

Towards a linguistic-based inductive bias for neural language modeling

Marcos Garcia¹, Iria de-Dios-Flores², Helena Pérez Puente¹ and David Gallego^{1,3}

¹*Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela*

²*Departament de Traducció i Ciències del Llenguatge, Universitat Pompeu Fabra*

³*Instituto de Psicología (IPsiUS), Universidade de Santiago de Compostela*

Abstract

This paper introduces the Project LINNA: “Towards a linguistic-based inductive bias for neural language modeling”, carried out at the CiTIUS Research Center of the Universidade de Santiago de Compostela, and funded by a ‘Consolidación Investigadora’ grant. The project aims to foster research towards more human-like language models by exploring new architectures and training methods and by using developmentally plausible data for training and evaluation. In this context, we plan to provide the scientific community with a new suite of resources and models, including diverse multilingual datasets and model weights, as well as empirical results regarding data-efficient language modeling.

Keywords

language modeling, language acquisition, inductive bias, small language models

1. Introduction

Current language models (LMs) based on artificial neural networks are able to generate high-quality text [1], and perform remarkably well in a variety of language processing tasks. The fact that these models are trained on simple objectives, such as predicting words in context, and that the generated text is not a simple copy of the training data [2], has revived the classic nature-nurture debate [3, 4]. On the one hand, some cognitive scientists argue that “modern language models implement genuine theories of language” [5, 6], while on the other hand, an opposing position claims that LMs are purely computational tools of no value for the scientific study of human language [7, 8, 9]. In an intermediate position, several researchers from different fields argue that while these computational tools are not plausible models of human language, they can be useful for cautiously informing models of how human language works [10, 11, 12, 13], since, among other things, they allow exploring the extent to which linguistic regularities can be generalized from data without explicit linguistic knowledge [14].

What is clear is that current LMs based on neural networks are significantly less efficient than the biological mechanism inherent in humans, which requires substantially less data to acquire natural languages [15, 9]. In this regard, humans possess an inductive bias that enables the efficient acquisition of linguistic regularities, while making it difficult to learn structures that do not adhere to natural language principles. Recent studies suggest that, at least in the USA, children are exposed to at most 8 million words per year, reaching roughly 100 million words by age 12 in the best-case scenario [16]. In contrast, the original English BERT model [17] was trained on approximately 3.3 billion words (4.3 billion tokens), while current LLMs like GPT-5 are estimated to be trained on about 114 trillion tokens.¹ In this light, the poverty of the stimulus argument [4] seems to remain highly relevant [18, 19].

Furthermore, in addition to the large amount of data required to computationally model human language, training current LMs involves tuning several billion parameters, resulting in extremely

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

✉ marcos.garcia.gonzalez@usc.gal (M. Garcia); iria.dedios@upf.edu (I. de-Dios-Flores); helenaperez@usc.gal (H. Pérez Puente); davidgallego.conde@usc.es (D. Gallego)

🆔 0000-0002-6557-0210 (M. Garcia); 0000-0002-5941-1707 (I. de-Dios-Flores); 0009-0005-6345-045X (H. Pérez Puente); 0000-0001-8951-1234 (D. Gallego)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://lifearchitect.ai/gpt-5/>

complex architectures and computationally expensive models. In this context, recent research has shown that a significant part of a pre-trained language model (about 70% of the attention heads and 20% of the hidden layers for in-context learning) can be removed with minimal impact on its performance [20]. It has also been shown that pre-training a model using psycholinguistic strategies, e.g., cloze tasks, can improve its alignment with reading time prediction and generalization [21].

Given these observations, this project aims to explore different methods to create a Language-Independent Neural Network Architecture (LINNA) by pruning and distilling a linguistically diverse multilingual model. The resulting weights of this language-independent architecture are used as a proxy for a linguistic-based inductive bias to train monolingual LMs using small-scale training data and strategies motivated by child language acquisition [22].

The hypothesis behind this approach is that a linguistically diverse multilingual model, pre-trained with a balanced set of languages from a variety of linguistic families and structures, can effectively encode some universal linguistic properties in its internal representations. By pruning and distilling the model using linguistic-based tasks, we aim at reducing the information overlap that exists across the network and to build a language-independent architecture whose weights can be seen as a proxy for a linguistic-based inductive bias.² The project is partially inspired by the BabyLM Challenge [23, 24].³

2. Objectives

The main objective of the project is to explore different methods aimed to create a language-independent neural network architecture whose weights could function as a proxy for a linguistic-based inductive bias. To achieve this goal, we have defined the following specific objectives:

- Curate a multilingual corpus by selecting a set of languages from different families and with divergent linguistic properties.
- Explore methods to train, prune and distill neural language models with a linguistically diverse multilingual corpus and linguistic-based datasets, and remove their language-specific vocabulary to create (partially) language-independent (LINNA) models.
- Select a set of languages to train the LINNA models, including both languages present in the initial pre-training and other languages from different families.
- Define a set of linguistic phenomena (semantic and syntactic) to evaluate the models, and compile and create datasets for the target languages.
- Design controlled evaluations to assess the models' capabilities and compare their performance to that of other models trained from scratch.

While these specific objectives are designed to achieve our main goal, they will also produce a set of resources, models, and empirical results of interest to the scientific community.

3. Methodology

To achieve both the specific objectives and the main objective of the project, we propose to adopt an interdisciplinary set of methodologies that come from both theoretical linguistics (e.g., linguistic universals and variation) and psycholinguistics (cloze tasks, acceptability judgments, etc.), as well as recent advances in AI and deep learning (Transformers, distillation and pruning, etc.). In addition to the management tasks (including research meetings, data management, dissemination activities, etc.), the project is divided into the following 6 work packages (WPs), each of them composed by several tasks:

²We use *inductive bias* to refer to the assumptions a model relies on to generalize to unseen data. Thus, we assume that the resulting language-agnostic model may constitute a form of linguistic-based inductive bias, as it derives from representations learned across typologically diverse languages during pre-training.

³<https://babylm.github.io/index.html>

1. **Compilation of a linguistically diverse corpus:** Curating a set of linguistically diverse languages from different families and with sufficient of training corpus available.
Summary of tasks: In this work package, we will analyze a set of structural properties of languages (such as SVO vs. SOV, agglutinative vs. fusional, and ergative-absolutive vs. nominative-accusative) using *The World Atlas of Language Structures* (WALS).⁴ Our purpose is to explore the availability and licenses of resources for a preliminary set of selected languages. Then, we will compile a balanced and linguistically diverse multilingual corpus, while keeping some target languages for monolingual model training (WP 4) out of the main corpus. For instance, we might choose Georgian as an example of an ergative language for pre-training (WP 2), while using Basque (also ergative) only to train and evaluate the monolingual models (WPs 5-7).
2. **Training encoder and decoder multilingual models:** Exploring neural architectures (e.g., LSTMs, Transformers) to train encoder and decoder models of different sizes using the multilingual corpus.
Summary of tasks: This work package focuses on processing and dividing the corpus into training, development, and test splits, specifically taking into account its multilingual configuration. Building on this setup, we will train and evaluate encoder and decoder models. Moreover, we will measure performance using metrics like perplexity and loss function.
3. **Create language-independent models (LINNAs):** Applying pruning and distillation techniques on the multilingual models to obtain the linguistic-based weights.
Summary of tasks: We will begin by compiling available datasets for multilingual evaluation of the linguistic abilities of language models, focusing on lexical semantic tasks [25, 26, 27], syntactic analysis [28, 29], and agreement using control dependencies [30]. Next, we aim at using the collected datasets to prune the multilingual models via structural pruning approaches [31, 20]. Furthermore, we will compare distillation techniques (in English), such as the standard approach used for Transformer models [32] and psycholinguistic-based distillation using a cloze prediction tasks [21], e.g., with the Provo Corpus [33]. After selecting the best performing method, we will apply it to other languages and remove the models' vocabulary, resulting in smaller models.
4. **Target language selection:** Selecting a set of languages to train the LINNA models, including both the languages of the initial pre-training (WPs 1 and 2) and other languages from different families, e.g. Basque (ergative), Galician (SVO), or Japanese (SOV), considering the availability of linguistic resources especially for languages not spoken by the research team.
Summary of tasks: WP 4 will start by using the outcomes of WP 1 and of recently released resources [34] to select languages with available data for both training and evaluating language models. We will select some languages present in the initial pre-training, and others not seen by the pre-trained models. Then, we will compile training data for the target languages and divide them into splits of different sizes, while also exploring the possibilities of building code-mixing and code-switching data for Galician-Spanish. After that, we will define a set of syntactic and semantic phenomena to evaluate the models, specifically looking at tasks related to lexical semantics (ambiguity, compositionality and idiomaticity, etc.) and to syntax (e.g., subject-verb agreement, inflected infinitive) including agreement with control verbs. Finally, we will compile available datasets for the target languages, including JBLiMP for Japanese [35], the linguistic evaluation suite for Basque [36], MultiBLiMP [29], and available resources for Galician [26, 37, 38, 30, 39, 40].
5. **Training monolingual models from the LINNA models:** Using the training corpora released in WP 4 and the various LINNA models from WP 3, train monolingual models with different amounts of training data and with strategies motivated by child language acquisition.
Summary of tasks: In this phase, we analyze the training corpora and explore curriculum learning strategies, ranking sentences according to their syntactic and lexical complexity (e.g., based on average dependency length and on normalized word frequency). Using this rank, we aim to define complexity levels for each language. We will also train and evaluate tokenizers for each of the target languages. Moreover, we will conduct continual pre-training of the original models

⁴<https://wals.info/>

(WP 2) with those target languages not seen in the initial pre-training. Then, we will perform continual pre-training of the LINNA models using standard encoder and decoder approaches, varying the number of training epochs and the levels of corpus complexity. Finally, we explore the possibility of training small bilingual Galician-Spanish models.⁵

6. **Linguistic evaluation of monolingual models:** Designing controlled evaluations and systematic analyses to assess each model’s capabilities and compare their performance to models trained from scratch across all target languages.

Summary of tasks: In the last work package, we will define metrics and evaluation procedures for each linguistic phenomenon, and run the evaluation of all models. Apart from training metrics (e.g., perplexity), we will evaluate syntactic and semantic capabilities using standard methods (e.g., Targeted Syntactic Evaluation [28], Word-in-Context for lexical semantics [25], or vector-based compositionality prediction [41]). We also intend to incorporate evaluation frameworks like PD accuracy for syntax [42], and substitutivity probes for semantic compositionality [27]. For each language, we will carry out systematic analysis of the impact of the (i) amount of training data, (ii) training procedures, and (iii) pre-trained model. Additionally, we would like to perform an initial evaluation of small bilingual Galician-Spanish models, aiming at observing if they produce human-like bilingual behavior. For example, we want to see if they produce similar patterns of code-mixing and code-switching as those observed in Galician-Spanish human speakers.

4. Expected outcomes

We expect that our project will contribute both new theoretical outcomes and a diverse collection of resources and models for the community. To this end, we defined specific milestones for each of the previously mentioned work packages:

- WP1: Release of a linguistically diverse multilingual corpus.
- WP2: Release of a variety of linguistically diverse multilingual models.
- WP3: Release of (i) datasets for evaluating and pruning language models based on linguistic phenomena, and (ii) several LINNA models.
- WP4: Release of (i) training corpora with different sizes, and (ii) evaluation resources for the target languages.
- WP5: Release, for each of the target languages (i) tokenizers, (ii) standard monolingual models, and (iii) models trained using linguistically motivated methods.
- WP6: Release of (i) new metrics for the linguistic evaluation of language models, and (ii) results of the systematic analyses performed.

Crucially, a key goal is to fill a gap in current resources by creating developmentally plausible datasets for Galician, which is missing from major datasets in this area [34].

5. Further work

As this project progresses, we are identifying several lines for future work, one of them related to bilingual scenarios already under initial development. In addition to exploring new methodologies, we see two clear paths forward. First, we aim to evaluate models’ performance in bilingual scenarios, specifically by addressing Galician and Spanish code-switching and code-mixing. By comparing model outputs to empirical data, we seek to better understand how models deal with structural patterns when working with more than one language, and whether their behavior is similar to that of bilingual speakers. Second, we plan to explore multimodal models that incorporate acoustic and visual inputs, simulating the human acquisition process more realistically.

⁵Given the focus on moderate-scale models and datasets, the project is computationally feasible using the resources available at our research centre.

Acknowledgments

This research is funded by MICIU/AEI/10.13039/501100011033 (grant reference CNS2024-154902).

Declaration on Generative AI

The authors used open-source AI tools to assist with English language polishing and error identification. The final text was thoroughly reviewed, edited, and approved by the authors.

References

- [1] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, N. A. Smith, All that's 'human' is not gold: Evaluating human evaluation of generated text, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 7282–7296. URL: <https://aclanthology.org/2021.acl-long.565/>. doi:10.18653/v1/2021.acl-long.565.
- [2] R. T. McCoy, P. Smolensky, T. Linzen, J. Gao, A. Celikyilmaz, How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN, Transactions of the Association for Computational Linguistics 11 (2023) 652–670. URL: <https://aclanthology.org/2023.tacl-1.38/>. doi:10.1162/tacl_a_00567.
- [3] J. N. Wood, Artificial intelligence tackles the nature–nurture debate, Nature Machine Intelligence 6 (2024) 381–382. doi:<https://doi.org/10.1038/s42256-024-00828-4>.
- [4] N. Chomsky, Knowledge of Language: Its Nature, Origin, and Use, Praeger, New York, 1986.
- [5] S. T. Piantadosi, Modern language models refute Chomsky's approach to language, From fieldwork to linguistic theory: A tribute to Dan Everett 15 (2023) 353–414. URL: https://pemt.ru/wp-content/uploads/2023/06/piantadosi_modern-lang.pdf.
- [6] F. Hill, Why transformers are obviously good models of language (2024). arXiv:2408.03855.
- [7] E. M. Bender, A. Koller, Climbing towards NLU: On meaning, form, and understanding in the age of data, in: D. Jurafsky, J. Chai, N. Schuster, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5185–5198. URL: <https://aclanthology.org/2020.acl-main.463/>. doi:10.18653/v1/2020.acl-main.463.
- [8] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), Association for Computing Machinery, New York, 2021, pp. 610–623. doi:10.1145/3442188.3445922.
- [9] J. Kodner, S. Payne, J. Heinz, Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi (2023) (2023). arXiv:2308.03228.
- [10] T. Linzen, M. Baroni, Syntactic structure from deep learning, Annual Review of Linguistics 7 (2021) 195–212. URL: <https://www.annualreviews.org/content/journals/10.1146/annurev-linguistics-032020-051035>.
- [11] A. Warstadt, S. R. Bowman, What artificial neural networks can tell us about human language acquisition, in: Algebraic Structures in Natural Language, CRC Press, 2022, pp. 17–60. arXiv:2208.07998.
- [12] M. C. Frank, Baby steps in evaluating the capacities of large language models, Nature Reviews Psychology 2 (2023) 451–452. doi:10.1038/s44159-023-00211-x.
- [13] R. Futrell, K. Mahowald, How Linguistics Learned to Stop Worrying and Love the Language Models, Behavioral and Brain Sciences (2025) 1–98. doi:10.1017/s0140525x2510112x.
- [14] E. Wilcox, M. Y. Hu, A. Mueller, A. Warstadt, L. Choshen, C. Zhuang, A. Williams, R. Cotterell, T. Linzen, Bigger is not always better: The importance of human-scale language modeling for

- psycholinguistics, *Journal of Memory and Language* 144 (2025). doi:<https://doi.org/10.1016/j.jml.2025.104650>.
- [15] T. Linzen, How can we accelerate progress towards human-like linguistic generalization?, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 5210–5217. URL: <https://aclanthology.org/2020.acl-main.465/>. doi:10.18653/v1/2020.acl-main.465.
 - [16] J. Gilkerson, J. A. Richards, S. F. Warren, J. K. Montgomery, C. R. Greenwood, D. Kimbrough Oller, J. H. Hansen, T. D. Paul, Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis, *American Journal of Speech-Language Pathology* 26 (2017) 248–265. doi:10.1044/2016_AJSLP-15-0169.
 - [17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423/>. doi:10.18653/v1/N19-1423.
 - [18] A. Yedetore, T. Linzen, R. Frank, R. T. McCoy, How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 9370–9393. URL: <https://aclanthology.org/2023.acl-long.521/>. doi:10.18653/v1/2023.acl-long.521.
 - [19] N. Lan, E. Chemla, R. Katzir, Large language models and the argument from the poverty of the stimulus, *Linguistic Inquiry* (2024) 1–28. doi:10.1162/ling_a_00533.
 - [20] H. Bansal, K. Gopalakrishnan, S. Dingliwal, S. Bodapati, K. Kirchhoff, D. Roth, Rethinking the role of scale for in-context learning: An interpretability-based case study at 66 billion scale, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 11833–11856. URL: <https://aclanthology.org/2023.acl-long.660/>. doi:10.18653/v1/2023.acl-long.660.
 - [21] T. Eisape, N. Zaslavsky, R. Levy, Cloze distillation: Improving neural language models with human next-word prediction, in: R. Fernández, T. Linzen (Eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Online, 2020, pp. 609–619. URL: <https://aclanthology.org/2020.conll-1.49/>. doi:10.18653/v1/2020.conll-1.49.
 - [22] E. Portelance, M. Jasbi, The roles of neural networks in language acquisition, *Language and Linguistics Compass* 18 (2024) e70001. doi:10.1111/lnc3.70001.
 - [23] A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, R. Cotterell (Eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Singapore, 2023. URL: <https://aclanthology.org/2023.conll-babylm.0/>.
 - [24] L. Charpentier, L. Choshen, R. Cotterell, M. O. Gul, M. Y. Hu, J. Liu, J. Jumelet, T. Linzen, A. Mueller, C. Ross, R. S. Shah, A. Warstadt, E. G. Wilcox, A. Williams (Eds.), *Proceedings of the First BabyLM Workshop*, Association for Computational Linguistics, Suzhou, China, 2025. URL: <https://aclanthology.org/2025.babylm-main.0/>. doi:10.18653/v1/2025.babylm-main.0.
 - [25] A. Raganato, T. Pasini, J. Camacho-Collados, M. T. Pilehvar, XL-WiC: A multilingual benchmark for evaluating semantic contextualization, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, 2020, pp. 7193–7206. URL: <https://aclanthology.org/2020.emnlp-main.584/>. doi:10.18653/v1/2020.emnlp-main.584.
 - [26] M. Garcia, T. Kramer Vieira, C. Scarton, M. Idiart, A. Villavicencio, Assessing the representations

- of idiomaticity in vector models with a noun compound dataset labeled at type and token levels, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Online, 2021, pp. 2730–2741. URL: <https://aclanthology.org/2021.acl-long.212/>. doi:10.18653/v1/2021.acl-long.212.
- [27] M. Garcia, T. Kramer Vieira, C. Scarton, M. Idiart, A. Villavicencio, Probing for idiomaticity in vector space models, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Association for Computational Linguistics, Online, 2021, pp. 3551–3564. URL: <https://aclanthology.org/2021.eacl-main.310/>. doi:10.18653/v1/2021.eacl-main.310.
- [28] R. Marvin, T. Linzen, Targeted syntactic evaluation of language models, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1192–1202. URL: <https://aclanthology.org/D18-1151/>. doi:10.18653/v1/D18-1151.
- [29] J. Jumelet, L. Weissweiler, A. Bisazza, MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs (2025). *arXiv:2504.02768*.
- [30] I. de Dios-Flores, J. Garcia Amboage, M. Garcia, Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 203–222. URL: <https://aclanthology.org/2023.acl-long.12/>. doi:10.18653/v1/2023.acl-long.12.
- [31] S. Anwar, K. Hwang, W. Sung, Structured pruning of deep convolutional neural networks, *ACM Journal on Emerging Technologies in Computing Systems (JETC)* 13 (2017) 1–18. doi:10.1145/3005348.
- [32] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, in: *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*, 2019. *arXiv:1910.01108*.
- [33] S. G. Luke, K. Christianson, The Provo Corpus: A large eye-tracking corpus with predictability norms, *Behavior research methods* 50 (2018) 826–833. doi:10.3758/s13428-017-0908-4.
- [34] J. Jumelet, A. Fourtassi, A. Haga, B. Bunzeck, B. Shandilya, D. Galvan-Sosa, F. G. Haznitrana, F. Padovani, F. Meyer, H. Hu, J. Etxaniz, L. Prevot, L. He, M. Grandury, M. Marcheva, N. Foroutan, N. Theodoropoulos, P. Sadeghi, S. Song, S. Salhan, S. Zhou, Y. Paniv, Z. Zhang, A. Bisazza, A. Warstadt, L. Choshen, BabyBabelLM: A multilingual benchmark of developmentally plausible training data, in: V. Demberg, K. Inui, L. Marquez (Eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Rabat, Morocco, 2026, pp. 3297–3329. URL: <https://aclanthology.org/2026.eacl-long.152/>. doi:10.18653/v1/2026.eacl-long.152.
- [35] T. Someya, Y. Oseki, JBLiMP: Japanese benchmark of linguistic minimal pairs, in: *Findings of the Association for Computational Linguistics: EACL 2023*, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1581–1594. URL: <https://aclanthology.org/2023.findings-eacl.117/>.
- [36] J. Etxaniz, O. Sainz, N. Perez, I. Aldabe, G. Rigau, E. Agirre, A. Ormazabal, M. Artetxe, A. Soroa, Latxa: An Open Language Model and Evaluation Suite for Basque, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 14952–14972. URL: <https://aclanthology.org/2024.acl-long.799/>.
- [37] M. Garcia, A. Crespo-Otero, A Targeted Assessment of the Syntactic Abilities of Transformer Models for Galician-Portuguese, in: *Computational Processing of the Portuguese Language. PROPOR 2022. Lecture Notes in Computer Science*, Springer, 2022, pp. 46–56. doi:https://doi.org/10.1007/978-3-030-98305-5_5.
- [38] H. Pérez Puente, O procesamento da estrutura sintáctica nos modelos de linguaxe neurais: a

- concordancia suxeito-verbo en galego e portugués, *Linguamática* 17 (2025) 119–128. URL: <https://linguamatica.com/index.php/linguamatica/article/view/498>. doi:10.21814/lm.17.2.498.
- [39] L. Castro, A. Temerko, M. Garcia, Compositionality and ambiguity in multiword expressions: A dataset for the evaluation of language models in galician, in: *Progress in Artificial Intelligence. EPIA 2024. Lecture Notes in Computer Science*, volume 14969, Springer, 2024, pp. 228–240. doi:https://doi.org/10.1007/978-3-031-73503-5_19.
- [40] M. V. Abuín, M. Garcia, WiC evaluation in Galician and Spanish: Effects of dataset quality and composition, in: L. Frermann, M. Stevenson (Eds.), *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, Association for Computational Linguistics, Suzhou, China, 2025, pp. 172–178. URL: <https://aclanthology.org/2025.starsem-1.13/>. doi:10.18653/v1/2025.starsem-1.13.
- [41] F. Miletic, S. Schulte im Walde, A systematic search for compound semantics in pretrained BERT architectures, in: A. Vlachos, I. Augenstein (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1499–1512. URL: <https://aclanthology.org/2023.eacl-main.110/>. doi:10.18653/v1/2023.eacl-main.110.
- [42] I. de Dios-Flores, M. Garcia, A computational psycholinguistic evaluation of the syntactic abilities of galician bert models at the interface of dependency resolution and training time, *Procesamiento del Lenguaje Natural* 69 (2022) 15–26. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6424>.