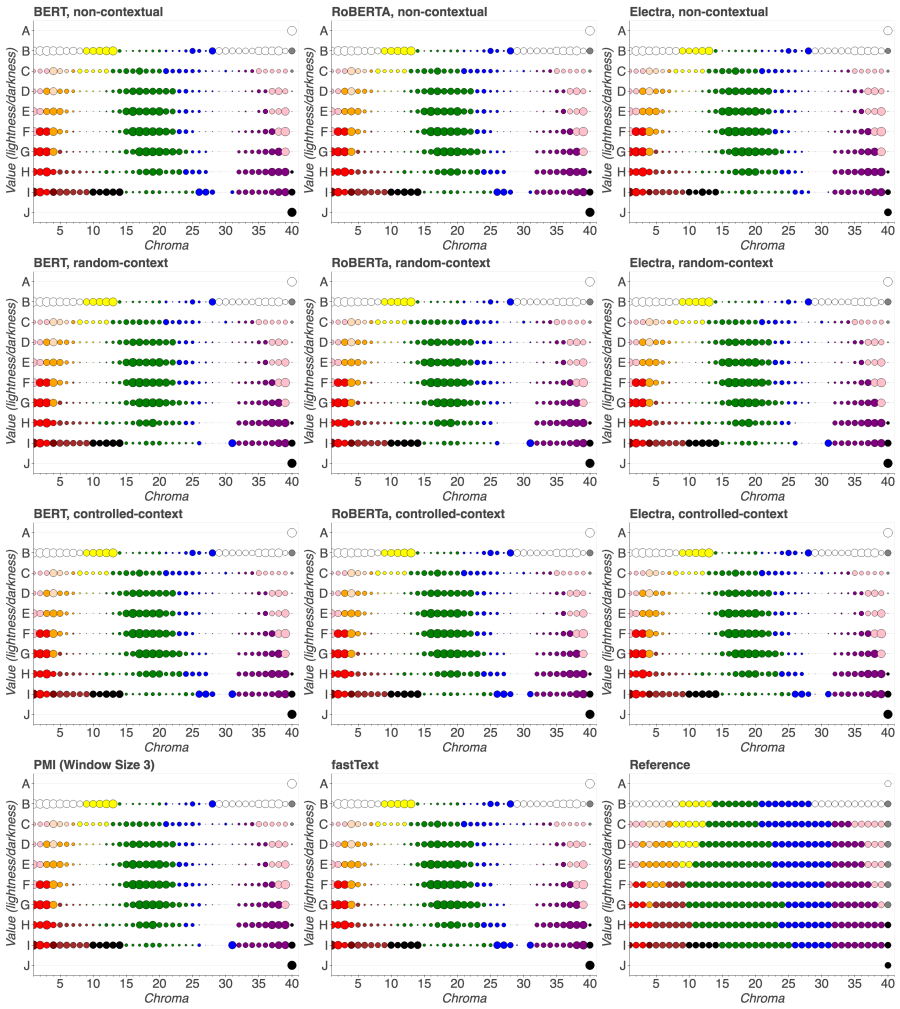


Figure A.11: Linear mapping results for each of the baselines and language models, under all extraction configurations, broken down by Munsell color chip. Each circle on the chart represents the ranking of the predicted color chip when ranked according to Pearson distance (1− Pearson's r) from gold – the larger the circle, the higher (better) the ranking. Circle colors reflect the modal color term assigned to the chips in the lexicon. Reference plot showing modal color of all chips also included.

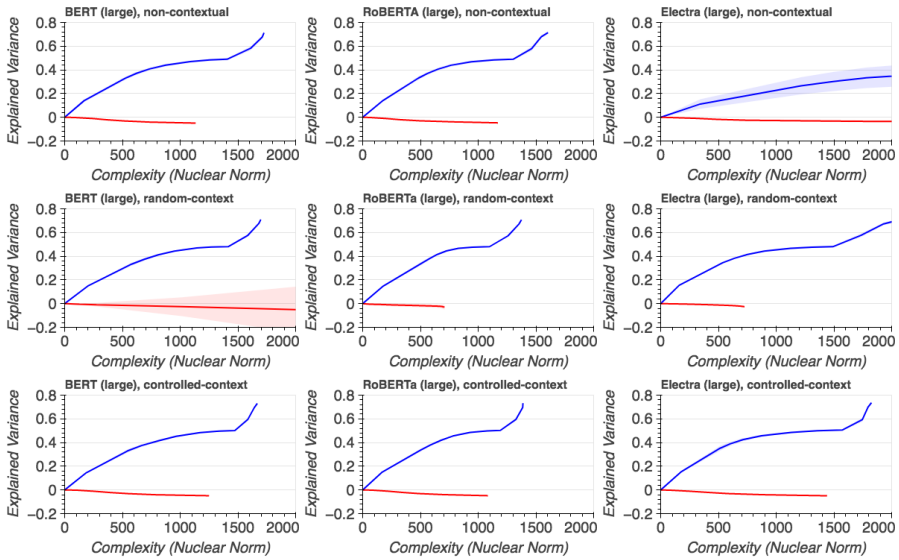


Figure A.12: Explained variance for the linear probes trained on the normal experimental condition (blue) and the control task (red) where color terms are randomly permuted. The means are indicated by the lines and standard deviation across layers is indicated by the bands.

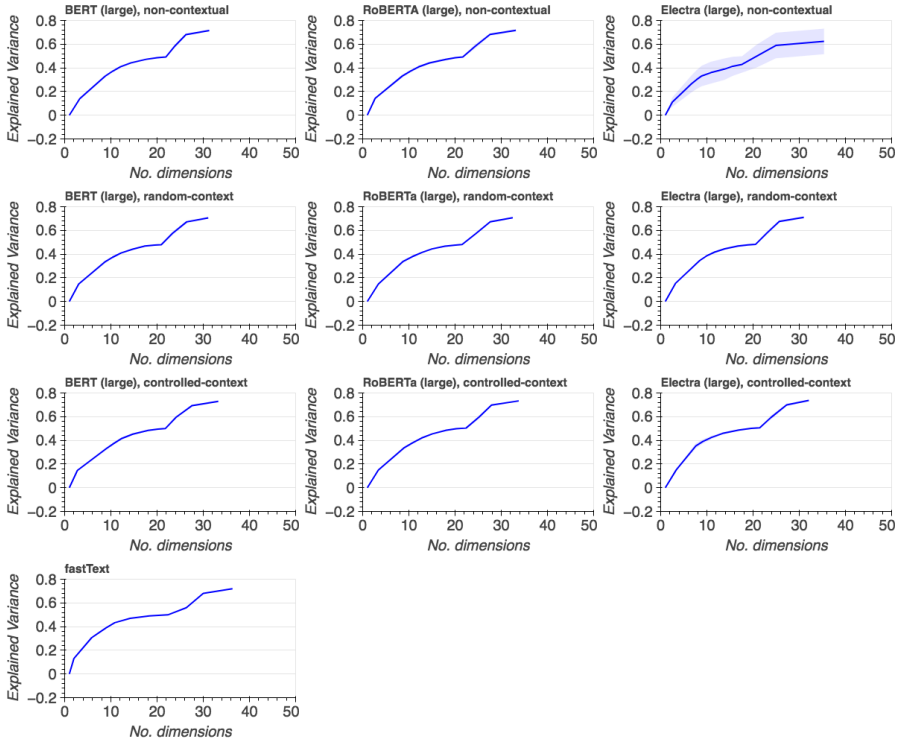


Figure A.13: The y-axis shows explained variance for the linear probes. The means are indicated by the lines and standard deviation across layers is indicated by the bands. The x-axis shows the number of regression matrix coefficients assigned 95% of the weight.

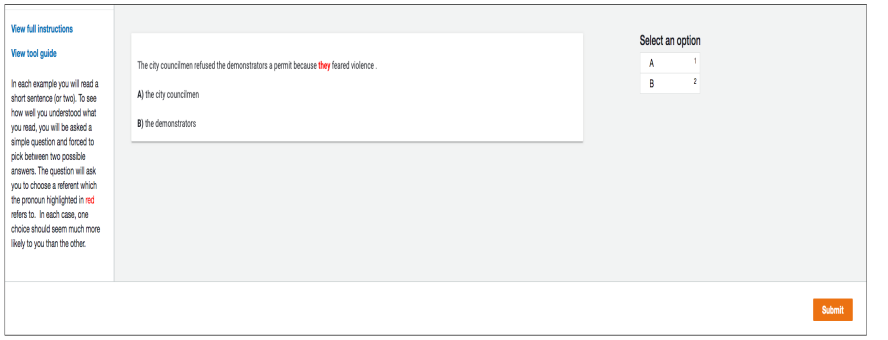


Figure A.14: Sample of Mturk template shown to annotators.

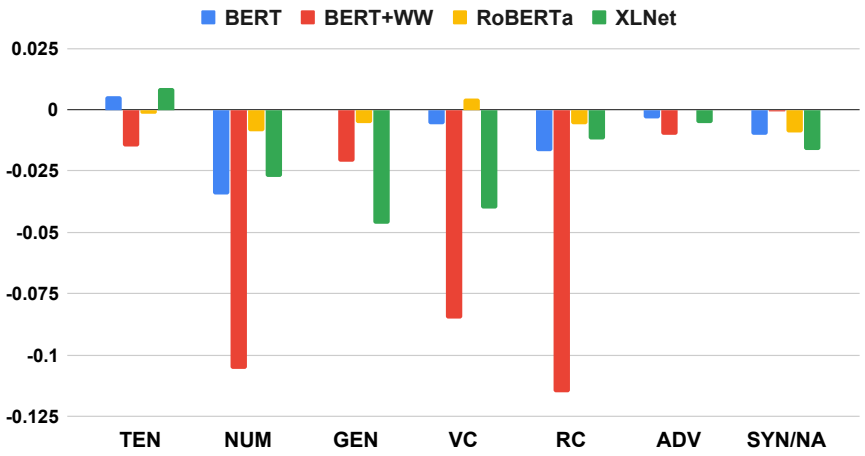


Figure A.15: The difference between average probability shift for the correct and the incorrect referents per perturbation. Y-axis values above zero mean the correct referent became more likely on average after a perturbation and vice versa.

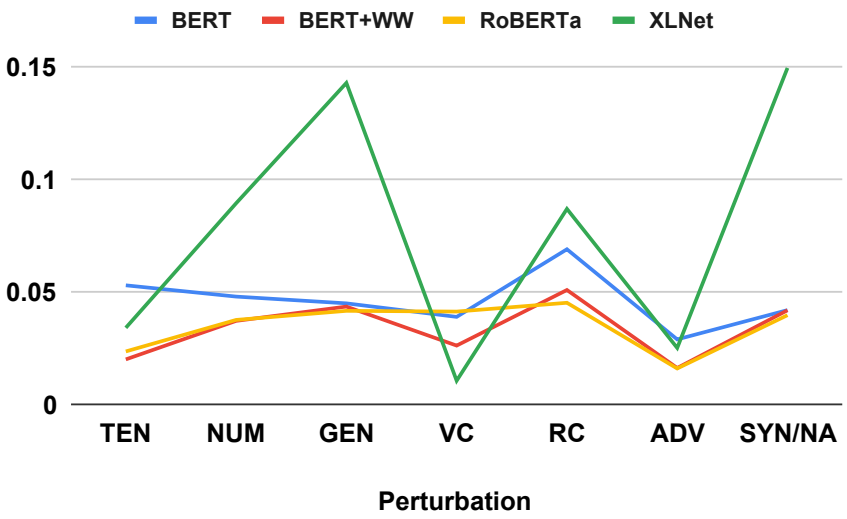


Figure A.16: The correlation of pronoun hidden state representation distance from the original for each perturbation.

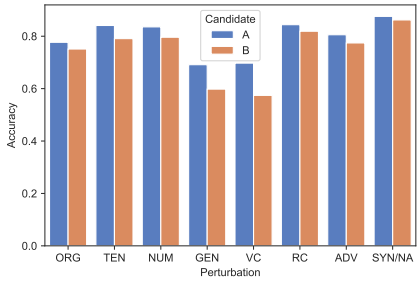


Figure A.17: Correlation (Spearman's ρ) between the probability of a candidate when it is the correct candidate and when it is the incorrect one. Candidates A and B are the first and second candidates in a WSC instance.

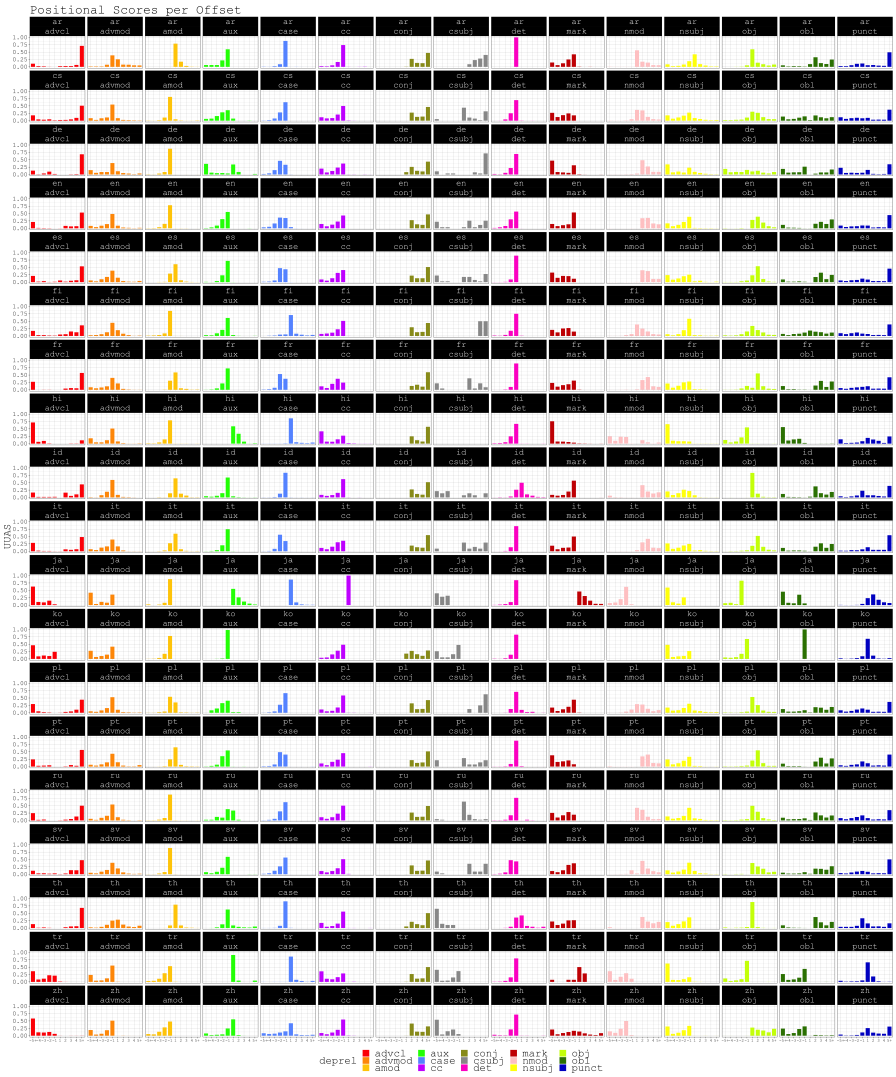


Figure A.18: Positional scores across relations for all languages.

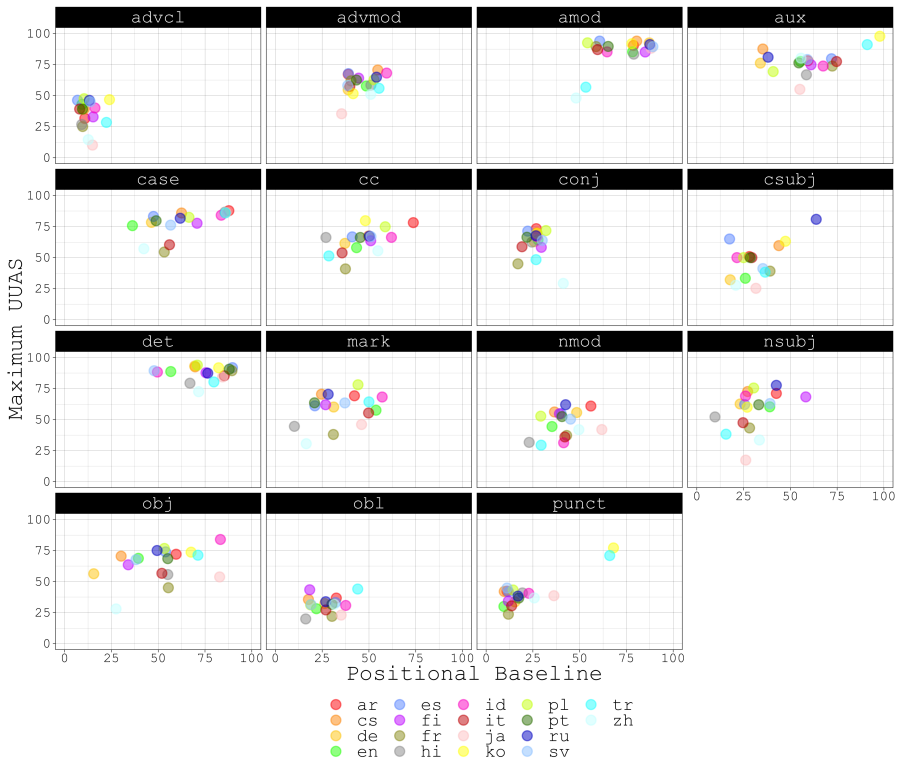


Figure A.19: Decoding UUAS as a function of best positional baselines.

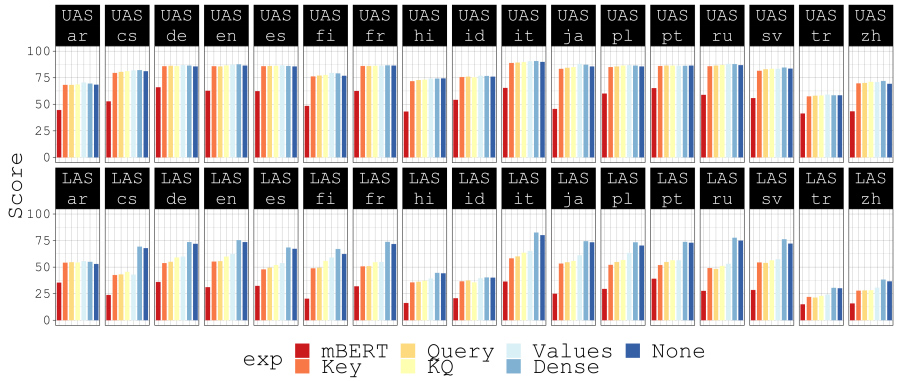


Figure A.20: Parsing scores across components and languages.

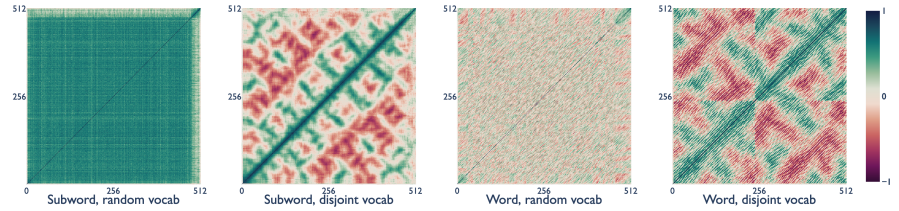


Figure A.21: Pearson correlations, when scrambling by subword/word, with/without disjoint vocabularies. Disjoint vocabularies appear to induce patterns in position-position correlations, while scrambling at a word level induces 'stripes' of oscillating magnitude; this is likely due to position embeddings learning connections to adjacent tokens.

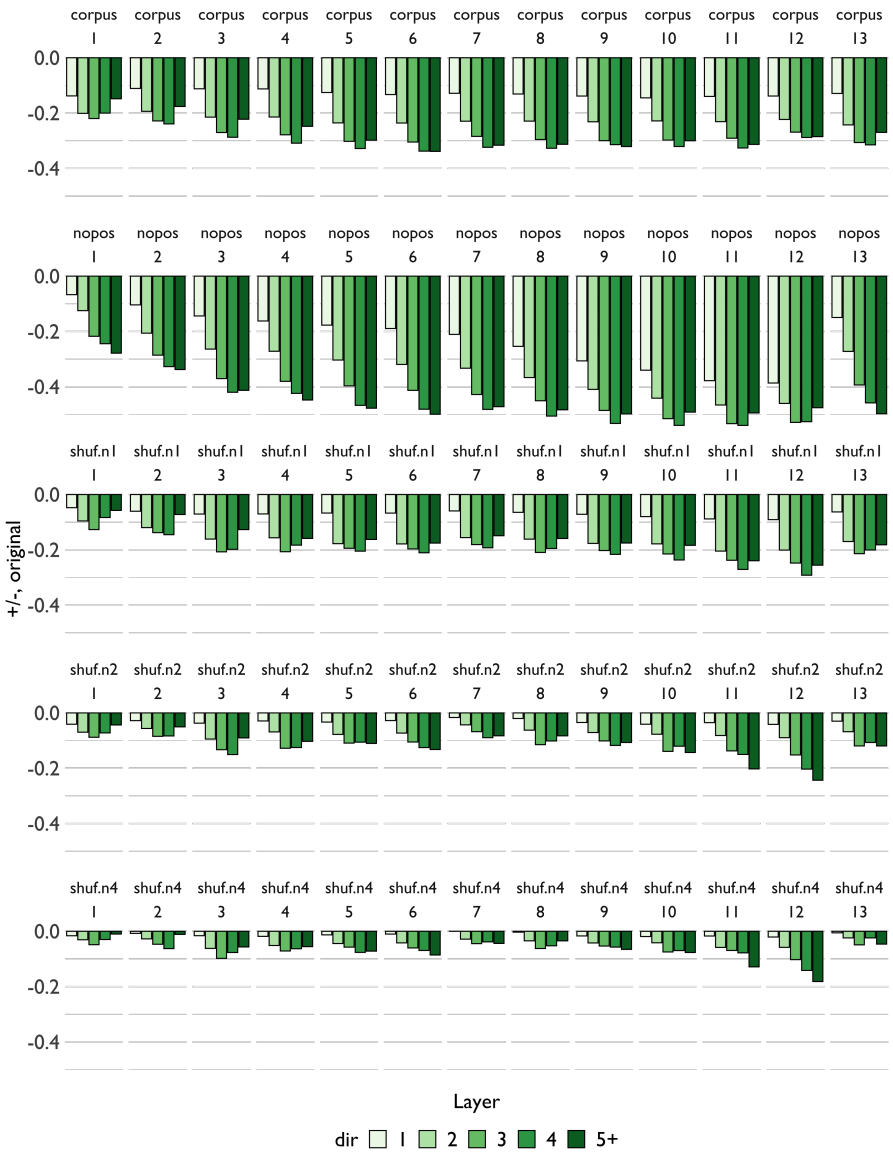


Figure A.22: Δ UAS, all models and layers across dependency lengths 1-5+, w.r.t. ORIG. Layer 13 represents a linear mix of all model layers.

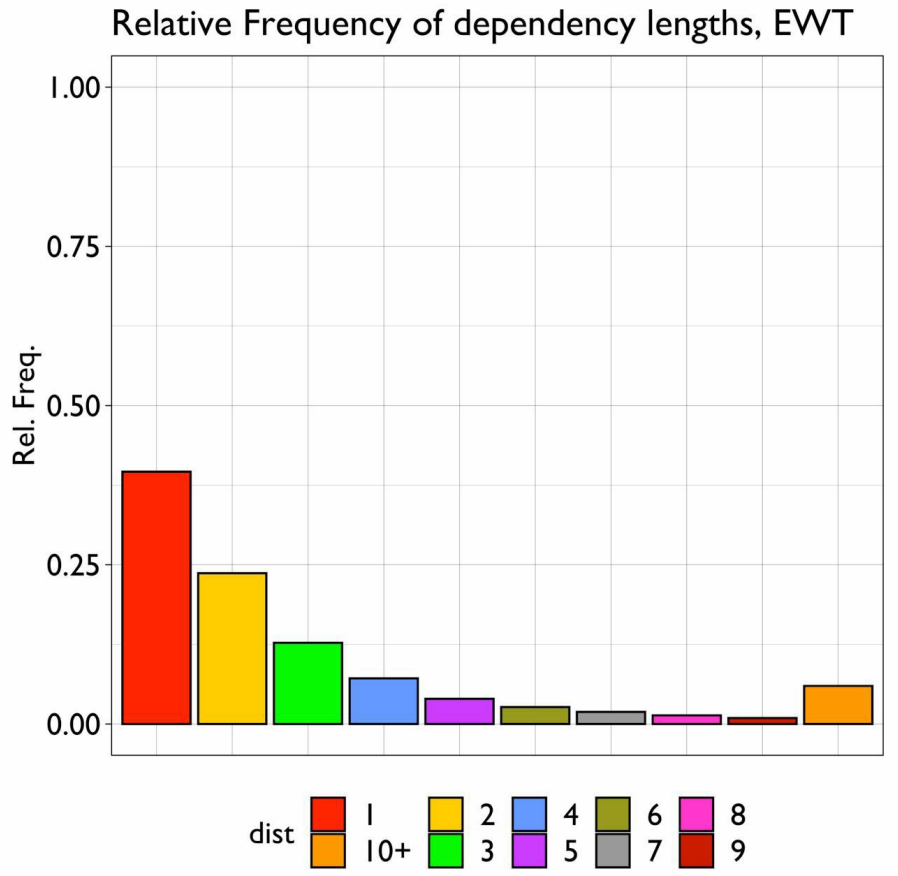


Figure A.23: Relative frequencies of dependency relations in $\text{UD}_{\texttt{English-EWT}}$, at a dependency lengths indicated by the x-axis

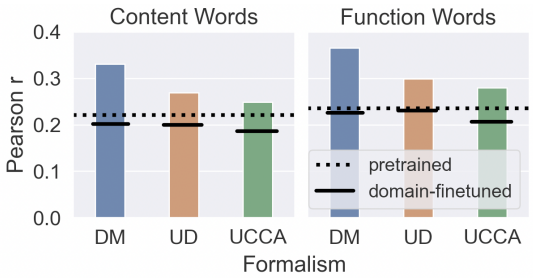


Figure A.24: Content word and function word brain decoding score (mean Pearson's r) for models fine-tuned by MLM with **guided attention** on each of the formalisms, as well the baseline models: *pretrained* BERT (dotted) and *domain-finetuned* BERT (solid).

BIBLIOGRAPHY

Abdou, Mostafa, Ana Valeria González, Mariya Toneva, Daniel Hershcovich, and Anders Søgaard (2021). "Does injecting linguistic structure into language models lead to better alignment with brain recordings?" In: *arXiv preprint arXiv:2101.12608*.

Abdou, Mostafa, Artur Kulmizev, Felix Hill, Daniel M Low, and Anders Søgaard (2019). "Higher-order Comparisons of Sentence Encoder Representations." In: *arXiv preprint arXiv:1909.00303*.

Abend, Omri and Ari Rappoport (Aug. 2013). "Universal Conceptual Cognitive Annotation (UCCA)." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 228–238. URL: https://www.aclweb.org/anthology/P13-1023.

Abnar, Samira, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema (2017). "Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity." In: *arXiv preprint arXiv:1711.09285*.

Abnar, Samira, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema (Aug. 2019a). "Blackbox Meets Blackbox: Representational Similarity & Stability Analysis of Neural Language Models and Brains." In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 191–203. URL: https://www.aclweb.org/anthology/W19-4820.

— (2019b). "Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains." In: *arXiv preprint arXiv:1906.01539*.

Abzianidze, Lasha and Johan Bos (2017). "Towards Universal Semantic Tagging." In: *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*. URL: https://www.aclweb.org/anthology/W17-6901.

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg (2016). "Fine-grained analysis of sentence embeddings using auxiliary prediction tasks." In: *arXiv preprint arXiv:1608.04207*.

Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa (2009). "A study on similarity and relatedness using distributional and wordnet-based approaches." In:

Alleman, Matteo, Jonathan Mamou, Miguel A Del Rio, Hanlin Tang, Yoon Kim, and SueYeon Chung (2021). "Syntactic Perturbations Reveal Representational Correlates of Hierarchical Phrase Structure in Pretrained Language Models." In: *arXiv preprint arXiv:2104.07578*.

Anderson, Andrew J, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni (2013). "Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts." In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1960–1970.

Anderson, Andrew J, Douwe Kiela, Stephen Clark, and Massimo Poesio (2017a). "Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns." In: *Transactions of the Association of Computational Linguistics* 5.1, pp. 17–30.

Anderson, Andrew James, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev DS Raizada (2017b). "Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation." In: *Cerebral Cortex* 27.9, pp. 4379–4395.

Antonello, Richard, Javier Turek, Vy Vo, and Alexander Huth (2021). "Low-Dimensional Structure in the Space of Language Representations is Reflected in Brain Responses." In: *arXiv preprint arXiv:2106.05426*.

Ashby, Jane, Keith Rayner, and Charles Clifton (2005). "Eye movements of highly skilled and average readers: Differential effects of frequency and predictability." In: *The Quarterly Journal of Experimental Psychology Section A* 58.6, pp. 1065–1086.

Baars, Bernard and Nicole M Gage (2013). *Fundamentals of cognitive neuroscience: a beginner's guide*. Academic Press.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate." In: *arXiv preprint arXiv:1409.0473*.

Bahlmann, Jörg, Antoni Rodriguez-Fornells, Michael Rotte, and Thomas F Münte (2007). "An fMRI study of canonical and noncanonical word order in German." In: *Human brain mapping* 28.10, pp. 940–949.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider (Aug. 2013). "Abstract Meaning Representation for Sembanking." In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 178–186. URL: https://www.aclweb.org/anthology/W13-2322.

Baroni, Marco (2021). "On the proper role of linguistically-oriented deep net analysis in linguistic theorizing." In: *arXiv preprint arXiv:2106.08694*.

Barrett, Maria, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard (2018). "Sequence classification with human attention." In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 302–312.

Barzilay, Regina and Lillian Lee (2004). "Catching the drift: Probabilistic content models, with applications to generation and summarization." In: *arXiv preprint cs/0405039*.

Beinborn, Lisa, Samira Abnar, and Rochelle Choenni (2019). "Robust evaluation of language-brain encoding experiments." In: *arXiv preprint arXiv:1904.02547*.

Belinkov, Yonatan (2021). "Probing classifiers: Promises, shortcomings, and alternatives." In: *arXiv preprint arXiv:2102.12452*.

Belinkov, Yonatan, Sebastian Gehrmann, and Ellie Pavlick (July 2020). "Interpretability and Analysis in Neural NLP." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Online: Association for Computational Linguistics, pp. 1–5. DOI: 10.18653/v1/2020.acl-tutorials.1. URL: https://www.aclweb.org/anthology/2020.acl-tutorials.1.

Belinkov, Yonatan and James Glass (Mar. 2019a). "Analysis Methods in Neural Language Processing: A Survey." In: *Transactions of the Association for Computational Linguistics* 7, pp. 49–72. DOI: 10.1162/tacl_a_00254. URL: https://www.aclweb.org/anthology/Q19-1004.

— (Mar. 2019b). "Analysis Methods in Neural Language Processing: A Survey." In: *Transactions of the Association for Computational Linguistics* 7, pp. 49–72. DOI: 10.1162/tacl_a_00254.

Belinkov, Yonatan, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass (Nov. 2017b). "Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks." In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 1–10. URL: https://www.aclweb.org/anthology/I17-1001.

— (2017a). "Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks." In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. 1, pp. 1–10.

Belinkov, Yonatan, Adam Poliak, Stuart M Shieber, Benjamin Van Durme, and Alexander M Rush (2019a). "Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference." In: *arXiv preprint arXiv:1907.04380*.

Belinkov, Yonatan, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush (June 2019b). "On Adversarial Removal of Hypothesis-only Bias in Natural Language Inference." In: *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 256–262. DOI: 10.18653/v1/S19-1028.

Bender, Emily M. and Alexander Koller (July 2020a). "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*. Online: Association for Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://www.aclweb.org/anthology/2020.acl-main.463.

Bender, Emily M and Alexander Koller (2020b). "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.

Bengio, Yoshua (2009). *Learning deep architectures for AI*. Now Publishers Inc.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin (Mar. 2003). "A Neural Probabilistic Language Model." In: *J. Mach. Learn. Res.* 3, pp. 1137–1155. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944966.

Berlin, Brent and Paul Kay (1991). *Basic color terms: Their universality and evolution*. Univ of California Press.

Bever, Thomas G (1970). "The cognitive basis for linguistic structures." In: *Cognition and the development of language*.

Bhattasali, Shohini, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale (May 2020). "The Alice Datasets: fMRI & EEG Observations of Natural Language Comprehension." English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 120–125. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.15.

Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. (2020a). "Experience grounds language." In: *arXiv preprint arXiv:2004.10151*.

Bisk, Yonatan et al. (Nov. 2020b). "Experience Grounds Language." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8718–8735. DOI: 10.18653/v1/2020.emnlp-main.703. URL: https://www.aclweb.org/anthology/2020.emnlp-main.703.

Blank, Idan, Zuzanna Balewski, Kyle Mahowald, and Evelina Fedorenko (2016). "Syntactic processing is distributed across the language system." In: *Neuroimage* 127, pp. 307–323.

Blevins, Terra, Omer Levy, and Luke Zettlemoyer (2018). "Deep rnns encode soft hierarchical syntax." In: *arXiv preprint arXiv:1805.04218*.

Bloomfield, Leonard (1926). "A set of postulates for the science of language." In: *Language* 2.3, pp. 153–164.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching word vectors with subword information." In:

Transactions of the Association for Computational Linguistics 5, pp. 135–146.

Bommasani, Rishi, Kelly Davis, and Claire Cardie (July 2020a). "Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4758–4781. DOI: 10.18653/v1/2020.acl-main.431. URL: https://www.aclweb.org/anthology/2020.acl-main.431.

— (2020b). "Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4758–4781.

Bommasani, Rishi, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. (2021). "On the opportunities and risks of foundation models." In: *arXiv preprint arXiv:2108.07258*.

Bos, Johan, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva (2017). "The Groningen meaning bank." In: *Handbook of linguistic annotation*. Springer, pp. 463–496.

Bouchacourt, Diane and Marco Baroni (2018). "How agents see things: On visual representations in an emergent language game." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 981–985. DOI: 10.18653/v1/D18-1119. URL: https://www.aclweb.org/anthology/D18-1119.

Bowman, Samuel R, Gabor Angeli, Christopher Potts, and Christopher D Manning (2015). "A large annotated corpus for learning natural language inference." In: *arXiv preprint arXiv:1508.05326*.

Boyce, Veronica, Richard Futrell, and Roger P Levy (2020). "Maze Made Easy: Better and easier measurement of incremental processing difficulty." In: *Journal of Memory and Language* 111, p. 104082.

Brennan, Jonathan R, Chris Dyer, Adhiguna Kuncoro, and John T Hale (2020). "Localizing syntactic predictions using recurrent neural network grammars." In: *Neuropsychologia* 146, p. 107479.

Brennan, Jonathan R and John T Hale (2019). "Hierarchical structure guides rapid linguistic predictions during naturalistic listening." In: *PloS one* 14.1, e0207741.

Brennan, Jonathan R and Liina Pylkkänen (2017). "MEG evidence for incremental sentence composition in the anterior temporal lobe." In: *Cognitive science* 41, pp. 1515–1531.

Brennan, Jonathan R, Edward P Stabler, Sarah E Van Wagenen, Wen-Ming Luh, and John T Hale (2016). "Abstract linguistic structure correlates with temporal activity during naturalistic comprehension." In: *Brain and language* 157, pp. 81–94.

Brennan, Jonathan and Liina Pylkkänen (2012). "The time-course and spatial distribution of brain activity associated with sentence processing." In: *Neuroimage* 60.2, pp. 1139–1148.

Bresnan, Joan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler (2015). *Lexical-functional syntax*. John Wiley & Sons.

Broca, Paul et al. (1861). "Perte de la parole, ramollissement chronique et destruction partielle du lobe antérieur gauche du cerveau." In: *Bull Soc Anthropol* 2.1, pp. 235–238.

Brown, Peter F, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin (1990). "A statistical approach to machine translation." In: *Computational linguistics* 16.2, pp. 79–85.

Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020). "Language models are few-shot learners." In: *arXiv preprint arXiv:2005.14165*.

Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran (2012). "Distributional semantics in technicolor." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 136–145.

Bugliarello, Emanuele and Naoaki Okazaki (July 2020). "Enhancing Machine Translation with Dependency-Aware Self-Attention." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1618–1627. DOI: 10.18653/v1/2020.acl-main.147. URL: https://www.aclweb.org/anthology/2020.acl-main.147.

Bulat, Luana, Stephen Clark, and Ekaterina Shutova (2017). "Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1081–1091.

Bullinaria, John A and Joseph P Levy (2007). "Extracting semantic representations from word co-occurrence statistics: A computational study." In: *Behavior research methods* 39.3, pp. 510–526.

— (2013). "Limiting factors for mapping corpus-based semantic representations to brain activity." In: *PloS one* 8.3, e57191.

Caucheteux, Charlotte, Alexandre Gramfort, and Jean-Remi King (2021a). "Disentangling syntax and semantics in the brain with deep networks." In: *International Conference on Machine Learning*. PMLR, pp. 1336–1348.

Caucheteux, Charlotte, Alexandre Gramfort, and Jean-Rémi King (2021b). "GPT-2's activations predict the degree of semantic comprehension in the human brain." In: *bioRxiv*.

Caucheteux, Charlotte and Jean-Rémi King (2020). "Language processing in brains and deep neural networks: computational convergence and its limits." In: *BioRxiv*.

Cawley, Gavin C and Nicola LC Talbot (2010). "On over-fitting in model selection and subsequent selection bias in performance evaluation." In: *The Journal of Machine Learning Research* 11, pp. 2079–2107.

Chaabouni, Rahma, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni (2021). "Communicating artificial neural networks develop efficient color-naming systems." In: *Proceedings of the National Academy of Sciences* 118.12.

Chastrette, M (1997). "Trends in structure-odor relationship." In: *SAR and QSAR in Environmental Research* 6.3-4, pp. 215–254.

Chen, Danqi, Jason Bolton, and Christopher D. Manning (Aug. 2016a). "A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 2358–2367. DOI: 10.18653/v1/P16-1223. URL: https://www.aclweb.org/anthology/P16-1223.

— (Aug. 2016b). "A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 2358–2367. DOI: 10.18653/v1/P16-1223.

Chen, Lin (1982). "Topological structure in visual perception." In: *Science* 218.4573, pp. 699–700.

Chen, Yun-Nung, Dilek Hakkani-Tur, Gokhan Tur, Asli Celikyilmaz, Jianfeng Gao, and Li Deng (2016c). "Knowledge as a teacher: Knowledge-guided structural attention networks." In: *arXiv preprint arXiv:1609.03286*.

Chi, Ethan A., John Hewitt, and Christopher D. Manning (July 2020). "Finding Universal Grammatical Relations in Multilingual BERT." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5564–5577. DOI: 10.18653/v1/2020.acl-main.493. URL: https://www.aclweb.org/anthology/2020.acl-main.493.

Chomsky, Noam (1956). "Three models for the description of language." In: *IRE Transactions on information theory* 2.3, pp. 113–124.

— (1957). *Syntactic Structures*. The Hague: Mouton and Co.

— (2014a). *Aspects of the Theory of Syntax*. Vol. 11. MIT press.

— (2014b). *The minimalist program*. MIT press.

Chowdhury, Shammur Absar and Roberto Zamparelli (Aug. 2019). "An LSTM Adaptation Study of (Un)grammaticality." In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 204–212. DOI: 10.18653/v1/W19-4821. URL: https://www.aclweb.org/anthology/W19-4821.

Christiansen, Morten H and Nick Chater (1999). "Toward a connectionist model of recursion in human linguistic performance." In: *Cognitive Science* 23.2, pp. 157–205.

Chronis, Gabriella and Katrin Erk (2020). "When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships." In: *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 227–244.

Chrupała, Grzegorz and Afra Alishahi (2019). "Correlating neural and symbolic representations of language." In: *arXiv preprint arXiv:1905.06401*.

Chrupała, Grzegorz, Lieke Gelderloos, and Afra Alishahi (2017). "Representations of language in a model of visually grounded speech signal." In: *arXiv preprint arXiv:1702.01991*.

Chu, Yoeng-Jin (1965). "On the shortest arborescence of a directed graph." In: *Scientia Sinica* 14, pp. 1396–1400.

Church, Kenneth Ward (1989). "A stochastic parts program and noun phrase parser for unrestricted text." In: *International Conference on Acoustics, Speech, and Signal Processing*, IEEE, pp. 695–698.

Church, Kenneth Ward and Patrick Hanks (1990). "Word association norms, mutual information, and lexicography." In: *Computational linguistics* 16.1, pp. 22–29.

Cichy, Radoslaw Martin, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva (2016). "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence." In: *Scientific reports* 6, p. 27755.

Clark, Christopher, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova (2019a). "BoolQ: Exploring the surprising difficulty of natural yes/no questions." In: *arXiv preprint arXiv:1905.10044*.

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D Manning (2019b). "What Does BERT Look At? An Analysis of BERT's Attention." In: *arXiv preprint arXiv:1906.04341*.

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning (Aug. 2019c). "What Does BERT Look at? An Analysis of BERT's Attention." In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 276–286. DOI: 10.18653/v1/W19-4828. URL: https://www.aclweb.org/anthology/W19-4828.

Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning (2020). "Electra: Pre-training text encoders as discriminators rather than generators." In: *arXiv preprint arXiv:2003.10555*.

Clouatre, Louis, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar (2021). "Demystifying Neural Language Models' Insensitivity to Word-Order." In: *arXiv preprint arXiv:2107.13955*.

Collins, Rebecca L (2011). "Content analysis of gender roles in media: Where are we now and where should we go?" In: *Sex roles* 64.3-4, pp. 290–298.

Comrie, Bernard (1989). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes (2017a). "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data." In: *CoRR* abs/1705.02364. arXiv: 1705.02364. URL: http://arxiv.org/abs/1705.02364.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes (2017b). "Supervised learning of universal sentence representations from natural language inference data." In: *arXiv preprint arXiv:1705.02364*.

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni (2018). "What you can cram into a single vector: Probing sentence embeddings for linguistic properties." In: *arXiv preprint arXiv:1805.01070*.

Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan A Sag (2005). "Minimal recursion semantics: An introduction." In: *Research on language and computation* 3.2-3, pp. 281–332.

Cordonnier, Jean-Baptiste, Andreas Loukas, and Martin Jaggi (Jan. 2020). "On the Relationship between Self-Attention and Convolutional Layers." In: *arXiv:1911.03584 [cs, stat]*.

Dagan, Ido, Oren Glickman, and Bernardo Magnini (2005). "The pascal recognising textual entailment challenge." In: *Machine Learning Challenges Workshop*. Springer, pp. 177–190.

Danilevsky, Marina, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen (Dec. 2020). "A Survey of the State of Explainable AI for Natural Language Processing." In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 447–459. URL: https://www.aclweb.org/anthology/2020.aacl-main.46.

Danks, Joseph H and Sam Glucksberg (1971). "Psychological scaling of adjective orders." In: *Journal of Memory and Language* 10.1, p. 63.

Dapretto, Mirella and Susan Y Bookheimer (1999). "Form and content: dissociating syntax and semantics in sentence comprehension." In: *Neuron* 24.2, pp. 427–432.

Dasgupta, Ishita, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman (2018). "Evaluating compositionality in sentence embeddings." In: *arXiv preprint arXiv:1802.04302*.

De Marneffe, Marie-Catherine, Mandy Simons, and Judith Tonhauser (2019). "The CommitmentBank: Investigating projection in natu-

rally occurring discourse." In: *proceedings of Sinn und Bedeutung*. Vol. 23. 2, pp. 107–124.

Desmond, Roger and Anna Danilewicz (2010). "Women are on, but not in, the news: Gender roles in local television news." In: *Sex Roles* 62.11-12, pp. 822–829.

Devereux, Barry, Colin Kelly, and Anna Korhonen (2010). "Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora." In: *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pp. 70–78.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019a). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

— (June 2019b). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Ding, Nai, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel (2016). "Cortical tracking of hierarchical linguistic structures in connected speech." In: *Nature neuroscience* 19.1, pp. 158–164.

Dinu, Georgiana, Angeliki Lazaridou, and Marco Baroni (2014). "Improving zero-shot learning by mitigating the hubness problem." In: *arXiv preprint arXiv:1412.6568*.

Dixon, Robert M. W. (2010/2012). *Basic Linguistic Theory*. 3 vols. Oxford University Press.

Dodge, Jesse, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith (2020). "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping." In: *arXiv preprint arXiv:2002.06305*.

Doumas, Leonidas AA, John E Hummel, and Catherine M Sandhofer (2008). "A theory of the discovery and predication of relational concepts." In: *Psychological review* 115.1, p. 1.

Dozat, Timothy and Christopher D Manning (2016). "Deep biaffine attention for neural dependency parsing." In: *arXiv preprint arXiv:1611.01734*.

Dozat, Timothy, Peng Qi, and Christopher D. Manning (Aug. 2017). "Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task." In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, pp. 20–

30. DOI: 10.18653/v1/K17-3002. URL: https://www.aclweb.org/anthology/K17-3002.

Duffy, Susan A, Robin K Morris, and Keith Rayner (1988). "Lexical ambiguity and fixation times in reading." In: *Journal of memory and language* 27.4, pp. 429–446.

Durrani, Nadir, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov (Nov. 2020). "Analyzing Individual Neurons in Pre-trained Language Models." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4865–4880. DOI: 10.18653/v1/2020.emnlp-main.395. URL: https://www.aclweb.org/anthology/2020.emnlp-main.395.

Dyer, Chris, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith (2016). "Recurrent neural network grammars." In: *arXiv preprint arXiv:1602.07776*.

Eickenberg, Michael, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion (2017). "Seeing it all: Convolutional network layers map the function of the human visual system." In: *NeuroImage* 152, pp. 184–194.

Ennis, Robert J and Qasim Zaidi (2019). "Geometrical structure of perceptual color space: mental representations and adaptation invariance." In: *Journal of vision* 19.12, pp. 1–1.

Ettinger, Allyson (2019). "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models." In: *arXiv preprint arXiv:1907.13528*.

Fedorenko, Evelina and Idan A Blank (2020). "Broca's area is not a natural kind." In: *Trends in cognitive sciences* 24.4, pp. 270–284.

Fedorenko, Evelina, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff (2020). "Lack of selectivity for syntax relative to word meanings throughout the language network." In: *Cognition* 203, p. 104348.

Fedorenko, Evelina, Alfonso Nieto-Castanon, and Nancy Kanwisher (2012). "Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses." In: *Neuropsychologia* 50.4, pp. 499–513.

Fedorenko, Evelina, Terri L Scott, Peter Brunner, William G Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher (2016). "Neural correlate of the construction of sentence meaning." In: *Proceedings of the National Academy of Sciences* 113.41, E6256–E6262.

Fedorenko, Evelina and Sharon L Thompson-Schill (2014). "Reworking the language network." In: *Trends in cognitive sciences* 18.3, pp. 120–126.

Feldbauer, Roman, Maximilian Leodolter, Claudia Plant, and Arthur Flexer (2018). "Fast approximate hubness reduction for large high-dimensional data." In: *2018 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, pp. 358–367.

Feldman, Joshua, Joe Davison, and Alexander M. Rush (2019). *Commonsense Knowledge Mining from Pretrained Models*. arXiv: 1909.00505 [cs.CL].

Feng, Shiwen, Jennifer Legault, Long Yang, Junwei Zhu, Keqing Shao, and Yiming Yang (2015). "Differences in grammatical processing strategies for active and passive sentences: An fMRI study." In: *Journal of Neurolinguistics* 33, pp. 104–117.

Ferreira, Fernanda, Karl GD Bailey, and Vittoria Ferraro (2002). "Good-enough representations in language comprehension." In: *Current directions in psychological science* 11.1, pp. 11–15.

Flickinger, Dan, Stephan Oepen, and Emily M Bender (2017). "Sustainable development and refinement of complex linguistic annotations at scale." In: *Handbook of Linguistic Annotation*. Springer, pp. 353–377.

Fodor, Jerry A (1983). *The modularity of mind*. MIT press.

Fodor, Jerry and Merrill Garrett (1966). "Some reflections on competence and performance." In: *Psycholinguistic papers*, pp. 135–179.

Forbes, Maxwell, Ari Holtzman, and Yejin Choi (2019). "Do Neural Language Representations Learn Physical Commonsense?" In: *arXiv preprint arXiv:1908.02899*.

Forster, Kenneth I, Christine Guerrera, and Lisa Elliot (2009). "The maze task: Measuring forced incremental sentence processing time." In: *Behavior research methods* 41.1, pp. 163–171.

Frank, Stefan L and Morten H Christiansen (2018). "Hierarchical and sequential processing of language: A response to: Ding, Melloni, Tian, and Poeppel (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience. Language, Cognition and Neuroscience." In: *Language, Cognition and Neuroscience* 33.9, pp. 1213–1218.

Frank, Stefan L, Leun J Otten, Giulia Galli, and Gabriella Vigliocco (2013). "Word surprisal predicts N400 amplitude during reading." In:

— (2015). "The ERP response to the amount of information conveyed by words in sentences." In: *Brain and language* 140, pp. 1–11.

Frassinelli, Diego (2015). "The effect of context on the activation and processing of word meaning over time." In:

Friederici, Angela D (2011). "The brain basis of language processing: from structure to function." In: *Physiological reviews* 91.4, pp. 1357–1392.

Friederici, Angela D and Sarah ME Gierhan (2013). "The language network." In: *Current opinion in neurobiology* 23.2, pp. 250–254.

Friederici, Angela D, Axel Mecklinger, Kevin M Spencer, Karsten Steinhauer, and Emanuel Donchin (2001). "Syntactic parsing preferences and their on-line revisions: A spatio-temporal analysis of

event-related brain potentials." In: *Cognitive Brain Research* 11.2, pp. 305–323.

Friederici, Angela D, Martin Meyer, and D Yves Von Cramon (2000). "Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information." In: *Brain and language* 74.2, pp. 289–300.

Friederici, Angela D, Erdmut Pfeifer, and Anja Hahne (1993). "Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations." In: *Cognitive brain research* 1.3, pp. 183–192.

Futrell, Richard and Roger P. Levy (2019). "Do RNNs learn human-like abstract word order preferences?" In: *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pp. 50–59. DOI: 10.7275/jb34-9986. URL: https://www.aclweb.org/anthology/W19-0106.

Futrell, Richard, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy (June 2019). "Neural language models as psycholinguistic subjects: Representations of syntactic state." 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 32–42. DOI: 10.18653/v1/N19-1004. URL: https://aclanthology.org/N19-1004.

Fyshe, Alona, Partha P Talukdar, Brian Murphy, and Tom M Mitchell (2014). "Interpretable semantic vectors from a joint model of brain- and text-based meaning." In: *Proceedings of the conference. Association for Computational Linguistics. Meeting*. Vol. 2014. NIH Public Access, p. 489.

Gałecki, Andrzej and Tomasz Burzykowski (2013). "Linear mixed-effects model." In: *Linear Mixed-Effects Models Using R*. Springer, pp. 245–273.

Gauthier, Jon and Anna Ivanova (2018). "Does the brain represent words? An evaluation of brain decoding studies of language understanding." In: *arXiv preprint arXiv:1806.00591*.

Gauthier, Jon and Roger Levy (2019). "Linking artificial and human neural representations of language." In: *arXiv preprint arXiv:1910.01244*.

Geschwind, Norman (1970). "The organization of language and the brain." In: *Science* 170.3961, pp. 940–944.

Gibson, Edward (1998). "Linguistic complexity: Locality of syntactic dependencies." In: *Cognition* 68.1, pp. 1–76.

Gibson, Edward, Leon Bergen, and Steven T Piantadosi (2013). "Rational integration of noisy evidence and prior semantic expectations in sentence interpretation." In: *Proceedings of the National Academy of Sciences* 110.20, pp. 8051–8056.

Gibson, Edward, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T Piantadosi, and Bevil R Conway (2017). "Color naming across

languages reflects color use." In: *Proceedings of the National Academy of Sciences* 114.40, pp. 10785–10790.

Goldberg, Yoav (2019). "Assessing BERT's Syntactic Abilities." In: *arXiv preprint arXiv:1901.05287*.

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. (2021). "Thinking ahead: prediction in context as a keystone of language in humans and machines." In: *bioRxiv*, pp. 2020–12.

Gonzalez-Garduno, Ana Valeria and Anders Søgaard (2017). "Using gaze to predict text readability." In: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 438–443.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton (2013). "Speech recognition with deep recurrent neural networks." In: *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, pp. 6645–6649.

Greenberg, Joseph (1963). "Some universals of grammar with particular reference to the order of meaningful elements." In: *In J. Greenberg, ed., Universals of Language. 73-113. Cambridge, MA.*

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni (June 2018). "Colorless Green Recurrent Networks Dream Hierarchically." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1195–1205. DOI: 10.18653/v1/N18-1108.

Gupta, Anurag (July 2003). "An Adaptive Approach to Collecting Multimodal Input." In: *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, pp. 31–36. DOI: 10.3115/1075178.1075182. URL: https://www.aclweb.org/anthology/P03-2005.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith (June 2018a). "Annotation Artifacts in Natural Language Inference Data." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 107–112. DOI: 10.18653/v1/N18-2017. URL: https://aclanthology.org/N18-2017.

— (June 2018b). "Annotation Artifacts in Natural Language Inference Data." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Hu-*

man Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational Linguistics, pp. 107–112. DOI: 10.18653/v1/N18-2017.

Hagoort, Peter (2005). "On Broca, brain, and binding: a new framework." In: *Trends in cognitive sciences* 9.9, pp. 416–423.

Hagoort, Peter, Colin Brown, and Jolanda Groothusen (1993). "The syntactic positive shift (SPS) as an ERP measure of syntactic processing." In: *Language and cognitive processes* 8.4, pp. 439–483.

Hahn, Michael and Frank Keller (2016). "Modeling human reading with neural attention." In: *arXiv preprint arXiv:1608.05604*.

Hajic, Jan, Eva Hajicová, Jarmila Panevová, Petr Sgall, Ondrej Bojar, Silvie Cinková, Eva Fucíková, Marie Mikulová, Petr Pajas, Jan Popelka, et al. (2012). "Announcing Prague Czech-English Dependency Treebank 2.0." In: *LREC*, pp. 3153–3160.

Hale, John (2001). "A probabilistic Earley parser as a psycholinguistic model." In: *Second meeting of the north American chapter of the association for computational linguistics*.

Hale, John, Luca Campanelli, Jixing Li, Shohini Bhattasali, Christophe Pallier, and Jonathan Brennan (2021). "Neuro-computational models of language processing." In: *Annual Review of Linguistics*.

Hale, John, Chris Dyer, Adhiguna Kuncoro, and Jonathan R Brennan (2018). "Finding syntax in human encephalography with beam search." In: *arXiv preprint arXiv:1806.04127*.

Hale, John, David Lutz, Wen-Ming Luh, and Jonathan Brennan (2015). "Modeling fMRI time courses with linguistic structure at various grain sizes." In: *Proceedings of the 6th workshop on cognitive modeling and computational linguistics*, pp. 89–97.

Henderson, John M, Wonil Choi, Matthew W Lowder, and Fernanda Ferreira (2016). "Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading." In: *Neuroimage* 132, pp. 293–300.

Henry, J Craig (2006). "Electroencephalography: basic principles, clinical applications, and related fields." In: *Neurology* 67.11, pp. 2092–2092.

Hernandez, Evan and Jacob Andreas (2021). "The Low-Dimensional Linear Geometry of Contextualized Word Representations." In: *arXiv preprint arXiv:2105.07109*.

Hershcovich, Daniel, Omri Abend, and Ari Rappoport (July 2017). "A Transition-Based Directed Acyclic Graph Parser for UCCA." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1127–1138. DOI: 10.18653/v1/P17-1104. URL: https://www.aclweb.org/anthology/P17-1104.

— (June 2019). "Content Differences in Syntactic and Semantic Representation." 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). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 478–488. DOI: 10.18653/v1/N19-1047. URL: https://www.aclweb.org/anthology/N19-1047.

Hewitt, John, Kawin Ethayarajh, Percy Liang, and Christopher D Manning (2021). "Conditional probing: measuring usable information beyond a baseline." In: *arXiv preprint arXiv:2109.09234*.

Hewitt, John and Percy Liang (Nov. 2019a). "Designing and Interpreting Probes with Control Tasks." 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 2733–2743. DOI: 10.18653/v1/D19-1275. URL: https://aclanthology.org/D19-1275.

— (2019b). "Designing and interpreting probes with control tasks." In: *arXiv preprint arXiv:1909.03368*.

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.

Hill, Felix, Roi Reichart, and Anna Korhonen (2015). "Simlex-999: Evaluating semantic models with (genuine) similarity estimation." In: *Computational Linguistics* 41.4, pp. 665–695.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory." In: *Neural computation* 9.8, pp. 1735–1780.

Hollenstein, Nora, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang (2019). "Advancing NLP with cognitive language processing signals." In: *arXiv preprint arXiv:1904.02682*.

Hollenstein, Nora, Adrian van der Lek, and Ce Zhang (Dec. 2020a). "CogniVal in Action: An Interface for Customizable Cognitive Word Embedding Evaluation." In: *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*. Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL), pp. 34–40. DOI: 10.18653/v1/2020.coling-demos.7. URL: https://www.aclweb.org/anthology/2020.coling-demos.7.

Hollenstein, Nora, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn (2021). "Multilingual language models predict human reading behavior." In: *arXiv preprint arXiv:2104.05433*.

Hollenstein, Nora, Marius Troendle, Ce Zhang, and Nicolas Langer (May 2020b). "ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation." English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 138–

146. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.18.

Hollenstein, Nora and Ce Zhang (2019). "Entity recognition at first sight: Improving NER with eye movement information." In: *arXiv preprint arXiv:1902.10068*.

Holtzman, Ari, Jan Buys, Maxwell Forbes, and Yejin Choi (2019). "The curious case of neural text degeneration." In: *arXiv preprint arXiv:1904.09751*.

Hou, Yifan and Mrinmaya Sachan (2021). "Bird's Eye: Probing for Linguistic Graph Structures with a Simple Information-Theoretic Approach." In: *arXiv preprint arXiv:2105.02629*.

Houle, Michael E (2015). "Inlierness, outlierness, hubness and discriminability: an extreme-value-theoretic foundation." In: *National Institute of Informatics Technical Report NII-2015-002E, Tokyo, Japan*.

Hoyle, Alexander Miserlis, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell (July 2019). "Unsupervised Discovery of Gendered Language through Latent-Variable Modeling." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1706–1716. URL: https://www.aclweb.org/anthology/P19-1167.

Htut, Phu Mon, Jason Phang, Shikha Bordia, and Samuel R Bowman (2019). "Do Attention Heads in BERT Track Syntactic Dependencies?" In: *arXiv preprint arXiv:1911.12246*.

Hu, Jennifer, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy (2020a). "A systematic assessment of syntactic generalization in neural language models." In: *arXiv preprint arXiv:2005.03692*.

Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson (Sept. 2020b). "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization." In: *arXiv:2003.11080 [cs]*. arXiv: 2003.11080. URL: http://arxiv.org/abs/2003.11080 (visited on 10/02/2020).

Hume, David (1938). *An Abstract of a Treatise of Human Nature, 1740*. CUP Archive.

Humphries, Colin, Jeffrey R Binder, David A Medler, and Einat Liebenthal (2006). "Syntactic and semantic modulation of neural activity during auditory sentence comprehension." In: *Journal of cognitive neuroscience* 18.4, pp. 665–679.

Hupkes, Dieuwke, Sara Veldhoen, and Willem Zuidema (2017). "Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure." In: *arXiv preprint arXiv:1711.10203*.

Huth, Alexander G, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant (2016). "Natural speech reveals the semantic maps that tile human cerebral cortex." In: *Nature* 532.7600, pp. 453–458.

Ilharco, Gabriel, Rowan Zellers, Ali Farhadi, and Hannaneh Hajishirzi (2020). "Probing text models for common ground with visual representations." In: *arXiv preprint arXiv:2005.00619*.

Ivanova, Angelina, Stephan Oepen, Lilja Øvrelid, and Dan Flickinger (July 2012). "Who Did What to Whom? A Contrastive Study of Syntacto-Semantic Dependencies." In: *Proceedings of the Sixth Linguistic Annotation Workshop*. Jeju, Republic of Korea: Association for Computational Linguistics, pp. 2–11. URL: https://www.aclweb.org/anthology/W12-3602.

Jackson, Frank (1982). "Epiphenomenal qualia." In: *The Philosophical Quarterly (1950-)* 32.127, pp. 127–136.

Jain, Sarthak and Byron C Wallace (2019). "Attention is not Explanation." 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. 3543–3556.

Jain, Shailee and Alexander Huth (2018). "Incorporating Context into Language Encoding Models for fMRI." In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/f471223d1a1614b58a7dc45c9d01df19-Paper.pdf.

Jain, Shailee, Shivangi Mahto, Javier S Turek, Vy A Vo, Amanda LeBel, and Alexander G Huth (2021). "Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech." In: *bioRxiv*, pp. 2020–10.

Jelinek, Frederick, Bernard Merialdo, Salim Roukos, and Martin Strauss (1991). "A dynamic language model for speech recognition." In: *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Jin, Huiyuan and Haitao Liu (June 2017). "How will text size influence the length of its linguistic constituents?" In: *Poznan Studies in Contemporary Linguistics* 53. DOI: 10.1515/psicl-2017-0008.

Jumelet, Jaap and Dieuwke Hupkes (Nov. 2018). "Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items." In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 222–231. DOI: 10.18653/v1/W18-5424.

Just, Marcel A and Patricia A Carpenter (1980). "A theory of reading: From eye fixations to comprehension." In: *Psychological review* 87.4, p. 329.

— (1992). "A capacity theory of comprehension: individual differences in working memory." In: *Psychological review* 99.1, p. 122.

Just, Marcel Adam, Patricia A Carpenter, Timothy A Keller, William F Eddy, and Keith R Thulborn (1996). "Brain activation modulated by sentence comprehension." In: *Science* 274.5284, pp. 114–116.

Kamp, Hans and Uwe Reyle (1993). "From discourse to logic: introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory." In: *Studies in linguistics and philosophy*.

Katz, Jerrold J and Jerry A Fodor (1963). "The structure of a semantic theory." In: *language* 39.2, pp. 170–210.

Kaushik, Divyansh and Zachary C Lipton (2018). "How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5010–5015.

Kay, Paul, Brent Berlin, Luisa Maffi, William R Merrifield, and Richard Cook (2009). *The world color survey*. CSLI Publications Stanford, CA.

Kay, Paul, Brent Berlin, and William Merrifield (1991). "Biocultural implications of systems of color naming." In: *Journal of Linguistic Anthropology* 1.1, pp. 12–25.

Kay, Paul and Chad K McDaniel (1978). "The linguistic significance of the meanings of basic color terms." In: *Language*, pp. 610–646.

Kell, Alexander JE, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott (2018). "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy." In: *Neuron* 98.3, pp. 630–644.

Kemighan, Mark D, Kenneth Church, and William A Gale (1990). "A spelling correction program based on a noisy channel model." In: *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Kennedy, Alan, Robin Hill, and Joël Pynte (2003). "The dundee corpus." In: *Proceedings of the 12th European conference on eye movement*.

Khashabi, Daniel, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth (2018). "Looking Beyond the Surface:A Challenge Set for Reading Comprehension over Multiple Sentences." In: *NAACL*.

Kim, Judy Sein, Brianna Aheimer, Verónica Montané Manrara, and Marina Bedny (2020). "Shared understanding of color among congenitally blind and sighted adults." In:

Klein, Dan and Christopher D. Manning (2004). "Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency." In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 479–486.

Kocijan, Vid, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz (2019). "A Surprisingly Robust Trick for Winograd Schema Challenge." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Kondratyuk, Dan and Milan Straka (2019a). "75 Languages, 1 Model: Parsing Universal Dependencies Universally." 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. 2779–2795.

— (Nov. 2019b). "75 Languages, 1 Model: Parsing Universal Dependencies Universally." 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 2779–2795. DOI: 10.18653/v1/D19-1279. URL: https://www.aclweb.org/anthology/D19-1279.

Kornblith, Simon, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton (2019). "Similarity of Neural Network Representations Revisited." In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 3519–3529. URL: http://proceedings.mlr.press/v97/kornblith19a.html.

Kovaleva, Olga, Alexey Romanov, Anna Rogers, and Anna Rumshisky (2019). "Revealing the Dark Secrets of BERT." en. 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 4364–4373. DOI: 10.18653/v1/D19-1445. URL: https://www.aclweb.org/anthology/D19-1445 (visited on 01/22/2021).

Kriegeskorte, Nikolaus (2015). "Deep neural networks: a new framework for modeling biological vision and brain information processing." In: *Annual review of vision science* 1, pp. 417–446.

Kriegeskorte, Nikolaus, Rainer Goebel, and Peter Bandettini (2006). "Information-based functional brain mapping." In: *Proceedings of the National Academy of Sciences* 103.10, pp. 3863–3868.

Kriegeskorte, Nikolaus, Marieke Mur, and Peter A Bandettini (2008). "Representational similarity analysis-connecting the branches of systems neuroscience." In: *Frontiers in systems neuroscience* 2, p. 4.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems* 25, pp. 1097–1105.

Kulmizev, Artur and Joakim Nivre (2021). "Schr\" odinger's Tree–On Syntax and Neural Language Models." In: *arXiv preprint arXiv:2110.08887*.

Kulmizev, Artur, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre (July 2020). "Do Neural Language Models Show Preferences for Syntactic Formalisms?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4077–4091. DOI: 10.18653/v1/

2020.acl-main.375. URL: https://www.aclweb.org/anthology/2020.acl-main.375.

Kutas, Marta and Steven A Hillyard (1980). "Reading senseless sentences: Brain potentials reflect semantic incongruity." In: *Science* 207.4427, pp. 203–205.

— (1984). "Brain potentials during reading reflect word expectancy and semantic association." In: *Nature* 307.5947, pp. 161–163.

Lai, Guokun, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy (2017). "Race: Large-scale reading comprehension dataset from examinations." In: *arXiv preprint arXiv:1704.04683*.

Lakoff, George and Mark Johnson (2008). *Metaphors we live by*. University of Chicago press.

Lauscher, Anne, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš (May 2020). "From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers." In: *arXiv:2005.00633 [cs]*. arXiv: 2005.00633. URL: http://arxiv.org/abs/2005.00633 (visited on 10/02/2020).

LeBel, Amanda, Shailee Jain, and Alexander G Huth (2021). "Voxel-wise encoding models show that cerebellar language representations are highly conceptual." In: *bioRxiv*.

LeCun, Yann, Yoshua Bengio, et al. (1995). "Convolutional networks for images, speech, and time series." In: *The handbook of brain theory and neural networks* 3361.10, p. 1995.

Leech, Geoffrey Neil (1992). "100 million words of English: the British National Corpus (BNC)." In:

Lenneberg, Eric H (1967). "The biological foundations of language." In: *Hospital Practice* 2.12, pp. 59–67.

Lerner, Yulia, Christopher J Honey, Lauren J Silbert, and Uri Hasson (2011). "Topographic mapping of a hierarchy of temporal receptive windows using a narrated story." In: *Journal of Neuroscience* 31.8, pp. 2906–2915.

Levesque, Hector, Ernest Davis, and Leora Morgenstern (2012). "The winograd schema challenge." In: *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015). "Improving distributional similarity with lessons learned from word embeddings." In: *Transactions of the Association for Computational Linguistics* 3, pp. 211–225.

Levy, Roger (2008). "Expectation-based syntactic comprehension." In: *Cognition* 106.3, pp. 1126–1177.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (July 2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for

Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://www.aclweb.org/anthology/2020.acl-main.703.

Li, Jixing, Wen-Ming Luh, Liina Pylkkänen, Yiming Yang, and John Hale (2020). "Modeling pronoun resolution in the brain." In: *bioRxiv*.

Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang (2019). "Visualbert: A simple and performant baseline for vision and language." In: *arXiv preprint arXiv:1908.03557*.

Limisiewicz, Tomasz, Rudolf Rosa, and David Mareček (2020). "Universal Dependencies according to BERT: both more specific and more general." In: *arXiv preprint arXiv:2004.14620*.

Lin, Yongjie, Yi Chern Tan, and Robert Frank (2019). "Open Sesame: Getting Inside BERT's Linguistic Knowledge." In: *arXiv preprint arXiv:1906.01698*.

Lindsey, Delwin T and Angela M Brown (2014). "The color lexicon of American English." In: *Journal of vision* 14.2, pp. 17–17.

Linzen, Tal and Marco Baroni (2021). "Syntactic structure from deep learning." In: *Annual Review of Linguistics* 7, pp. 195–212.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg (2016). "Assessing the ability of LSTMs to learn syntax-sensitive dependencies." In: *Transactions of the Association for Computational Linguistics* 4, pp. 521–535.

Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith (June 2019a). "Linguistic Knowledge and Transferability of Contextual 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1073–1094. DOI: 10.18653/v1/N19-1112. URL: https://www.aclweb.org/anthology/N19-1112.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019b). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." In: *arXiv preprint arXiv:1907.11692*.

— (July 2019c). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." In: *arXiv:1907.11692 [cs]*.

Locke, John (1847). *An essay concerning human understanding*. Kay & Troutman.

Lopopolo, Alessandro, Stefan L Frank, Antal Van den Bosch, and Roel M Willems (2017). "Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain." In: *PloS one* 12.5, e0177794.

Loula, Joao, Marco Baroni, and Brenden M Lake (2018). "Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks." In: *arXiv preprint arXiv:1807.07545*.

Luce, Paul A and David B Pisoni (1998). "Recognizing spoken words: The neighborhood activation model." In: *Ear and hearing* 19.1, p. 1.

Luck, Steven J (2014). *An introduction to the event-related potential technique*. MIT press.

Lucy, Li and Jon Gauthier (2017). "Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning." In: *arXiv preprint arXiv:1705.11168*.

Malmaud, Jonathan, Roger Levy, and Yevgeni Berzak (Nov. 2020). "Bridging Information-Seeking Human Gaze and Machine Reading Comprehension." In: *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 142–152. DOI: 10.18653/v1/2020.conll-1.11. URL: https://www.aclweb.org/anthology/2020.conll-1.11.

Manning, Christopher D, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy (2020). "Emergent linguistic structure in artificial neural networks trained by self-supervision." In: *Proceedings of the National Academy of Sciences*.

Marblestone, Adam H, Greg Wayne, and Konrad P Kording (2016). "Toward an integration of deep learning and neuroscience." In: *Frontiers in computational neuroscience* 10, p. 94.

Martin, Andrea E (2020). "A compositional neural architecture for language." In: *Journal of Cognitive Neuroscience* 32.8, pp. 1407–1427.

Martin, Andrea E and Leonidas AA Doumas (2017). "A mechanism for the cortical computation of hierarchical linguistic structure." In: *PLoS biology* 15.3, e2000663.

Marvin, Rebecca and Tal Linzen (2018). "Targeted Syntactic Evaluation of Language Models." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202.

— (2019). "Targeted Syntactic Evaluation of Language Models." In: *Proceedings of the Society for Computation in Linguistics* 2.1, pp. 373–374.

Matthies, Franz and Anders Søgaard (2013). "With blinkers on: Robust prediction of eye movements across readers." In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 803–807.

McCann, Bryan, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher (2018a). "The Natural Language Decathlon: Multitask Learning as Question Answering." In: *arXiv preprint arXiv:1806.08730*.

— (2018b). "The Natural Language Decathlon: Multitask Learning as Question Answering." In: *arXiv preprint arXiv:1806.08730*.

McClelland, James L (1988). "Connectionist models and psychological evidence." In: *Journal of memory and language* 27.2, pp. 107–123.

McClelland, James L, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze (2020). "Placing language in an integrated understanding system: Next steps toward human-level performance in

neural language models." In: *Proceedings of the National Academy of Sciences* 117.42, pp. 25966–25974.

McClelland, James L, David E Rumelhart, PDP Research Group, et al. (1986). "Parallel distributed processing." In: *Explorations in the Microstructure of Cognition* 2, pp. 216–271.

McCoy, R Thomas, Ellie Pavlick, and Tal Linzen (2019). "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference." In: *arXiv preprint arXiv:1902.01007*.

McCulloch, Warren S and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity." In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133.

McDonald, Ryan, Koby Crammer, and Fernando Pereira (2005a). "Online Large-Margin Training of Dependency Parsers." In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 91–98.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič (2005b). "Non-Projective Dependency Parsing using Spanning Tree Algorithms." In: *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 523–530.

McRae, Ken, George S Cree, Mark S Seidenberg, and Chris McNorgan (2005). "Semantic feature production norms for a large set of living and nonliving things." In: *Behavior research methods* 37.4, pp. 547–559.

Merrill, William, Yoav Goldberg, Roy Schwartz, and Noah A Smith (2021). "Provable Limitations of Acquiring Meaning from Ungrounded Form: What will Future Language Models Understand?" In: *arXiv preprint arXiv:2104.10809*.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient estimation of word representations in vector space." In: *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Stefan Kombrink, Anoop Deoras, Lukar Burget, and Jan Cernocky (2011). "Rnnlm-recurrent neural network language modeling toolkit." In: *Proc. of the 2011 ASRU Workshop*, pp. 196–201.

Miller, George A (1995). "WordNet: a lexical database for English." In: *Communications of the ACM* 38.11, pp. 39–41.

Mitchell, Tom M, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just (2008). "Predicting human brain activity associated with the meanings of nouns." In: *science* 320.5880, pp. 1191–1195.

Mollica, Francis, Matthew Siegelman, Evgeniia Diachek, Steven T Piantadosi, Zachary Mineroff, Richard Futrell, Hope Kean, Peng Qian, and Evelina Fedorenko (2020). "Composition is the core driver of the language-selective network." In: *Neurobiology of Language* 1.1, pp. 104–134.

Murphy, Brian, Marco Baroni, and Massimo Poesio (2009). "EEG responds to conceptual stimuli and corpus semantics." In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 619–627.

Murphy, Brian, Partha Talukdar, and Tom Mitchell (2012). "Learning effective and interpretable semantic models using non-negative sparse embedding." In: *Proceedings of COLING 2012*, pp. 1933–1950.

Murphy, Brian, Leila Wehbe, and Alona Fyshe (2018). "Decoding language from the brain." In: *Language, cognition, and computational models*, pp. 53–80.

Nastase, Samuel A, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J Honey, Yaara Yeshurun, Mor Regev, et al. (2020). "Narratives: fMRI data for evaluating models of naturalistic language comprehension." In: *bioRxiv*.

— (2021). "Narratives: fMRI data for evaluating models of naturalistic language comprehension." In: *bioRxiv*, pp. 2020–12.

Nili, Hamed, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte (2014). "A toolbox for representational similarity analysis." In: *PLoS computational biology* 10.4, e1003553.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman (May 2020a). "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection." English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4034–4043. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.497.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Dan Zeman (2020b). "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection." In: *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*.

Nivre, Joakim et al. (2016). "Universal Dependencies v1: A Multilingual Treebank Collection." In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Nivre, Joakim et al. (2019). *Universal Dependencies 2.4*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. URL: http://hdl.handle.net/11234/1-2988.

O'Connor, Joe and Jacob Andreas (2021). "What Context Features Can Transformer Language Models Use?" In: *arXiv preprint arXiv:2106.08367*.

Oepen, Stephan, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajic, Angelina Ivanova, and Zdenka Urešová (2016). "Semantic Dependency Parsing (SDP) graph banks release 1.0 LDC2016T10." In: *Web Download*.

Olson, David R and Nikola Filby (1972). "On the comprehension of active and passive sentences." In: *Cognitive Psychology* 3.3, pp. 361–381.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli (2019). "fairseq: A fast, extensible toolkit for sequence modeling." In: *arXiv preprint arXiv:1904.01038*.

Palatucci, Mark M, Dean A Pomerleau, Geoffrey E Hinton, and Tom Mitchell (2009). "Zero-shot learning with semantic output codes." In:

Pallier, Christophe, Anne-Dominique Devauchelle, and Stanislas Dehaene (2011). "Cortical representation of the constituent structure of sentences." In: *Proceedings of the National Academy of Sciences* 108.6, pp. 2522–2527.

Papadimitriou, Isabel, Ethan A Chi, Richard Futrell, and Kyle Mahowald (2021). "Deep Subjecthood: Higher-Order Grammatical Features in Multilingual BERT." In: *arXiv preprint arXiv:2101.11043*.

Parikh, Ankur P, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit (2016). "A decomposable attention model for natural language inference." In: *arXiv preprint arXiv:1606.01933*.

Parviz, Mehdi, Mark Johnson, Blake Johnson, and Jon Brock (2011). "Using language models and Latent Semantic Analysis to characterise the N400m neural response." In: *Proceedings of the Australasian Language Technology Association Workshop 2011*, pp. 38–46.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation." In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Pereira, Francisco, Matthew Botvinick, and Greg Detre (2013). "Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments." In: *Artificial intelligence* 194, pp. 240–252.

Pereira, Francisco, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko (2018). "Toward a universal decoder of linguistic meaning from brain activation." In: *Nature communications* 9.1, p. 963.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (June 2018a). "Deep Contextualized Word Representations." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long*

Papers). New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018b). "Deep contextualized word representations." In: *arXiv preprint arXiv:1802.05365*.

Petroni, Fabio, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel (2019). *Language Models as Knowledge Bases?* arXiv: 1909.01066 [cs.CL].

Pham, Thang M, Trung Bui, Long Mai, and Anh Nguyen (2020). "Out of Order: How important is the sequential order of words in a sentence in Natural Language Understanding tasks?" In: *arXiv preprint arXiv:2012.15180*.

Pilehvar, Mohammad Taher and Jose Camacho-Collados (2018). "WiC: the word-in-context dataset for evaluating context-sensitive meaning representations." In: *arXiv preprint arXiv:1808.09121*.

Pimentel, Tiago, Naomi Saphra, Adina Williams, and Ryan Cotterell (Nov. 2020a). "Pareto Probing: Trading Off Accuracy for Complexity." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 3138–3153. DOI: 10.18653/v1/2020.emnlp-main.254. URL: https://aclanthology.org/2020.emnlp-main.254.

— (2020b). "Pareto probing: Trading off accuracy for complexity." In: *arXiv preprint arXiv:2010.02180*.

Pimentel, Tiago, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell (July 2020c). "Information-Theoretic Probing for Linguistic Structure." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4609–4622. DOI: 10.18653/v1/2020.acl-main.420. URL: https://www.aclweb.org/anthology/2020.acl-main.420.

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001.

Poeppel, David (2012). "The maps problem and the mapping problem: two challenges for a cognitive neuroscience of speech and language." In: *Cognitive neuropsychology* 29.1-2, pp. 34–55.

— (2014). "The neuroanatomic and neurophysiological infrastructure for speech and language." In: *Current opinión in neurobiology* 28, pp. 142–149.

Poliak, Adam, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme (June 2018a). "Hypothesis Only Baselines in Natural Language Inference." In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. New

Orleans, Louisiana: Association for Computational Linguistics, pp. 180–191. DOI: 10.18653/v1/S18-2023. URL: https://aclanthology.org/S18-2023.

Poliak, Adam, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme (2018b). "Hypothesis only baselines in natural language inference." In: *arXiv preprint arXiv:1805.01042*.

Pollard, Carl and Ivan A Sag (1988). *Information-based syntax and semantics: Vol. 1: fundamentals*. Center for the Study of Language and Information.

Prince, Alan and Paul Smolensky (2008). *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Provenzi, Edoardo (2020). "Geometry of color perception. Part 1: structures and metrics of a homogeneous color space." In: *The Journal of Mathematical Neuroscience* 10.1, pp. 1–19.

Pylkkänen, Liina (2019). "The neural basis of combinatory syntax and semantics." In: *Science* 366.6461, pp. 62–66.

— (2020). "Neural basis of basic composition: what we have learned from the red–boat studies and their extensions." In: *Philosophical Transactions of the Royal Society B* 375.1791, p. 20190299.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). "Improving language understanding with unsupervised learning." In: *Technical report, OpenAI*.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language models are unsupervised multitask learners." In: *OpenAI Blog* 1.8.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2019). "Exploring the limits of transfer learning with a unified text-to-text transformer." In: *arXiv preprint arXiv:1910.10683*.

Raganato, Alessandro, Jörg Tiedemann, et al. (2018). "An analysis of encoder representations in transformer-based machine translation." In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. The Association for Computational Linguistics.

Rahman, Altaf and Vincent Ng (2012). "Resolving complex cases of definite pronouns: the winograd schema challenge." In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pp. 777–789.

Rajpurkar, Pranav, Robin Jia, and Percy Liang (2018). "Know what you don't know: Unanswerable questions for SQuAD." In: *arXiv preprint arXiv:1806.03822*.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (2016). "Squad: 100,000+ questions for machine comprehension of text." In: *arXiv preprint arXiv:1606.05250*.

Raney, Gary E, Spencer J Campbell, and Joanna C Bovee (2014). "Using eye movements to evaluate the cognitive processes involved in text comprehension." In: *Journal of visualized experiments: JoVE* 83.

Rao, Rajesh PN and Dana H Ballard (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects." In: *Nature neuroscience* 2.1, pp. 79–87.

Ravichander, Abhilasha, Yonatan Belinkov, and Eduard Hovy (2020). "Probing the probing paradigm: Does probing accuracy entail task relevance?" In: *arXiv preprint arXiv:2005.00719*.

Ravishankar, Vinit, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre (Apr. 2021). "Attention Can Reflect Syntactic Structure (If You Let It)." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 3031–3045. DOI: 10.18653/v1/2021.eacl-main.264. URL: https://aclanthology.org/2021.eacl-main.264.

Ravishankar, Vinit and Anders Søgaard (Nov. 2021). "The Impact of Positional Encodings on Multilingual Compression." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 763–777. URL: https://aclanthology.org/2021.emnlp-main.59.

Rayner, Keith (1998). "Eye movements in reading and information processing: 20 years of research." In: *Psychological bulletin* 124.3, p. 372.

Rayner, Keith and Susan A Duffy (1986). "Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity." In: *Memory & cognition* 14.3, pp. 191–201.

Reddy, Aniketh Janardhan and Leila Wehbe (2021). "Syntactic representations in the human brain: beyond effort-based metrics." In: *bioRxiv*, pp. 2020–06.

Regier, Terry, Paul Kay, and Naveen Khetarpal (2007). "Color naming reflects optimal partitions of color space." In: *Proceedings of the National Academy of Sciences* 104.4, pp. 1436–1441.

Reif, Emily, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim (2019a). "Visualizing and Measuring the Geometry of BERT." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf.

— (2019b). "Visualizing and measuring the geometry of BERT." In: *Advances in Neural Information Processing Systems* 32, pp. 8594–8603.

Reitter, David, Frank Keller, and Johanna D Moore (2011). "A computational cognitive model of syntactic priming." In: *Cognitive science* 35.4, pp. 587–637.

Roark, Brian (2001). "Probabilistic top-down parsing and language modeling." In: *Computational linguistics* 27.2, pp. 249–276.

Rocktäschel, Tim, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom (2015). "Reasoning about entailment with neural attention." In: *arXiv preprint arXiv:1509.06664*.

Roemmele, Melissa, Cosmin Adrian Bejan, and Andrew S Gordon (2011). "Choice of plausible alternatives: An evaluation of commonsense causal reasoning." In: *2011 AAAI Spring Symposium Series*.

Rosa, Rudolf and David Mareček (2019). "Inducing syntactic trees from bert representations." In: *arXiv preprint arXiv:1906.11511*.

Rosenblatt, Frank (1958). "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6, p. 386.

Rossiter, Karen J (1996). "Structure- odor relationships." In: *Chemical reviews* 96.8, pp. 3201–3240.

Ruan, Yu-Ping, Xiaodan Zhu, Zhen-Hua Ling, Zhan Shi, Quan Liu, and Si Wei (2019). "Exploring Unsupervised Pretraining and Sentence Structure Modelling for Winograd Schema Challenge." In: *arXiv preprint arXiv:1904.09705*.

Rubinstein, Dana, Effi Levi, Roy Schwartz, and Ari Rappoport (2015). "How well do distributional models capture different types of semantic knowledge?" In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 726–730.

Rubner, Yossi, Carlo Tomasi, and Leonidas Guibas (2000). "The earth mover's distance as a metric for image retrieval." In: *IJCV*.

Rumelhart, David E and James L McClelland (1985). *On learning the past tenses of English verbs.* Tech. rep. CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE.

Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi (2019). "WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale." In: *arXiv preprint arXiv:1907.10641*.

Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi (2020). "Winogrande: An adversarial winograd schema challenge at scale." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 8732–8740.

Saysani, Armin, Michael C Corballis, and Paul M Corballis (2018). "Colour envisioned: Concepts of colour in the blind and sighted." In: *Visual Cognition* 26.5, pp. 382–392.

Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview." In: *Neural networks* 61, pp. 85–117.

Schrimpf, Martin, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko (2020a). "Artificial Neural Networks Accurately Predict Language Processing in the Brain." In: *bioRxiv*. DOI: 10.1101/2020.06.26.174482. eprint: https://www.biorxiv.org/content/early/2020/06/27/2020.06.26.174482.full.pdf. URL: https://www.biorxiv.org/content/early/2020/06/27/2020.06.26.174482.

Schrimpf, Martin, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko (2020b). "The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing." In: *BioRxiv*.

Schrimpf, Martin, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo (2020c). "Integrative benchmarking to advance neurally mechanistic models of human intelligence." In: *Neuron*.

Schütze, Hinrich (1992). "Dimensions of Meaning." In: *SC*, pp. 787–796.

— (1993). "Word space." In: *Advances in neural information processing systems*, pp. 895–902.

Schwartz, Dan, Mariya Toneva, and Leila Wehbe (2019). "Inducing brain-relevant bias in natural language processing models." In: *Advances in Neural Information Processing Systems*, pp. 14123–14133.

Schwartz, Roy, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith (2017). "The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task." In: *arXiv preprint arXiv:1702.01841*.

Schwenk, Holger and Jean-Luc Gauvain (2002). "Connectionist language modeling for large vocabulary continuous speech recognition." In: *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE, pp. I–765.

Serrano, Sofia and Noah A Smith (2019). "Is attention interpretable?" In: *arXiv preprint arXiv:1906.03731*.

Sgall, Petr, Eva Hajicová, and Jarmila Panevová (1986). *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects. Academia*.

Shepard, Roger N and Susan Chipman (1970). "Second-order isomorphism of internal representations: Shapes of states." In: *Cognitive psychology* 1.1, pp. 1–17.

Shi, Xing, Inkit Padhi, and Kevin Knight (2016). "Does string-based neural MT learn source syntax?" In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1526–1534.

Sigurd, Bengt, Mats Eeg-Olofsson, and Joost Van Weijer (2004). "Word length, sentence length and frequency – Zipf revisited." In: *Studia Linguistica* 58.1, pp. 37–52. DOI: https://doi.org/10.1111/j.0039-3193.2004.00109.x. eprint: https://onlinelibrary.wiley.

com/doi/pdf/10.1111/j.0039-3193.2004.00109.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0039-3193.2004.00109.x.

Silveira, Natalia, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning (2014). "A Gold Standard Dependency Corpus for English." In: *LREC*. Citeseer, pp. 2897–2904.

Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. (2016). "Mastering the game of Go with deep neural networks and tree search." In: *nature* 529.7587, pp. 484–489.

Sinha, Koustuv, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela (2021). "Masked language modeling and the distributional hypothesis: Order word matters pre-training for little." In: *arXiv preprint arXiv:2104.06644*.

Sinha, Koustuv, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams (2020). "Unnatural language inference." In: *arXiv preprint arXiv:2101.00010*.

Soares, José M, Ricardo Magalhães, Pedro S Moreira, Alexandre Sousa, Edward Ganz, Adriana Sampaio, Victor Alves, Paulo Marques, and Nuno Sousa (2016). "A hitchhiker's guide to functional magnetic resonance imaging." In: *Frontiers in neuroscience* 10, p. 515.

Søgaard, Anders (2016). "Evaluating word embeddings with fMRI and eye-tracking." In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 116–121.

Stanojević, Miloš, Shohini Bhattasali, Donald Dunagan, Luca Campanelli, Mark Steedman, Jonathan Brennan, and John Hale (2021). "Modeling incremental language comprehension in the brain with Combinatory Categorial Grammar." In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 23–38.

Steedman, Mark and Jason Baldridge (2011). "Combinatory categorial grammar." In: *Non-Transformational Syntax: Formal and Explicit Models of Grammar. Wiley-Blackwell*, pp. 181–224.

Stehwien, Sabrina, Lena Henke, John Hale, Jonathan Brennan, and Lars Meyer (2020). "The little prince in 26 languages: towards a multilingual neuro-cognitive corpus." In: *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pp. 43–49.

Straka, Milan, Jan Hajic, and Jana Straková (2016). "UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4290–4297.

Strubell, Emma and Andrew McCallum (July 2018). "Syntax Helps ELMo Understand Semantics: Is Syntax Still Relevant in a Deep

Neural Architecture for SRL?" In: *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*. Melbourne, Australia: Association for Computational Linguistics, pp. 19–27. DOI: 10.18653/v1/W18-2904. URL: https://www.aclweb.org/anthology/W18-2904.

Strubell, Emma, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum (2018a). "Linguistically-Informed Self-Attention for Semantic Role Labeling." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 5027–5038. DOI: 10.18653/v1/D18-1548. URL: https://www.aclweb.org/anthology/D18-1548.

— (2018b). "Linguistically-Informed Self-Attention for Semantic Role Labeling." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5027–5038.

Sudre, Gustavo, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell (2012). "Tracking neural coding of perceptual and semantic features of concrete nouns." In: *NeuroImage* 62.1, pp. 451–463.

Sun, Chen, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid (2019). "Videobert: A joint model for video and language representation learning." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7464–7473.

Swayamdipta, Swabha, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith (2018). "Syntactic Scaffolds for Semantic Structures." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3772–3782. DOI: 10.18653/v1/D18-1412. URL: https://www.aclweb.org/anthology/D18-1412.

Tabor, Whitney and Michael K Tanenhaus (1999). "Dynamical models of sentence processing." In: *Cognitive Science* 23.4, pp. 491–515.

Tai, Kai Sheng, Richard Socher, and Christopher D Manning (2015). "Improved semantic representations from tree-structured long short-term memory networks." In: *arXiv preprint arXiv:1503.00075*.

Taylor, Wilson L (1953). ""Cloze procedure": A new tool for measuring readability." In: *Journalism quarterly* 30.4, pp. 415–433.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick (2019). "Bert rediscovers the classical nlp pipeline." In: *arXiv preprint arXiv:1905.05950*.

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, et al. (2018). "What do you learn from context? Probing for sentence structure in contextualized word representations." In:

Toneva, Mariya, Tom M Mitchell, and Leila Wehbe (2020). "Combining computational controls with natural text reveals new aspects of meaning composition." In: *bioRxiv*.

Toneva, Mariya and Leila Wehbe (2019a). "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)." In: *arXiv preprint arXiv:1905.11833*.

— (2019b). "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)." In: *NeurIPS*.

Traxler, Matthew J (2014). "Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing." In: *Trends in cognitive sciences* 18.11, pp. 605–611.

Trichelair, Paul, Ali Emami, Jackie Chi Kit Cheung, Adam Trischler, Kaheer Suleman, and Fernando Diaz (2018). "On the evaluation of common-sense reasoning in natural language understanding." In: *arXiv preprint arXiv:1811.01778*.

Trinh, Trieu H and Quoc V Le (2018). "A simple method for commonsense reasoning." In: *arXiv preprint arXiv:1806.02847*.

Tsuchiya, Masatoshi (2018). "Performance impact caused by hidden bias of training data for recognizing textual entailment." In: *arXiv preprint arXiv:1804.08117*.

Turc, Iulia, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "Well-read students learn better: On the importance of pre-training compact models." In: *arXiv preprint arXiv:1908.08962*.

Turney, Peter D and Patrick Pantel (2010). "From frequency to meaning: Vector space models of semantics." In: *Journal of artificial intelligence research* 37, pp. 141–188.

Van Gompel, Roger PG (2007). *Eye movements: A window on mind and brain*. Elsevier.

Van Rij, Jacolien, Hedderik Van Rijn, and Petra Hendriks (2013). "How WM load influences linguistic processing in adults: a computational model of pronoun interpretation in discourse." In: *Topics in cognitive science* 5.3, pp. 564–580.

Van Schijndel, Marten and Tal Linzen (2018). "Can Entropy Explain Successor Surprisal Effects in Reading?" In: *arXiv preprint arXiv:1810.11481*.

— (2021). "Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty." In: *Cognitive Science* 45.6, e12988.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need." In: *Advances in Neural Information Processing Systems*, pp. 5998–6008.

Voita, Elena, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov (July 2019a). "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned." In: *Proceedings of the 57th Annual Meeting of the Association for Compu-*

tational Linguistics. Florence, Italy: Association for Computational Linguistics, pp. 5797–5808. DOI: 10.18653/v1/P19-1580. URL: https://aclanthology.org/P19-1580.

— (2019b). "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned." In: *arXiv preprint arXiv:1905.09418*.

Voita, Elena and Ivan Titov (Nov. 2020). "Information-Theoretic Probing with Minimum Description Length." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 183–196. DOI: 10.18653/v1/2020.emnlp-main.14. URL: https://www.aclweb.org/anthology/2020.emnlp-main.14.

Vulić, Ivan, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen (2020). "Probing pretrained language models for lexical semantics." In: *arXiv preprint arXiv:2010.05731*.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019a). "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems." In: *CoRR* abs/1905.00537. arXiv: 1905.00537. URL: http://arxiv.org/abs/1905.00537.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman (2019b). "Superglue: A stickier benchmark for general-purpose language understanding systems." In: *arXiv preprint arXiv:1905.00537*.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman (2018a). "Glue: A multi-task benchmark and analysis platform for natural language understanding." In: *arXiv preprint arXiv:1804.07461*.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (2018b). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355. DOI: 10.18653/v1/W18-5446.

Wang, Alex et al. (2019c). *`jiant` 1.2: A software toolkit for research on general-purpose text understanding models*. http://jiant.info/.

Wang, Benyou, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen (2021). "ON POSITION EMBEDDINGS IN BERT." In: p. 21.

Wang, Shaonan, Jiajun Zhang, Haiyan Wang, Nan Lin, and Chengqing Zong (2020). "Fine-grained neural decoding with distributed word representations." In: *Information Sciences* 507, pp. 256–272.

Wang, Yu-An and Yun-Nung Chen (2020). "What do position embeddings learn? an empirical study of pre-trained language model positional encoding." In: *arXiv preprint arXiv:2010.04903*.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman (2020). "BLiMP: The Benchmark of Linguistic Minimal Pairs for English." In: *Transactions of the Association for Computational Linguistics* 8, pp. 377–392. DOI: 10.1162/tacl_a_00321. URL: https://www.aclweb.org/anthology/2020.tacl-1.25.

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman (Mar. 2019). "Neural Network Acceptability Judgments." In: *Transactions of the Association for Computational Linguistics* 7, pp. 625–641. DOI: 10.1162/tacl_a_00290. URL: https://www.aclweb.org/anthology/Q19-1040.

Wehbe, Leila, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell (Nov. 2014a). "Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses." In: *PLOS ONE* 9.11. Ed. by Kevin Paterson, e112575. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0112575. URL: http://dx.plos.org/10.1371/journal.pone.0112575.

Wehbe, Leila, Ashish Vaswani, Kevin Knight, and Tom M. Mitchell (2014b). "Aligning context-based statistical models of language with brain activity during reading." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 233–243. URL: http://www.aclweb.org/anthology/D14-1030.

Weir, Nathaniel, Adam Poliak, and Benjamin Van Durme (2020). *Probing Neural Language Models for Human Tacit Assumptions*. arXiv: 2004.04877 [cs.CL].

Wernicke, Carl (1874). *The aphasic symptom complex: a psychological study on an anatomical basis*. Cohn.

Wiegreffe, Sarah and Yuval Pinter (2019). "Attention is not not Explanation." 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. 11–20.

Wilcox, Ethan Gotlieb, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy (2020). "On the predictive power of neural language models for human real-time comprehension behavior." In: *arXiv preprint arXiv:2006.01912*.

Wilcoxon, Frank (1992). "Individual comparisons by ranking methods." In: *Breakthroughs in statistics*. Springer, pp. 196–202.

Wolf, Thomas et al. (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing." In: *ArXiv* abs/1910.03771.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. (2016). "Google's neural machine translation system: Bridging the gap between human and machine translation." In: *arXiv preprint arXiv:1609.08144*.

Wu, Zhaofeng, Hao Peng, and Noah A Smith (2021). "Infusing fine-tuning with semantic dependencies." In: *Transactions of the Association for Computational Linguistics* 9, pp. 226–242.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio (2015). "Show, attend and tell: Neural image caption generation with visual attention." In: *International conference on machine learning*, pp. 2048–2057.

Xu, Yilun, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon (2020). "A theory of usable information under computational constraints." In: *arXiv preprint arXiv:2002.10689*.

Yamins, Daniel LK and James J DiCarlo (2016). "Using goal-driven deep learning models to understand sensory cortex." In: *Nature neuroscience* 19.3, pp. 356–365.

Yamins, Daniel LK, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo (2014). "Performance-optimized hierarchical models predict neural responses in higher visual cortex." In: *Proceedings of the national academy of sciences* 111.23, pp. 8619–8624.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le (2019). "XLNet: Generalized Autoregressive Pretraining for Language Understanding." In: *arXiv preprint arXiv:1906.08237*.

Yngve, Victor H (1960). "A model and an hypothesis for language structure." In: *Proceedings of the American philosophical society* 104.5, pp. 444–466.

Zaslavsky, Noga, Charles Kemp, Terry Regier, and Naftali Tishby (2018). "Efficient compression in color naming and its evolution." In: *Proceedings of the National Academy of Sciences* 115.31, pp. 7937–7942.

Zeman, Daniel, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov (2018). "CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies." In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pp. 1–21.

Zhang, Meishan, Zhenghua Li, Guohong Fu, and Min Zhang (June 2019a). "Syntax-Enhanced Neural Machine Translation with Syntax-Aware 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1151–1161. DOI: 10.18653/v1/N19-1118. URL: https://www.aclweb.org/anthology/N19-1118.

Zhang, Sheng, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme (2018). "Record: Bridging the gap be-

tween human and machine commonsense reading comprehension." In: *arXiv preprint arXiv:1810.12885*.

Zhang, Yuan, Jason Baldridge, and Luheng He (2019b). "PAWS: Paraphrase adversaries from word scrambling." In: *arXiv preprint arXiv:1904.01130*.

Zhao, Yiyun and Steven Bethard (2020). "How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope." en. In: p. 19.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books." In: *The IEEE International Conference on Computer Vision (ICCV)*.