Monolingual and Bilingual Language Acquisition in Language Models

Mihir Sharma, Ryan Ding, Raj Sanjay Shah, Sashank Varma
Georgia Institute of Technology
{msharma95, rding62, rajsanjayshah, varma}@gatech.edu

Abstract

An emerging research question in human language understanding is how well computational language models (LMs) align with child language acquisition. The starting point for our work is the study of bilingual language acquisition by Hoff et al. (2012). That study found that monolingual children consistently outperformed bilingual children on single-language understanding tasks, with a consistent and measurable lag of less than three months. However, that study also found that bilingual children have a comparable overall vocabulary size to monolingual children. Here, we investigate these questions computationally. We train monolingual and bilingual LMs using the BabyBERTa architecture (Huebner et al., 2021). We use English, German, and Spanish data collected from the CHILDES dataset, with each model trained for 10 epochs on approximately 2.8M tokens in each epoch. We document the "development" of these models over training and look for whether bilingual models show the same lag and overall vocabulary size effects as children. We hypothesize that we will see a similar lag in LMs as in human children. In this case, we argue that LMs are an effective tool with which to computationally model human language acquisition and that they serve as a suitable basis for future research aligning LM performance with human performance.

1 Background

We computationally model monolingual and bilingual learning to identify whether LMs, like humans, demonstrate a "lag" in language understanding. Hoff et al. (2012) evaluated 56 children learning only English (monolingual development) and 47 children learning English and Spanish simultaneously (bilingual development). The data were collected at different stages of development. To computationally model the same "lag" during development, we pre-train several LMs on different monolingual and bilingual language combinations with similar vocabulary sizes. We then establish a temporal measurement of growth by measuring the accuracy of each model for intermediate checkpoints during the training process. Through these checkpoint evaluations, we will be able to see if there is an alignment between the number of model training steps and human growth.

2 Bilingual Dataset

We pre-train our language models using conversational transcripts of children. This cross-lingual pre-training data is representative of child-level grammar, extensive in its overall vocabulary, and similar in content across languages. We draw from CHILDES, a dataset aggregating an extensive collection of conversational transcripts with children in over 20 different languages (Macwhinney, 2000). CHILDES contains child-directed speech and child-level grammar, encompasses a variety of children's ages, and contains similar content across languages, making it a realistic simulation for studying human language acquisition effects. Additionally, the CHILDES repository has already been used to simulate human-like language understanding when training transformer-based models (Huebner et al., 2021). We first create various training corpora across English, Spanish, and German. We selected these three languages due to their greater representation in the dataset, with each language having higher total token counts compared to several of the other languages available. The training data are taken directly from CHILDES transcripts and processed for pre-training.

3 Monolingual and Bilingual Simulation

For each pair of languages L1 and L2, we train the monolingual and bilingual model combinations defined in Table 2. The monolingual training is self-explanatory. For the bilingual training, we use three different data splits. For the *L1-L2 shuffled* training, the two languages are randomly shuffled

Task	Description
Named Entity Recognition (NER)	Identify named entities (people, places, etc.). (Sang, 2002) (Sang and Meulder, 2003)
Part of Speech Tagging (POS)	Assign parts of speech to tokens. (Zeman et al., 2019)
News Classification (NC)	Categorize news articles given their headline and content.
Cross-lingual Natural Language Inference (XNLI)	Determine whether one sentence entails another. (Conneau et al., 2018)
Paraphrase Adversaries from Word Scrambling (PAWS-X)	Determine if one sentence paraphrases another. (Yang et al., 2019)
Query-Ad Matching (QASDM)	Determine whether an ad is relevant to a query.
Web Page Ranking (WPR)	Determine whether a web page is relevant to a query.
Question-Answer Matching (QAM)	Predict whether a question and answer are a pair.

Table 1: Evaluation benchmark: XGLUE tasks used to test the model checkpoints, as well as a short description of each task.

Simulation type	Language
Monolingual	L1
Monolingual	L2
Bilingual	L1–L2 (shuffled)
Bilingual	L1, L2 (sequential)
Bilingual	L2, L1 (sequential)

Table 2: Different simulation types and the respective pre-training languages. L1 and L2 refer to the two languages under consideration.

within the training and evaluation datasets. For *L1, L2 sequential* training, the two languages are evenly split but sequenced in two blocks, i.e. with all L1 data preceding all L2 data in the training and evaluation datasets; for *L2, L1 sequential*, the ordering of the blocks is reversed. Each dataset's training and evaluation splits are ≈2.8 million tokens and ≈120 thousand tokens in size, respectively.

4 Model Architecture and Training

To pre-train the models, we perform a Masked Language Modeling task on untrained instances of BabyBERTa. This is a newer model intended as a suitable testbed for the alignment to human language acquisition (Huebner et al., 2021). It is a scaled-down version of RoBERTa, possessing fewer layers and trained on a much smaller vocabulary size, modifications which significantly reduce the training time and compute power. Pre-training is performed using the *simpletransformers.ai* MLM framework (Rajapakse). All models are initialized with random weights and trained for 10 epochs on datasets of ≈2.8M tokens in each epoch, following a similar procedure as the BabyBERTa models. During pre-training, the model's weights are checkpointed at every 100 training steps, which are then tested on language understanding benchmarks, to understand how the model improves over time.

5 Model Evaluation

We use XGLUE, a benchmark tailored for the cross-language pre-trained models, to evaluate performance (Liang et al., 2020). We use a subset of the tasks as seen in Table 1. We evaluate the performance of each available intermediate checkpoint to observe developmental trends. Each checkpoint's benchmark results are used as a proxy for language understanding and acquisition. This enables us to draw parallels to age-based language understanding assessments.

6 Preliminary Results

We have conducted preliminary investigations of the full set of models. Currently, we see that model perplexity is lower for the monolingual models after pre-training than the bilingual models, supporting the "lag" prediction. Moreover, the sequential bilingual models appear to have lower perplexity than the shuffled models. We are currently in the process of testing the full set of models with XGLUE and will report the results in the poster.

7 Potential Implications

Our results hope to indicate that lag may be quantified as a time value during model pre-training, which could estimate chronological measurement during model language acquisition. Additionally, if the models do demonstrate a similar lag to children, this could open future work applying insights from children to creating multilingual models. Lastly, we intend to release the checkpointed models to the public as tools for future experiments on computational and human language acquisition.

References

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *EMNLP*.

Erika Hoff, Cynthia Core, Silvia Place, Rosario Rumiche, Melissa Señor, and Marisol Parra. 2012. Dual language exposure and early bilingual development. *Journal of Child Language*, 39(1):1–27. Publisher: Cambridge University Press.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. ArXiv:2004.01401 [cs].

Brian Macwhinney. 2000. The CHILDES project: tools for analyzing talk. *Child Language Teaching and Therapy*, 8.

Thilina Rajapakse. Simple Transformers.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. *ArXiv*, cs.CL/0209010.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *ArXiv*, cs.CL/0306050.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *ArXiv*, abs/1908.11828.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Maria Liovina, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka

Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.