DECODING THE ENCODED – LINGUISTIC SECRETS OFLANGUAGE MODELS: A SYSTEMATIC LITERATURE REVIEW

H. Avetisyan and D. Broneske

Research Area Research Infrastructure and Methods, The German Centre for Higher Education Research and Science Studies (DZHW)

ABSTRACT

Language models' growing role in natural language processing neces- sitates a deeper understanding of their linguistic knowledge. Linguistic probing tasks have become crucial for model explainability, designed to evaluate models' understanding of vari-ous linguistic phenomena. ***Objective****: This systematic review critically assesses the linguistic knowledge of language models via linguistic probing, providing a comprehensive overview ofthe understood linguistic phenomena and identifying future research areas.* ***Method****: We performed an extensive search of relevant academic databases and analyzed 57 articles pub- lished between October 2018 and October 2022.* ***Results****: While language models exhibit extensive linguistic knowledge, limitations persist in their comprehension of specific phe- nomena. The review also points to a need for consensus on evaluating language models' linguistic knowledge and the linguistic terminology used.* ***Conclusion****: Our review offers an extensive look into linguistic knowledge of language models through linguistic probing tasks. This study underscores the importance of understanding these models' linguistic capabilities for effective use in NLP applications and for fostering more explainable AI systems.*

KEYWORDS

LLMs, linguistic knowledge, probing, analysis of LMs.

1. INTRODUCTION AND BACKGROUND

BERTology and its Significance The recent advancements in deep learning techniques have led to the popularity of neural language models, particularly transformer-based models, which have achieved state-of-the-art performance on various natural language understanding tasks. Despite their exceptional performance, how these models acquire and utilize linguistic information is still uncertain.

Transformers, first introduced by [23], have demonstrated outstanding results in a wide range of NLP tasks, such as machine translation [23, 25], question answering [67], text classification [26], and semantic role labeling [27]. The remarkable development of pre-trained language models [67, 28, 29] has sparked questions about what specific aspects of language these models capture and do not capture. As a result, an area of research called BERTology has emerged.

David C. Wyld et al. (Eds): MLNLP, NWCOM, DTMN, ASOFT, SIGPRO, AIFZ, CSITY, CLSB - 2023
pp. 67-88, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131606

BERTology is an important study area for understanding the inner workings of large pre-trained models such as BERT. The BERT model is one of the most significant models in NLP, and the term ”BERTology” is widely used in the research community. Prominent recent research has included studies on the self-attention mechanism [30–33], the role of individual neurons in the BERT model [34], gradient-based methods [35], and the visualization of attention weights [36]. For a more comprehensive overview of the field, please refer to [83], who present a general overview of BERTology, and [68], who review analysis methods in neural language processing.

The importance of understanding linguistic information in language models (LMs) and the need for accurate and comprehensive linguistic information in LMs have been emphasized by [4], [18], [19], [20] [21].

More profound and fine-grained linguistic knowledge can help improve the models' performance and increase their explainability. Recent research in NLP and machine learning has underlined the critical need for explainability and interpretability [38–41]. Understanding complex models' decision-making processes, particularly with intricate data, poses a significant challenge, potentially impeding research progress. [73] further underscored the necessity of comprehending BERT models' inner workings to facilitate improvements and strides towards general AI.

Investigation of linguistic information through probing.

This review explores linguistic knowledge encoded in LMs, focusing specifically on studies employing probing tasks for their investigations. Probing tasks, as detailed by [42, 8, 43, 10], offer an intricate lens through which to scrutinize the capabilities of these models.

To elaborate, *probing* involves training a simplified classifier on the static representations produced by a language model, each targeting a particular linguistic task [44, 45]. The classifier ingests single-word representations, and the precision of the probing subsequently establishes the extent of linguistically relevant knowledge encoded in these representations. This operation can be carried out in a zero-shot scenario [46, 47], make use of structural probes [48], or, more conventionally, entail the training and assessment of uncomplicated classifiers on diagnostic tasks [45, 49].

The concept of *”linguistic probing”* is geared towards discerning how much a pre-trained model has gleaned about a given linguistic abstraction from the raw data, as portrayed by [50–52], and [68]. According to [69], this operation unfolds in several stages:

(1) choosing an annotated dataset that transforms the theoretical abstraction of interest into a predictive task, such as the Penn Treebank [53] adapted to Stanford dependencies or the DM corpus from [54]'s shared task [55]; (2) pre-training the model, for instance, RoBERTa or BERT; (3) training a ”ceiling” model with refined representations that serve as a benchmark for optimal performance with pre-trained representations; (4) training a supervised ”probe” model with the pre-trained, static representations, typically utilizing a simple, low-capacity model like a linear classifier; and (5) contrasting the performance of the probe model on unseen data with the ceiling model, thereby estimating how much the pre-trained model is inherently capable of performing the task or what pertinent features it can reveal to the probe model.

It should be clarified that this review focuses not on the specific probing techniques or the detailed processes used in the studies. Instead, the principal objective is to analyze the nature

and extent of linguistic knowledge absorbed by LMs.

This review's choice of papers investigating linguistic features of LMs, specifically through probing, was deliberate. Probing tasks, due to their fine-grained analysis and specific focus on linguistic knowledge, provide an excellent lens to study these models' complex, often opaque, learning processes. Moreover, probing provides a controlled envi- ronment for inspecting LMs, enabling exploring particular linguistic features in isolation. This targeted focus facilitates understanding whether and to what extent these models have successfully learned to encode specific linguistic information during training.

However, we also recognize that probing is one tool in the evaluation toolkit, and its results must be interpreted in context([14], [15], [16]). To complement probing, future work might investigate model competence using tasks that require integrating multiple types of linguistic knowledge.

Objectives.

The purpose of this systematic literature review is to summarize state of the art in using linguistic knowledge of pre-trained language models since 2018[1]. The review aims to identify gaps in current research and propose areas for further study. Additionally, the review aims to provide a theoretical basis for further developing pre- trained language models and contributes to research on the explainability of LMs. The research questions of this review are:

- RQ1: Which linguistic phenomena have been investigated in pre-trained language mod-els since 2018?
- RQ2: Which language models were mostly investigated?
- RQ3: Which natural languages were investigated? Which languages might need moreresearch?
- RQ4: Which mode of language is the current focus of the research?
- RQ5: Which linguistic phenomena should be considered while probing for linguistic knowledge in language models?

  The scope and goals of the review are:
- Summary of existing knowledge on the linguistic knowledge of language models.
- Identify gaps in current research to suggest areas for further investigation.
- Providing a framework/background for further development of pre-trained language models.
- Contribution to the newly developed area of interpretability and explainability of com- putational linguistics, seeking to understand how models capture natural language phenomena [57, 68, 58].
- Attempt to create a comprehensive overview assisting and encouraging future studies on the linguistic knowledge of language models.

By addressing these questions and goals, this review aims to advance our understanding of the linguistic knowledge encoded in LMs and contribute to the broader goal of creating more explainable and interpretable AI systems.

2. Related secondary studies

While this systematic literature review is the first comprehensive examination of the lin- guistic features represented in LMs using probing methods, it is essential to acknowledge and highlight other relevant research in this field.

[10] used an edge probing technique to determine the degree to which the BERT model contains linguistic information, revealing a sequence of linguistic processing capabilities, from part-of-speech tagging to coreference resolution.

[11] offered a comprehensive survey of over 150 existing literature pieces on the BERT model. The study explored the model's architecture, pre-training, and fine-tuning pro- cesses, and discussed its overparameterization issue, suggesting compression techniques such as knowledge distillation and pruning.

[12] assessed the transition from traditional Distributional Semantic Models (DSMs) to deep learning-based representations in NLP. The paper argued that while models like BERT perform well in lexical semantics, debates over type representations persist, and handling complex commonsense knowledge remains a challenge.

[13] evaluated deep neural networks' syntactic capabilities in NLP tasks. While DNNs demonstrated substantial syntactic understanding, they fell short of human competence.

[1]Release year of BERT. See [67]

These studies provide valuable insights into different aspects of LMs, including their representational abilities, architecture, parameterization, compression techniques, and syn- tactic and semantic knowledge. While this systematic literature review focuses specifically on probing methods and linguistic phenomena, these secondary studies contribute to the broader understanding of LMs and their linguistic capabilities.

3. REVIEW METHODS

This systematic literature review was conducted according to guidelines set by [1], [5], [6], and divided into three phases: review planning, conducting, and reporting. The review structure follows the PRISMA 2020 Checklist [84].

3.1.Eligibility Criteria

Inclusion criteria. The following types of papers will be included:

- Studies that used linguistic probing techniques to investigate the linguistic knowledgeof a pre-trained language model.
- The language of the article is either English or German.
- The article was published after January 2018.
- Full content of the article is available online for academic use.
- The study is related to the specified search keywords.
- The study is a peer-reviewed publication.

Exclusion criteria. The following types of papers will be excluded:

- Studies that did not use linguistic probing techniques to investigate the linguistic knowledge of a pre-trained language model.
- Articles written in languages other than English or German.
- Articles published before January 2018.
- Full content of the article is unavailable.
- The study is not related to the specified search keywords.
- The study is not a peer-reviewed publication.

The choice of these inclusion and exclusion criteria helps us to ensure that our systematic literature review is focused on a specific area of research (1) and captures the most relevant

and current information (2). Additionally, the review is able to focus on recent studies that are most relevant to the current state of the field (3). This is particularly important since BERT, being one of the first pre-trained language models, was introduced in 2018 by [67] after which the field of BERTology slowly started to emerge. Requiring that the full content of the article is available online and that the study is related to the specified search keywords also helps to ensure that the review includes only accessible, high-quality, and relevant studies (4). By requiring that the studies be peer-reviewed (5), the review can exclude studies that may not have undergone rigorous review and evaluation by experts in the field, helping to ensure that the review includes only studies that have been evaluated and considered trustworthy by other experts in the area.

3.2.Information sources

The review involved a comprehensive automatic search of multiple databases, including *ACM, Scopus, and IEEE* and a semi-manual search of *ACL, BlackboxNLP, and COLING 2022*. The sites were accessed as seen in Table 1 in the Appendix.

3.3.Search strategy

The search strategy was developed following the PRESS 2015 Checklist mcgowan2016press and the critical search points from the Cochrane Handbook higgins2019cochrane. The search terms were organized into four clusters: TOPIC, TASK, DOMAIN, and METHOD. The final search terms used can be found in Table 2 in the Appendix. The search was limited to articles published in English or German between October 2018 and October 2022.

3.4.Selection process

The initial search identified 455 studies, which were then screened for duplicates, resulting in 408 articles for further evaluation. The inclusion and exclusion criteria were applied to select a final set of 57 studies for inclusion in the review. Two independent reviewers conducted the selection process, with any discrepancies resolved through discussion. The study selection process followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Studies focusing solely on developing and improving probing techniques without investigating linguistic information and investigating probing for factual knowledge were excluded. Language-vision models were also not considered due to the limited scope of the review.

3.5.Data collection process

Two independent reviewers performed data extraction using a pre-designed data extraction form. The extracted data included the title, year of publication, authors, main findings, model investigated, analyzed linguistic phenomena and their linguistic level, mode of lan- guage (written vs. spoken), the natural language analyzed, and the main findings of the studies reviewed. The relevance of each extracted data field to the research questions is provided in Table 3 in the Appendix.

3.6.Synthesis methods

A thematic analysis was conducted on the data obtained from the selected studies to identify prevalent trends and themes. This comprehensive synthesis analyzed various as- pects, including the models under investigation, linguistic phenomena and levels, language modes,

key findings, and the natural languages examined.

This rigorous examination helped identify and categorize key findings and trends in the literature, contributing to a holistic understanding of linguistic features in pre-trained language models and addressing the review's research questions.

A summary table was constructed to enhance data presentation, providing an overview of the included studies[2]. This resource enables easy comparison and reference, benefiting researchers interested in the linguistic knowledge of pre-trained language models.

This synthesis methodology, grounded in robust data analysis, offers valuable insights into pre-trained language models' linguistic capabilities.

4. RESULTS AND DISCUSSION

After synthesizing the corpus, we can now answer our research questions. Subsequently, we present identified research gaps that open further directions for future work. The relevant papers that were included in our survey will be referred to with their ID numbers (see Table 4 in the Appendix.

2 The extracted data will be published upon completion.

4.1. Answers to research questions

Based on the findings of our literature review, we will now address the research questions posed in this review.

Findings related to RQ1: Which linguistic phenomena have been investigated in pre-trained language models since 2018?

Research on pre-trained language mod- els (PTLMs) has made significant strides since 2018, offering new insights into their ca- pabilities and their encoded linguistic knowledge. Analyzing the provided dataset offers a comprehensive overview of the predominant research areas and the linguistic phenomena under investigation in PTLMs.

Our analysis shows that the area of **semantics** has seen considerable interest in PTLM research. Numerous phenomena such as metaphor identification [ID1], noun properties and entailment [ID2], semantic attributes and values [ID3], predicate-argument structures and semantic role labeling [ID4], idiomatic meanings of noun compounds [ID5], lexical entailment and negation [ID6], and similes [ID8] have been investigated. These studies significantly contribute to understanding how PTLMs encode and represent semantic re- lationships.

Similarly, investigations within the domain of **syntax** have been prevalent, with a focus on syntactic dependencies [ID9, ID10, ID11, ID12] and part-of-speech tagging [ID13, ID14]. These research initiatives illuminate PTLMs' grasp of sentence structure and their ability to generate grammatically correct sentences. Several studies also probe PTLMs' understanding of hypernymy, hyponymy, synonymy, and meronymy [ID15, ID16], revealing their ability to encode and use both semantic and syntactic relationships.

Discourse analysis and **pragmatics** have been explored less frequently, yet they provide valuable insights into PTLMs' ability to comprehend context, speaker intentions, and discourse coherence [ID17]. Similarly, studying **morphology** and **orthography** can shed light on PTLMs' knowledge of word forms, inflection, and spelling variations [ID18, ID19]. Analyzing typological properties can gauge PTLMs' cross-linguistic generalization of linguistic knowledge [ID20], and research on **textual analysis** and **speech recognition** can improve the robustness and context-awareness of PTLMs [ID21]. For a detailed visual representation of investigated linguistic levels and specific linguistic features alongside with the ID numbers of the corresponding papers, consider Figure 3 in A and for the distribution of papers investigating different linguistic levels, see Figure 1.

Although PTLMs can handle syntactic and semantic information reasonably well, abias towards syntax is noticeable [ID18]. This bias is evident in the implicit embedding of syntax trees in deep models' vector geometry [ID17], contrasting with specific models' failure to encapsulate the dependency tree structure, such as in the case of Multimodal-BERT [ID34]. This skew towards syntax underlines a significant challenge in NLP: creating models that excel in both syntactic and semantic understanding.

Remarkably, state-of-the-art LMs such as BERT, mBERT, and ELMo exhibit an im- pressive, incomplete capability to encapsulate linguistic phenomena. They adeptly en- code metaphorical [ID1] and syntactic knowledge [ID17, ID18, ID35], underlined by the structured dependency relationships and linguistic features embedded within these models [ID16]. However, limitations are evident in these models' understanding of specific lin- guistic phenomena such as negation, sarcasm, irony, modality, idioms, and non-standard spelling [ID3] and in identifying semantic equivalence between distinct lexical references to the same concept [ID4].

There are notable disparities in how LMs handle different levels of linguistic knowl- edge. They exhibit proficiency in extracting semantic proto-role properties [ID25], encoding contextualized representations of predicates [ID9], and handling agentive positions [ID19]. Still, they struggle with consistently capturing semantics [ID56] and show instability in cat- egories involving lexical semantics, logic, and predicate-argument structure [ID52]. These disparities highlight the ongoing challenge of developing models that understand linguistic nuances as well as humans do.

However, the flexibility of these models indicates promising future progress. For in- stance, while having a minimal impact on core and low-level information, fine-tuning methodologies improve the performance of models like BERT on downstream tasks [ID7, ID42]. Larger models outperform smaller ones at character-level tasks [ID20], and the emergence of the feature-based approach in lexical semantics [ID6] reveals potential direc- tions for future advancements.

Regarding cross-lingual proficiency, these models demonstrate good knowledge trans- ferability between languages and datasets, provided annotation consistency exists [ID1]. However, potential biases are discernible, such as mBERT's tendency to mirror the infor- mation distribution for languages typologically similar to English [ID28] and its limited semantic consistency across languages [ID39].

Despite the progress made in machine-based language comprehension, a significant gap exists between the performance of these models and human linguistic proficiency, especially for tasks requiring a deeper understanding of linguistic knowledge [ID14, ID45, ID47]. The challenge lies in bridging these 'gaps' in the linguistic knowledge obtained from single

pretraining objectives and in exploring the combination of diverse pretraining goals for a more holistic understanding [ID21].

Future work should not only aim to build models that can perform specific tasks but also to create systems capable of nuanced, human-like understanding of language. Although this might be challenging, it represents an exciting frontier for exploration and innovation in the field.

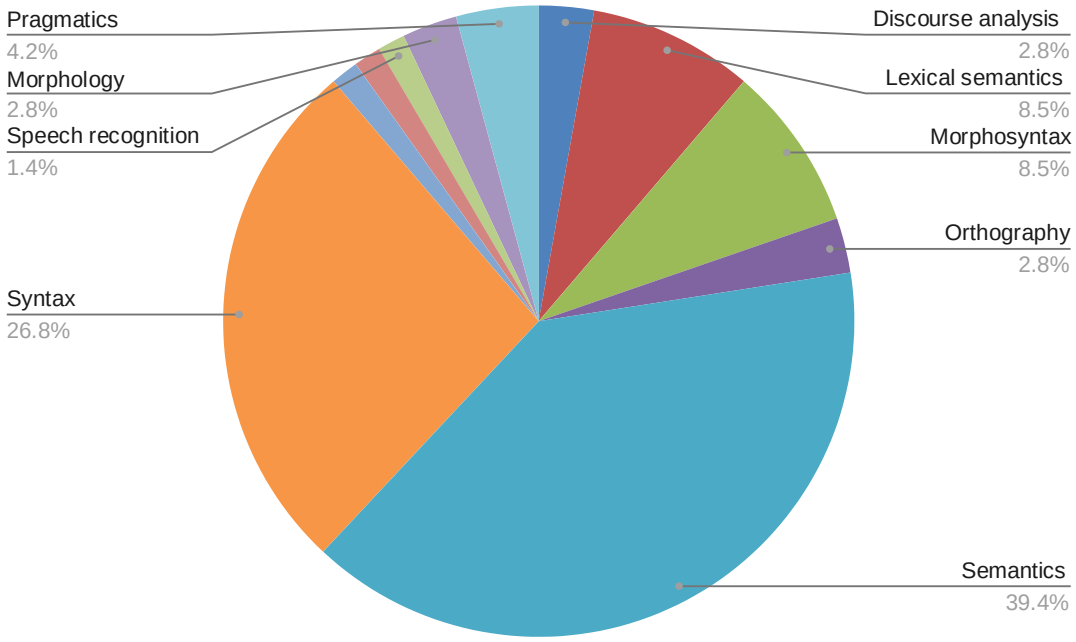

Fig. 1. The distribution of investigated linguistic levels.

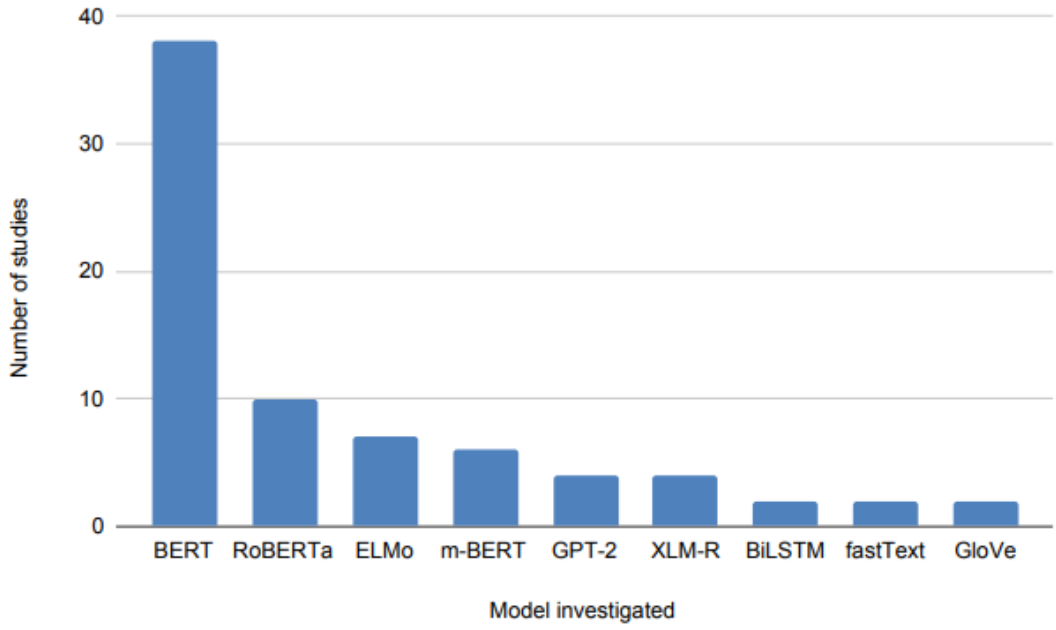

Fig. 2. Language models investigated. (Only models investigated by more than one study are depicted.)

Findings related to RQ2: Which language models were mostly investigated?

In recent years, numerous pre-trained language models (LMs) have been the subject of an investigation to explore their linguistic capabilities. Analyzing the dataset provided has revealed valuable insights into the LMs that garnered the most attention within the research community.

Our findings highlight that the research community has primarily focused on probing the linguistic knowledge of three prominent LMs: BERT, RoBERTa, ELMo, and mBERT. These models have been extensively studied across various linguistic tasks, including semantic analysis, syntactic parsing, and discourse analysis.

While BERT, Roberta, ELMo, and mBERT have received the most attention, other LMs have also been explored to a lesser extent. These include GPT2, XLM-R, BiLSTM, fastText, and GloVe. While not as extensively studied as the models mentioned above, these LMs have also made valuable contributions to our understanding of language models' linguistic capabilities. For a more detailed overview, please, consider Figure 2.

The comprehensive exploration of these LMs has shed light on their proficiency in understanding semantics, syntax, and discourse. However, it is crucial to acknowledge that the full extent of the linguistic knowledge encoded within these models is yet to be fully comprehended. The field of pre-trained language models continues to evolve rapidly, with the emergence of novel models regularly. Hence, future investigations may involve exploring the linguistic properties of these newer models and comparing them to the more established ones.

Findings related to RQ3: Which natural languages were investigated? Which languages might need more research?

Numerous studies have extensively explored English LMs and their variants, to assess their linguistic knowledge and performance on various tasks. This emphasis on English reflects its dominance as a global language and the availability of large-scale English corpora for training LMs.

However, in addition to English, several other languages have been investigated, in- cluding German, French, Spanish, Chinese, Italian, Portuguese, Dutch, Russian, Turkish, Arabic, and Finnish. Research on these languages has aimed to evaluate the effectiveness of LMs in capturing language-specific linguistic phenomena and understanding the nuances of diverse linguistic structures.

However, the results highlight specific languages that may require further research attention. Languages such as Hindi, Japanese, Korean, Swahili, and Indonesian have had relatively fewer studies investigating their linguistic properties within the context of pre- trained language models. Conducting more research on these languages would provide valuable insights into how well LMs capture the intricacies of their specific linguistic characteristics. For a more detailed overview, please, consider Figure 4 in the Appendix.

Furthermore, low-resource languages also warrant additional investigation. These lan- guages may include regional dialects, indigenous languages, or languages with fewer digital resources. Exploring the behavior and performance of LMs on these languages can help bridge the existing gaps and promote inclusivity in natural language processing research and applications.

Findings related to RQ4: Which mode of language is the current focus of the research?

Based on our review, it is evident that most studies have focused on written text rather than speech. The results strongly emphasize exploring the linguistic capabilities of pre-trained language models in written languages. These studies have delved into various linguistic phenomena in written language already thoroughly discussed in 4.1.

However, it is essential to acknowledge that research on the linguistic capabilities of pre-trained language models in speech is also significant, although relatively less prevalent than written text. Therefore, further exploration of this modality is necessary to understand the models' performance in transcribing spoken language and capturing phonetic, prosodic, and intonational features.

Findings related to RQ5: Which linguistic phenomena should be consideredwhile probing for linguistic knowledge in language models?

The proficiency of LMs in comprehending and producing natural language heavily depends on their ability to handle various linguistic phenomena, as highlighted by [70]. This idea resonates with the "rediscovery hypothesis" put forth by [18], which emphasizes the importance of in-depth linguistic understanding for maximizing language model performance.

In light of our research findings, we suggest conducting a focused investigation into the following linguistic phenomena.

Nuances of expression: This includes comprehension of sarcasm, irony [ID3], id- iomatic meanings [ID11], and non-standard spelling. Improving models' handling of these nuances is key to robust language understanding.

Cross-linguistic transferability: Investigating the generalizability and adaptability of LMs across diverse languages can uncover ways to enhance cross-lingual performance [ID1, ID29].

Semantic complexity: Complex semantic structures, including semantic roles and predicate-argument structures, often pose challenges to LMs [ID9, ID25]. Probing these areas can yield insights for improved semantic understanding.

Interplay of syntax and pragmatics: LMs need to handle pragmatic phenomena like negation, conversational implicatures, and other speech acts efficiently for real-world communication [ID13, ID23].

Morphological features: Research into morphological phenomena, including inflec- tional morphology and derivational processes, can enhance models' capability for morpho- logical ambiguity resolution [ID15, ID44].

Discourse understanding: Language models' ability to understand hierarchical struc- tures, coherence relations, and global coherence within texts significantly influences their capacity for coherent language generation [ID22].

Semantic richness: Investigating more complex lexical phenomena can help improve the depth and accuracy of models' semantic representations [ID6, ID49].

Language typology: The study of distinct typological features can uncover potential biases and improve language model performance across diverse language families [ID39].

Contextual pragmatics: Enhancing models' understanding of discourse through in- depth studies of contextual pragmatics, including anaphora resolution and presupposition projection, can further improve their contextual awareness [ID7].

This comprehensive investigation can facilitate the development of linguistically richer, more accurate, and explainable LMs, propelling advancements in the field of NLP.

4.2.Implications for future research

The findings outlined in the preceding section have illuminated our understanding of the current state of linguistic knowledge encoded within LMs. These insights also provide a robust foundation for guiding future research within NLP, mainly as we aim to advance the capabilities of LMs and enhance their linguistic comprehension.

Delving deeper into complex linguistic phenomena like negation, sarcasm, irony, id- iomatic expressions, and non-standard spelling [ID3] is vital. The observed limitations of current models highlight the necessity for innovative training methodologies, augmentation of training data with diverse linguistic examples, and the design of architectures tailored to handle such intricate linguistic challenges.

Simultaneously, the gap in models' abilities to understand semantic structures [ID4], coupled with their underperformance in semantic tasks compared to syntactic ones [ID56], accentuates the need for research to enhance semantic representation within these models. The stratified nature of LMs provides a ripe opportunity for exploration. Unraveling the rationale behind the distribution of linguistic knowledge across various layers [ID18, ID35] and strategizing how this knowledge can be effectively utilized for a range of downstream tasks is an enticing research direction.

Furthermore, the ability of LMs to transfer linguistic knowledge across languages [ID1, ID28, ID39] presents exciting prospects for building models with better cross-lingual generalizability. Future research could focus on facilitating cross-lingual knowledge transfer, addressing language representation biases, and enhancing the semantic consistency of multilingual models.

An intriguing line of inquiry also examines how different pretraining objectives could be harmoniously combined, given the gaps in linguistic knowledge derived from single pretraining objectives [ID21]. This could entail the development of novel pretraining techniques that leverage a broad spectrum of linguistic goals, capturing both syntactic and semantic knowledge to cultivate more comprehensive and robust LMs.

Moreover, future research should aim to diversify the linguistic coverage of LMs, addressing the mentioned underrepresented languages and focusing on low-resource languages to enable the development of more inclusive and robust natural language processing systems.

Lastly, the inclusion of linguistic formalism into NLP research ([4], [18], [19], [20], [21]) promises to enhance the transparency and systematicity of research. Similarly, focusing on finer-grained linguistic features in future probing studies and exploring less-studied linguistic areas will assist in a more profound understanding of language models' capabilities. By addressing these issues and harnessing the potential of these opportunities, we can aspire to develop LMs that offer a more nuanced and sophisticated understanding of human language.

5. CONCLUSION

In this systematic literature review, we evaluated the linguistic knowledge of LMs through linguistic probing tasks. Our findings indicate that LMs possess a high level of linguistic expertise, but there are still limitations and gaps in their understanding of specific linguistic phenomena. We also identified a need for consensus in evaluating their linguistic knowledge and terminology.

This review provides an unbiased overview of current research on language models' linguistic capabilities and their potential for natural language processing applications and explainable AI systems. We highlight the importance of future research focusing on fine-grained linguistic analysis, incorporating linguistic formalism, and considering additional linguistic features. We recommend providing clear annotation guidelines to address these limitations and ensure NLP research transparency and systematicity.

In conclusion, this review underscores the significance of linguistic knowledge in enhancing language models' accuracy, performance, and interpretability. By addressing the identified limitations and adopting a linguistically-informed approach, researchers can advance our understanding of LMs and contribute to developing more robust and explainable AI systems. Exploring fine-grained linguistic features, incorporating linguistic formalism, and

considering a wide range of linguistic phenomena will drive innovation and practical applications across domains and industries.

6. LIMITATIONS

While this systematic literature review provides valuable insights into the linguistic knowledge of LMs, certain limitations should be acknowledged. As with any review, there is the potential for subjectivity or misunderstanding in the filtering and data extraction process. Future studies could benefit from involving a larger team of reviewers and implementing inter-rater reliability measures to address this limitation to ensure consistency.

Another limitation is the lack of emphasis on systematic linguistic features in some of the studies included in this review. Future studies should strive for a more linguistically-informed approach to overcome this limitation, incorporating linguistic formalism and considering a broader range of linguistic phenomena.

Furthermore, this review focused primarily on published studies in English and Ger- man. While efforts were made to include a diverse range of studies, valuable research in other languages may not be captured in this review. Future reviews should consider ex- panding the scope to include studies in various languages to ensure a more comprehensive understanding of language models' linguistic capabilities.

Lastly, it is essential to note that LMs and linguistic probing are rapidly evolving. The studies included in this review were conducted until a specific date, and new research may have emerged since then.

REFERENCES

[1] B. Kitchenham and S. Charters, Guidelines for performing systematic literature reviews in software engineering. Citeseer, 2007.

[2] A. Pollock and E. Berge, "How to do a systematic review", International Journal of Stroke, vol. 13, no.2, pp. 138-156, 2018. SAGE Publications Sage UK: London, England.

[3] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J.Devereaux, J. Kleijnen, and D. Moher, "The PRISMA statement for reporting systematic reviews andmeta-analyses of studies that evaluate health care interventions: explanation and elaboration", Journalof clinical epidemiology, vol. 62, no. 10, pp. e1-e34, 2009. Elsevier.

[4] I. Kuznetsov and I. Gurevych, "A matter of framing: The impact of linguistic formalism on probingresults", arXiv preprint arXiv:2004.14999, 2020.

[5] J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch, Cochranehandbook for systematic reviews of interventions. John Wiley Sons, 2019.

[6] J. McGowan, M. Sampson, D. M. Salzwedel, E. Cogo, V. Foerster, and C. Lefebvre, "PRESS peerreview of electronic search strategies: 2015 guideline statement", Journal of clinical epidemiology, vol.75, pp. 40-46, 2016. Elsevier.

[7] A. Ettinger, A. Elgohary, and P. Resnik, "Probing for semantic evidence of composition by means ofsimple classification tasks", Proceedings of the 1st workshop on evaluating vector-space representations for nlp, pp. 134-139, 2016.

[8] Y. Belinkov, "Probing classifiers: Promises, shortcomings, and advances", Computational Linguistics,vol. 48, no. 1, pp. 207-219, 2022. MIT Press.

[9] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. Van Durme, S. R. Bowman,D. Das, et al., "What do you learn from context? probing for sentence structure in contextualized word representations", arXiv preprint arXiv:1905.06316, 2019.

[10] I. Tenney, D. Das, and E. Pavlick, "BERT Rediscovers the Classical NLP Pipeline", Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4593-4601, 2019.

[11] A. Rogers, O. Kovaleva, and A. Rumshisky, ”A primer in bertology: What we know about how bert works”, Transactions of the Association for Computational Linguistics, vol. 8, pp. 842-866, 2020. MIT Press.
[12] E. Pavlick, ”Semantic structure in deep learning”, Annual Review of Linguistics, vol. 8, pp. 447-471, 2022. Annual Reviews.
[13] T. Linzen and M. Baroni, ”Syntactic structure from deep learning”, Annual Review of Linguistics, vol. 7, pp. 195-212, 2021. Annual Reviews.
[14] A. Miaschi, C. Alzetta, D. Brunato, F. Dell’Orletta, and G. Venturi, ”Probing Tasks Under Pressure”,CLiC-it, 2021.
[15] A. Warstadt, Y. Cao, I. Grosu, W. Peng, H. Blix, Y. Nie, A. Alsop, S. Bordia, H. Liu, A. Parrish, et al., ”Investigating BERT’s knowledge of language: five analysis methods with NPIs”, arXiv preprintarXiv:1909.02597, 2019.
[16] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, ”What you can cram into a single vector: Probing sentence embeddings for linguistic properties”, arXiv preprint arXiv:1805.01070, 2018.
[17] I. K. Raharjana, D. Siahaan, and C. Fatichah, ”User stories and natural language processing: A systematic literature review.” IEEE Access, vol. 9, pp. 53811-53826, 2021.
[18] V. Nikoulina, M. Tezekbayev, N. Kozhakhmet, M. Babazhanova, M. Gall´e, and Z. Assylbekov, ”Therediscovery hypothesis: Language models need to meet linguistics.” Journal of Artificial IntelligenceResearch, vol. 72, pp. 1343-1384, 2021.
[19] M. Baroni, ”On the proper role of linguistically-oriented deep net analysis in linguistic theorizing.”arXiv preprint arXiv:2106.08694, 2021.
[20] I. Kuznetsov, ”The Role of Linguistics in Probing Task Design.” Technische Universit¨at, 2021.
[21] B. Li, ”Integrating Linguistic Theory and Neural Language Models.” Ph.D. dissertation, University ofToronto (Canada), 2022.
[22] Y. Belinkov and J. Glass, ”Analysis methods in neural language processing: A survey.” Transactionsof the Association for Computational Linguistics, vol. 7, pp. 49-72, 2019.
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin,”Attention is all you need.” Advances in neural information processing systems, vol. 30, 2017.
[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805, 2018.
[25] M. Ott, Gilles Deleuze zur Einf¨uhrung. Junius Verlag, 2018.
[26] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, and others, ”Improving language understandingby generative pre-training.” OpenAI, 2018.
[27] E. Strubell, P. Verga, D. Andor, D. Weiss, and A. McCallum, ”Linguistically-informed self-attentionfor semantic role labeling.” arXiv preprint arXiv:1804.08199, 2018.
[28] M. Lewis, M. Ghazvininejad, G. Ghosh, A. Aghajanyan, S. Wang, and L. Zettlemoyer, ”Pre-trainingvia paraphrasing.” Advances in Neural Information Processing Systems, vol. 33, pp. 18470-18481, 2020.
[29] L. Lan, D. Xu, G. Ye, C. Xia, S. Wang, Y. Li, and H. Xu, ”Positive RT-PCR test results in patientsrecovered from COVID-19.” Jama, vol. 323, no. 15, pp. 1502-1503, 2020.
[30] L. D. Zambrano, S. Ellington, P. Strid, R. R. Galang, T. Oduyebo, V. T. Tong, K. R. Woodworth,J. F. Nahabedian III, E. Azziz-Baumgartner, S. M. Gilboa, and others, ”Update: characteristics ofsymptomatic women of reproductive age with laboratory-confirmed SARS-CoV-2 infection by pregnancy status—United States, January 22–October 3, 2020.” Morbidity and Mortality Weekly Report, vol. 69,no. 44, pp. 1641, 2020.
[31] M. Bi´nkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K.Simonyan, ”High fidelity speech synthesis with adversarial networks.” arXiv preprint arXiv:1909.11646,2019.
[32] S. P. H. Alexander, A. Christopoulos, A. P. Davenport, E. Kelly, A. Mathie, J. A. Peters, E. L. Veale,J. F. Armstrong, E. Faccenda, S. D. Harding, and others, ”The Concise Guide to PHARMACOLOGY2019/20: G protein-coupled receptors.” British journal of pharmacology, vol. 176, pp. S21-S141, 2019.
[33] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, ”Revealing the dark secrets of BERT.” arXivpreprint arXiv:1908.08593, 2019.

[34] D. A. Case, H. M. Aktulga, K. Belfon, I. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. CheathamIII, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, and others, ”Amber 2021.” University of California,San Francisco, 2021.

[35] M. Aringer, K. Costenbader, D. Daikh, R. Brinks, M. Mosca, R. Ramsey-Goldman, J. S. Smolen, D.Wofsy, D. T. Boumpas, D. L. Kamen, et al., ”2019 European League Against Rheumatism/AmericanCollege of Rheumatology classification criteria for systemic lupus erythematosus,” Arthritis & rheumatology, vol. 71, no. 9, pp. 1400–1412, 2019. Wiley Online Library.

[36] J. Vig, ”A multiscale visualization of attention in the transformer model,” arXiv preprintarXiv:1906.05714, 2019.

[37] A. Nott, I. R. Holtman, N. G. Coufal, J. C. M. Schlachetzki, M. Yu, R. Hu, C. Z. Han, M. Pena, J.Xiao, Y. Wu, et al., ”Brain cell type–specific enhancer–promoter interactome maps and disease-risk association,” Science, vol. 366, no. 6469, pp. 1134–1139, 2019. American Association for the Advancementof Science.

[38] M. T. Ribeiro, S. Singh, C. Guestrin, ”Model-agnostic interpretability of machine learning,” arXivpreprint arXiv:1606.05386, 2016.

[39] L. F. R. Ribeiro, M. Schmitt, H. Sch¨utze, I. Gurevych, ”Investigating pretrained language models forgraph-to-text generation,” arXiv preprint arXiv:2007.08426, 2020.

[40] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu,T. Berah, J. W. Eaton, et al., ”Estimating the effects of non-pharmaceutical interventions on COVID-19in Europe,” Nature, vol. 584, no. 7820, pp. 257–261, 2020. Nature Publishing Group.

[41] F. C. Krause, E. Linardatos, D. M. Fresco, M. T. Moore, ”Facial emotion recognition in major depressive disorder: A meta-analytic review,” Journal of Affective Disorders, vol. 293, pp. 320–328, 2021.Elsevier.

[42] S. J. Ettinger, E. C. Feldman, E. Cote, Textbook of Veterinary Internal Medicine-Inkling E-Book, 2016.Elsevier health sciences.

[43] Z. Agic, I. Vulic, ”JW300: A wide-coverage parallel corpus for low-resource languages,” In ˇ Proceedingsof the Association for Computational Linguistics, 2019.

[44] T. Linzen, ”Issues in evaluating semantic spaces using word analogies,” arXiv preprintarXiv:1606.07736, 2016.

[45] Y. Belinkov, Y. Bisk, ”Synthetic and natural noise both break neural machine translation,” arXivpreprint arXiv:1711.02173, 2017.

[46] W. S. El-Deiry, R. M. Goldberg, H.-J. Lenz, A. F. Shields, G. T. Gibney, A. R. Tan, J. Brown, B.Eisenberg, E. I. Heath, S. Phuphanich, et al., ”The current state of molecular testing in the treatmentof patients with solid tumors, 2019,” CA: a cancer journal for clinicians, vol. 69, no. 4, pp. 305–343,2019. Wiley Online Library.

[47] B. Rochwerg, S. Einav, D. Chaudhuri, J. Mancebo, T. Mauri, Y. Helviz, E. C. Goligher, S. Jaber, J.-D.Ricard, N. Rittayamai, et al., ”The role for high flow nasal cannula as a respiratory support strategyin adults: a clinical practice guideline,” Intensive care medicine, vol. 46, no. 12, pp. 2226–2237, 2020.Springer.

[48] J. Hewitt and C. D. Manning, ”A structural probe for finding syntax in word representations,” InProceedings of the 2019 Conference of the North American Chapter of the Association for ComputationalLinguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4129–4138, 2019.

[49] J. Hewitt and P. Liang, ”Designing and Interpreting Probes with Control Tasks,” Proceedings of the2019 Con, 2019.

[50] C. Lu, L. Niu, N. Chen, K. Jin, T. Yang, P. Xiu, Y. Zhang, F. Gao, H. Bei, S. Shi, et al., ”Enhancingradiation tolerance by controlling defect mobility and migration pathways in multicomponent singlephase alloys,” Nature communications, vol. 7, no. 1, pp. 1–8, 2016. Nature Publishing Group.

[51] R. Janot, ”ADI n. 227.175/2017-AsJConst/SAJ/PGR,” Revista Brasileira de Direito Animal, vol. 12,no. 03, 2017.

[52] M. Giulianelli, J. Harding, F. Mohnert, D. Hupkes, and W. Zuidema, ”Under the hood: Using diagnosticclassifiers to investigate and improve how language models track agreement information,” arXiv preprintarXiv:1808.08079, 2018.

[53] R. A. Marcus, ”Elektronentransferreaktionen in der Chemie-Theorie und Experiment (NobelVortrag),” Angewandte Chemie, vol. 105, no. 8, pp. 1161–1172, 1993. Wiley Online Library.
[54] N. Xue, H. T. Ng, S. Pradhan, R. Prasad, C. Bryant, and A. Rutherford, ”The conll-2015 shared taskon shallow discourse parsing,” In Proceedings of the Nineteenth Conference on Computational NaturalLanguage Learning-Shared Task, pp. 1–16, 2015.
[55] S. Oepen, M. Kuhlmann, Y. Miyao, D. Zeman, S. Cinkov´a, D. Flickinger, J. Hajic, and Z. Uresova,”Semeval 2015 task 18: Broad-coverage semantic dependency parsing,” In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 915–926, 2015.
[56] J. E. Ebinger, J. Fert-Bober, I. Printsev, M. Wu, N. Sun, J. C. Prostko, E. C. Frias, J. L. Stewart, J.E. Van Eyk, J. G. Braun, et al., ”Antibody responses to the BNT162b2 mRNA vaccine in individualspreviously infected with SARS-CoV-2,” Nature medicine, vol. 27, no. 6, pp. 981–984, 2021. NaturePublishing Group.
[57] M. Alishahi, M. Farzaneh, F. Ghaedrahmati, A. Nejabatdoust, A. Sarkaki, and S. E. Khoshnam,”NLRP3 inflammasome in ischemic stroke: as possible therapeutic target,” International Journal ofStroke, vol. 14, no. 6, pp. 574–591, 2019. SAGE Publications Sage UK: London, England.
[58] J. L. Prince-Guerra, O. Almendares, L. D. Nolen, J. KL. Gunn, A. P. Dale, S. A. Buono, M. DeutschFeldman, S. Suppiah, L. Hao, Y. Zeng, et al., ”Evaluation of Abbott BinaxNOW rapid antigen test forSARS-CoV-2 infection at two community-based testing sites—Pima County, Arizona, November 3–17,2020,” Morbidity and Mortality Weekly Report, vol. 70, no. 3, pp. 100, 2021. Centers for Disease Controland Prevention.
[59] E. Kalyaeva, O. Durandin, and A. Malafeev, ”Behavior of Modern Pre-trained Language Models Usingthe Example of Probing Tasks,” In Proceedings of the International Conference on Recent Advances inNatural Language Processing (RANLP 2021), pp. 664–670, 2021.
[60] E. Aghazadeh, M. Fayyaz, and Y. Yaghoobzadeh, ”Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages,” In Proceedings of the 60th Annual Meeting ofthe Association for Computational Linguistics (Volume 1: Long Papers), pp. 2037–2050, 2022. Association for Computational Linguistics.
[61] [No author given. Please provide the author details for a complete citation.]
[62] J. R. Searle, F. Kiefer, M. Bierwisch, et al., Speech act theory and pragmatics, Vol. 10. 1980. Springer.
[63] M. Foucault, Archaeology of knowledge. 2013. Routledge.
[64] N. Fairclough, ”The dialectics of discourse,” Textus, vol. 14, no. 2, pp. 231–242, 2001. Citeseer.
[65] T. A. Van Dijk, ”The study of discourse,” Discourse as structure and process, vol. 1, no. 34, pp. 703–52,1997.
[66] G. N. Leech, Principles of pragmatics. 2016. Routledge.
[67] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ”Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
[68] Y. Belinkov and J. Glass, ”Analysis methods in neural language processing: A survey,” Transactionsof the Association for Computational Linguistics, vol. 7, pp. 49–72, 2019. MIT Press.
[69] Z. Wu, H. Peng, and N. A. Smith, ”Infusing finetuning with semantic dependencies,” Transactions ofthe Association for Computational Linguistics, vol. 9, pp. 226–242, 2021. MIT Press.
[70] I. A. Sag, ”Linguistic theory and natural language processing,” In Natural language and speech, pp.69–83, 1991. Springer.
[71] L. Birnbaum, ”Let’s Put the AI Back in NLP,” In Theoretical Issues in Natural Language Processing3, 1987.
[72] M. Silverstein, ”Linguistic theory: syntax, semantics, pragmatics,” Annual review of Anthropology, pp.349–382, 1972. JSTOR.
[73] C. Schuster and S. Hegelich, ”From BERT‘s Point of View: Revealing the Prevailing Contextual Differences,” In Findings of the Association for Computational Linguistics: ACL 2022, pp. 1120–1138,2022.
[74] M. Apidianaki and A. Gar´ı Soler, ”ALL Dolphins Are Intelligent and SOME Are Friendly: ProbingBERT for Nouns’ Semantic Properties and their Prototypicality,” In Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pp. 79–94, 2021.

[75] M. Garcia, T. K. Vieira, C. Scarton, M. Idiart, and A. Villavicencio, ”Probing for idiomaticity

invector space models," In Proceedings of the 16th Conference of the European Chapter of the Associationfor Computational Linguistics, pp. 3551–3564, 2021. Association for Computational Linguistics (ACL).

[76] A. Geiger, K. Richardson, and C. Potts, "Neural Natural Language Inference Models Partially EmbedTheories of Lexical Entailment and Negation," In Proceedings of the Third BlackboxNLP Workshop onAnalyzing and Interpreting Neural Networks for NLP, pp. 163–173, 2020.

[77] Q. He, S. Cheng, Z. Li, R. Xie, and Y. Xiao, "Can Pre-trained Language Models Interpret Similes asSmart as Human?", arXiv preprint arXiv:2203.08452, 2022.

[78] E. Hernandez and J. Andreas, "The Low-Dimensional Linear Geometry of Contextualized Word Representations," In Proceedings of the 25th Conference on Computational Natural Language Learning, pp.82–93, 2021.

[79] J. Kunz and M. Kuhlmann, "Test Harder than You Train: Probing with Extrapolation Splits," InProceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks forNLP, pp. 15–25, 2021.

[80] K. Richardson and A. Sabharwal, "What does my qa model know? devising controlled probes usingexpert knowledge," Transactions of the Association for Computational Linguistics, vol. 8, pp. 572–588,2020. MIT Press.

[81] A. Ravichander, E. Hovy, K. Suleman, A. Trischler, and J. C. K. Cheung, "On the systematicity ofprobing contextualized word representations: The case of hypernymy in BERT," In Proceedings of theNinth Joint Conference on Lexical and Computational Semantics, pp. 88–102, 2020.

[82] A. Kumar, M. N. Sundararaman, and J. Vepa, "What BERT Based Language Models Learn in SpokenTranscripts: An Empirical Study," arXiv preprint arXiv:2109.01009, 2021.

[83] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in BERTology: What we know about how BERTworks, Transactions of the Association for Computational Linguistics, vol. 8, MIT Press One RogersStreet, Cambridge, MA 02142-1209, USA, 2021, pp. 842-866.

[84] D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L.A. Stewart, Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement,Systematic reviews, vol. 4, no. 1, BioMed Central, 2015, pp. 1-9.

AUTHORS

Hayastan Avetisyan completed her BA in Translation Studies in Yerevan, Armenia, and pursued an MA in Linguistics in Hannover, Germany, in 2020. In 2021, she joined the AI4S2 project at the Department of Research Infrastructure and Methods, focusing on NLP, ML, and AI interpretability. Her research explores leveraging linguistic knowledge to enhance the development and interpretation of language models. Currently pursuing her Ph.D., she investigates the utilization of AI in research methodologies.

David Broneske is the head of the department Infrastructure and Methods at the Ger- man Centre for Higher Education Research and Science Studies (DZHW), Hannover. He received his PhD in Computer Science from University of Magdeburg, where he also pur- sued his Master and Bachelor Studies in Computer Science. His research interests include main-memory database systems, interdisciplinary data management, and the application of artificial intelligence in various domains.

A Appendix

Table 1. Information sources

Source	Date	Search
ACM	04.10.22	automatic
Scopus	30.09.22	automatic
IEEE	03.10.22	automatic
ACL	05.10.22	semi-manual
BlackboxNLP	05.10.22	semi-manual
COLING 2022	13.10.22	semi-manual

Table 2. Keyword searches

Source	Keyword search
ACM	[[Full Text: ”linguistic features”] OR [Full Text: ”linguistic information”] OR [Full Text: ”linguistic”] OR [Full Text: ”linguistics”] OR [Full Text: ”linguistic knowledge”] OR [Full Text: ”semantic”] OR [Full Text: ”syntactic”] OR [Full Text: ”lexical”] OR [Full Text: ”pragmatic”]] AND [[Title: ”probing”] OR [Title: ”probe”] OR [Title: ”probes”] OR [Title: ”probed”]] AND [[Full Text: ”pre-trained”] OR [Full Text: ”language”] OR [Full Text: ”models”] OR [Full Text: ”lms”]] AND [Publication Date: (01/01/2018 TO *)]
Scopus	(TITLE-ABS-KEY (”linguistic features” OR ”linguistic information” OR ”linguis- tic” OR ”linguistics” OR ”linguistic knowledge” OR ”semantic” OR ”syntactic” OR ”lexical” OR ”pragmatic”) AND TITLE (”probing” OR ”probe” OR ”probes” OR ”probed”) AND TITLE-ABS-KEY (”pre-trained” OR ”language” OR ”models” OR”LMs”)) AND PUBYEAR ¿ 2017 AND (LIMIT-TO (SUBJAREA , ”COMP”))AND (LIMIT-TO (LANGUAGE , ”English”))
IEEE	(”Full Text & Metadata”:”linguistic features” OR ”Full Text & Metadata”:”linguistic information” OR ”Full Text & Metadata”:”linguistic” OR ”Full Text & Meta- data”:”linguistics” OR ”Full Text & Metadata”:”linguistic knowledge” OR ”Full Text& Metadata”:”semantic” OR ”Full Text & Metadata”:”syntactic” OR ”Full Text & Metadata”:”lexical” OR ”Full Text & Metadata”:”pragmatic”) AND (”Document Title”:”probing” OR ”Document Title”:”probe” OR ”Document Title”:”probes” OR ”Document Title”:”probed”) AND (”Full Text & Metadata”:”pre-trained” OR ”Full Text & Metadata”:”language” OR ”Full Text & Metadata”:”models” OR ”Full Text & Metadata”:”lms”)
BlackboxNLP	1) ”prob*” in title 2) no ”prob*” in title
ACL	1) ”probing” / ”probe” in title 2) ”probing” / ”probe” in abstract
COLING 2022	1) ”probing” / ”probe” in title 2) ”probing” / ”probe” in abstract

Table 3. Extraction form (following [17])

Data	Description	Relevance
Title		Overview
Year		Overview
Authors		Overview
Model	Which model was investigated?	RQ2
Ling. phenomena	Which linguistic information was investigated?	RQ1, RQ5
Ling. level	What linguistic level was the analysis at?	RQ1, RQ5
Language and its mode	What language was investigated? Was it written or spo- ken?	RQ3, RQ4

Fig. 3. The distribution of investigated linguistic levels and linguistic features with the ID numbers of thecorresponding papers.

Table 4. List of the studies reviewed for the survey.

ID	Author(s)	Title
1	Aghazadeh, Ehsan; Fayyaz, Mohsen; Yaghoobzadeh, Yadollah	Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages
2	Apidianaki, Marianna; Gar\'\i Soler, Aina	ALL Dolphins Are Intelligent and SOME Are Friendly: Probing BERT for Nouns' Semantic Properties and their Prototypicality
3	Barnes, Jeremy; Øvrelid, Lilja; Velldal, Erik	Sentiment Analysis Is Not Solved! Assessing and Probing Sentiment Classification
4	Beloucif, Meriem; Biemann, Chris	Probing Pre-trained Language Models for Semantic Attributes and their Values
5	Biddle, Rhys; Rybinski, Maciek; Li, Qian; Paris, Cecile; Xu, Guandong	Harnessing Privileged Information for Hyperbole Detection
6	Branco, Ant´onio; Ant´onio Rodrigues, Joao; Salawa, Malgorzata; Branco, Ruben; Saedi, Chakaveh	Comparative Probing of Lexical Semantics Theories for Cognitive Plausibility and Technological Usefulness
7	Cai, J.; Zhu, Z.; Nie, P.; Liu, Q.	A Pairwise Probe for Understanding BERT Fine-Tuning on Machine Reading Comprehension
8	Chen, Weijie; Chang, Yongzhu; Zhang, Rongsheng; Pu, Jiashu; Chen, Guandan; Le Zhang; Xi, Yadong; Chen, Yijiang; Su, Chang	Probing Simile Knowledge from Pre-trained Language Models
9	Conia, Simone; Navigli, Roberto	Probing for Predicate Argument Structures in Pretrained Language Models
10	Dai, Yuqian; Kamps, Marc de; Sharoff, Serge	BERTology for Machine Translation: What BERT Knows about Linguistic Difficulties for Translation
11	Garcia, M.; Vieira, T. K.; Scarton, C.; Idiart, M.; Villavicencio, A.	Probing for idiomaticity in vector space models

12	Geiger, Atticus; Richardson, Kyle; Potts, Christopher	Neural Natural Language Inference Models Partially Embed Theories of Lexical Entailment and Negation
13	Hartmann, Mareike; Lhoneux, Miryam de; Hershcovich, Daniel; Kementchedjhieva, Yova; Nielsen, Lukas; Qiu, Chen; Søgaard, Anders	A Multilingual Benchmark for Probing Negation-Awareness with Minimal Pairs
14	He, Qianyu; Cheng, Sijie; Li, Zhixu; Xie, Rui; Xiao, Yanghua	Can Pre-trained Language Models Interpret Similes as Smart as Human?
15	Hennigen, L. T.; Williams, A.; Cotterell, R.	Intrinsic Probing through Dimension Selection
16	Hernandez, Evan; Andreas, Jacob	The Low-Dimensional Linear Geometry of Contextualized Word Representations
17	Hewitt, J.; Manning, C. D.	A structural probe for finding syntax in word representations
18	Hou, Yifan; Sachan, Mrinmaya	Bird's Eye: Probing for Linguistic Graph Structures with a Simple Information-Theoretic Approach
19	Kalyaeva, E.; Durandin, O.; Malafeev, A.	Behavior of Modern Pre-trained Language Models Using the Example of Probing Tasks
20	Kaushal, Ayush; Mahowald, Kyle	What do tokens know about their characters and how do they know it?
21	Kim, N.; Patel, R.; Poliak, A.; Wang, A.; Xia, P.; McCoy, R. T.; Tenney, I.; Ross, A.; Linzen, T.; van Durme, B.; Bowman, S. R.; Pavlick, E.	Probing what different NLP tasks teach machines about function word comprehension
22	Koto, F.; Lau, J. H.; Baldwin, T.	Discourse Probing of Pretrained Language Models
23	Kumar, Ayush; Narayanan Sundararaman, Mukuntha; Vepa, Jithendra	What BERT Based Language Model Learns in Spoken Transcripts: An Empirical Study
24	Kunz, Jenny; Kuhlmann, Marco	Test Harder than You Train: Probing with Extrapolation Splits
25	Kuznetsov, I.; Gurevych, I.	A matter of framing: The impact of linguistic formalism on probing results
26	Li, Bai; Zhu, Zining; Thomas, Guillaume; Rudzicz, Frank; Xu, Yang	Neural reality of argument structure constructions
27	Li, Bingzhi; Wisniewski, Guillaume; Crabb´e, Benoit	How Distributed are Distributed Representations? An Observation on the Locality of Syntactic Information in Verb Agreement Tasks
28	Limisiewicz, Tomasz; Mareˇcek, David	Examining Cross-lingual Contextual Embeddings with Orthogonal Structural Probes
29	Limisiewicz, Tomasz; Mareˇcek, David	Introducing Orthogonal Constraint in Structural Probes
30	Lin, Ruixi; Ng, Hwee Tou	Does BERT Know that the IS-A Relation Is Transitive?
31	Maudslay, Rowan Hall; Cotterell, Ryan	Do Syntactic Probes Probe Syntax? Experiments with Jabberwocky Probing
32	Michael, Julian; Botha, Jan A.; Tenney, Ian	Asking without Telling: Exploring Latent Ontologies in Contextual Representations
33	Mikhailov, V.; Serikov, O.; Artemova, E.	Morph Call: Probing Morphosyntactic Content of Multilingual Transformers
34	Milewski, Victor; Lhoneux, Miryam	Finding Structural Knowledge in

	de; Moens, Marie-Francine	Multimodal-BERT
35	Mohebbi, H.; Modarressi, A.; Pilehvar, M. T.	Exploring the Role of BERT Token Representations to Explain Sentence Probing Results
36	Mu¨ller-Eberstein, Max; van der Goot, Rob; Plank, Barbara	Probing for Labeled Dependency Trees
37	Patil, Rajaswa; Dhillon, Jasleen; Mahurkar, Siddhant; Kulkarni, Saumitra; Malhotra, Manav; Baths, Veeky	VyA Colorless Green Benchmark for Syntactic Evaluation in Indic Languages
38	Puccetti, Giovanni; Miaschi, Alessio; Dell'Orletta, Felice	How Do BERT Embeddings Organize Linguistic Knowledge?
39	Rama, Taraka; Beinborn, Lisa; Eger, Steffen	Probing Multilingual BERT for Genetic and Typological Signals
40	Ravichander, Abhilasha; Belinkov, Yonatan; Hovy, Eduard	Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance?
41	Ravichander, Abhilasha; Hovy, Eduard; Suleman, Kaheer; Trischler, Adam; Cheung, Jackie Chi Kit	On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT
42	Richardson, K.; Hu, H.; Moss, L. S.; Sabharwal, A.	Probing natural language inference models through semantic fragments
43	Richardson, K.; Sabharwal, A.	What does my qa model know? Devising controlled probes using expert knowledge
44	S¸ahin, G¨ozde Gu¨l; Vania, Clara; Kuznetsov, Ilia; Gurevych, Iryna	LINSPECTOR: Multilingual Probing Tasks for Word Representations
45	Senel, Lutfi Kerem; Schu¨tze, Hinrich	Does She Wink or Does She Nod? A Challenging Benchmark for Evaluating Word Understanding of Language Models
46	Shapiro, Naomi; Paullada, Amandalynne; Steinert-Threlkeld, Shane	A multilabel approach to morphosyntactic probing
47	Sinha, Koustuv; Jia, Robin; Hupkes, Dieuwke; Pineau, Joelle; Williams, Adina; Kiela, Douwe	Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little
48	Stanczak, Karolina; Ponti, Edoardo; Torroba Hennigen, Lucas; Cotterell, Ryan; Augenstein, Isabelle	Same Neurons, Different Languages: Probing Morphosyntax in Multilingual Pre-trained Models
49	Tan, Minghuan; undefined; Jiang, Jing	Does BERT Understand Idioms? A Probing-Based Empirical Study of BERT Encodings of Idioms
50	Tayyar Madabushi, Harish; Romain, Laurence; Divjak, Dagmar; Milin, Petar	CxGBERT: BERT meets Construction Grammar
51	Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; Thomas McCoy, R.; Kim, N.; van Durme, B.; Bowman, S. R.; Das, D.; Pavlick, E.	What do you learn from context? Probing for sentence structure in contextualized word representations
52	Tikhonova, Maria; Mikhailov, Vladislav; Pisarevskaya, Dina; Malykh, Valentin; Shavrina, Tatiana	Ad astra or astray: Exploring linguistic knowledge of multilingual BERT through NLI task
53	Vries, Wietse de; van Cranenburgh, Andreas; Nissim, Malvina	What's so special about BERT's layers? A closer look at the NLP pipeline in monolingual and multilingual models
54	Vuli´c, I.; Ponti, E. M.; Litschko,	Probing pretrained language models for

	R.; Glavaˇs, G.; Korhonen, A.	lexical semantics
55	Wang, Yile; Cui, Leyang; Zhang, Yue	Does Chinese BERT Encode Word Structure?
56	Wu, Zhaofeng; Peng, Hao; Smith, Noah A.	Infusing Finetuning with Semantic Dependencies
57	Zhao, M.; Dufter, P.; Yaghoobzadeh, Y.; Schu ̈tze, H.	Quantifying the contextualization of word representations with semantic class probing

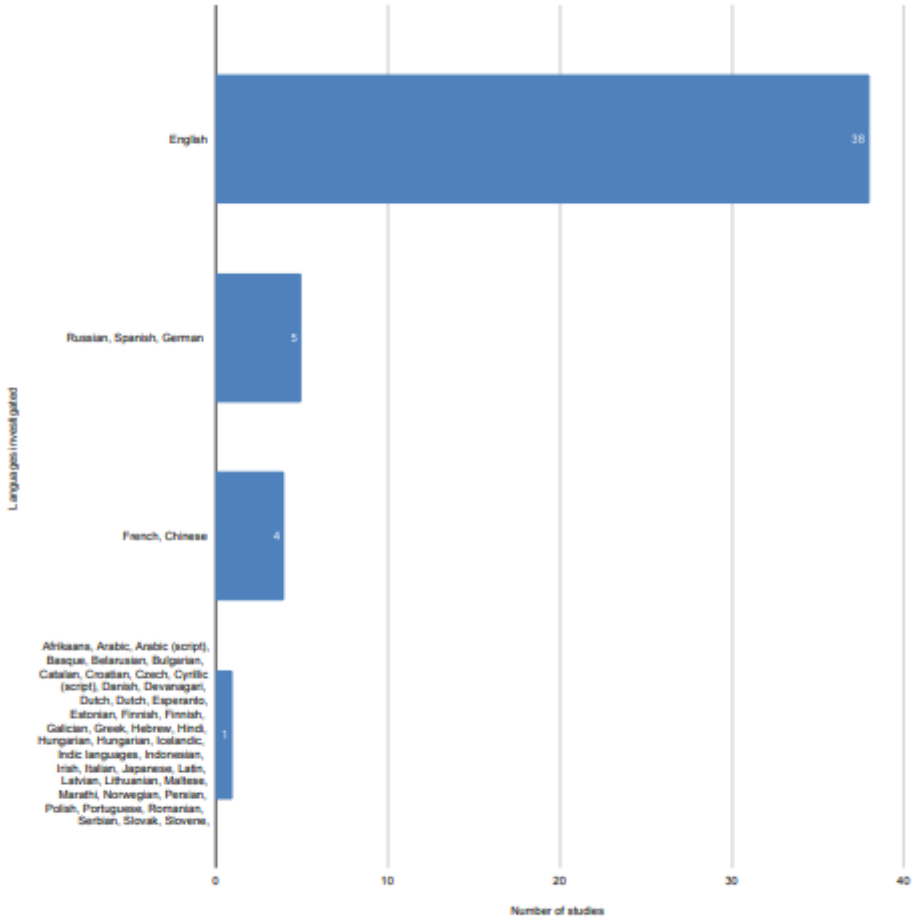

Fig. 4. Languages investigated.

© 2023 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.