BAMBINO-LM: (Bilingual-)Human-Inspired Continual Pre-training of BabyLM

Zhewen Shen[1] **Aditya Joshi** [1] **Ruey-Cheng Chen**[2]
[1] University of New South Wales, Sydney, Australia
[2] Canva, Sydney, Australia
zhewen.shen@student.unsw.edu.au, aditya.joshi@unsw.edu.au, rcchen@canva.com

Abstract

Children from bilingual backgrounds benefit from interactions with parents and teachers to re-acquire their heritage language. In this paper, we investigate how this insight from behavioral study can be incorporated into the learning of small-scale language models. We introduce BAMBINO-LM, a continual pre-training strategy for BabyLM that uses a novel combination of alternation and PPO-based perplexity reward induced from a parent Italian model. Upon evaluation on zero-shot classification tasks for English and Italian, BAMBINO-LM improves the Italian language capability of a BabyLM baseline. Our ablation analysis demonstrates that employing both the alternation strategy and PPO-based modeling is key to this effectiveness gain. We also show that, as a side effect, the proposed method leads to a similar degradation in L1 effectiveness as human children would have had in an equivalent learning scenario. Through its modeling and findings, BAMBINO-LM makes a focused contribution to the pre-training of small-scale language models by first developing a human-inspired strategy for pre-training and then showing that it results in behaviours similar to that of humans.

1 Introduction

The recently held **BabyLM** challenge (Warstadt et al., 2023) explores pretraining of language models using a constrained dataset analogous to the linguistic exposure of a 13-year-old English-speaking child. In this paper, we extend the BabyLM challenge to a bilingual setting, drawing inspiration from parent-child interactions in heritage language acquisition (Lohndal et al., 2019). Immigrant children in western societies, who may have acquired their home language at a young age, can sometimes need to re-acquire the same language during the school years when the language becomes a minority. These heritage speakers typically benefit from an extended exposure to the minority language at home or in the community, owing largely to feedback and stimuli provided by parents and family members (Montrul, 2010). This observation about child bilingualism is in line with the behaviorist theory for child language development (Demirezen, 1988).

Inspired by this line of work, we ask the following research question in the context of computational language modeling:

> *Can a small-scale language model trained on the majority language (e.g. English) be continually pre-trained on the minority language, leveraging the feedback of a second model that is fluent in the latter language?*

To address this question, we introduce '**Bilingual language Acquisition Modeling Based on INterleaved Optimization of Language Models (BAMBINO-LM)**', a novel continual pretraining strategy that uses a combination of alternation and proximal policy optimization (PPO) using a reward from a second model playing the parent role (i.e., a large language model pre-trained in the minority language). We experiment with BabyLM trained on *English*, and continually pretrain this model on an assumed second language, *Italian*.In its connection to cognitive processing, our work makes the following contributions:

- BAMBINO-LM draws inspiration from bilingual language acquisition and learns from interactions with a second model by **incorporating a perplexity-based reward for language model pre-training**. To the best of our knowledge, this is the first work to use PPO-based modeling for language acquisition in BabyLM.
- We show that BAMBINO-LM can acquire Italian to a reasonable degree with some expected

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 1–7

degradation in its English capability. The findings hint at **a common learning trajectories for second language acquisition shared by language models and humans**.

2 Related Work

Pre-training small-scale language models is an emerging field that has garnered some interest from the language acquisition community. BabyBERTa (Huebner et al., 2021) is an early adaptation to this scenario. Warstadt et al. (2023) introduce the BabyLM challenge to provide an atypically small dataset for benchmarking small-scale language models. This shared task enables research in not only language acquisition but also sample-efficient pre-training. In the case of our paper, we do not focus on sample efficiency but instead describe ways to enhance the ability of a second language via continual pre-training.

Our work is conducted in a setup similar to Yadavalli et al. (2023), where a tiered first/second language acquisition process is attempted. Samuel (2023) also experiments with a teacher-student setting but only tests the approach on English tasks. Evanson et al. (2023) is another closely related work, which investigates the learning trajectory of large-scale language models by probing their syntactic and semantic capabilities at each step.

Conventionally in language generation for natural language processing, the design of feedback signals is commonly discussed in the context of knowledge distillation (Calderon et al., 2023). Recently, reinforcement learning from human feedback (RLHF) utilizes human preferences for serving reward signals when dealing with sparse training labels (Christiano et al., 2017; Stiennon et al., 2020), and has been shown successful for generative tasks such as dialogues and summarization. This approach is further extended in Bai et al. (2022) by using AI feedback (RLAIF) to remove the dependency on human preference data, leading to better scalability and signal availability. The approach we take in this paper mostly falls within the latter camp, but generally departs from all prior efforts in the way the parent model's perplexity is used to signal the conformity of the child model's generation. This is in contrast to sequence-level knowledge distillation (Kim and Rush, 2016) where teacher's generation is used to guide the learning process.

3 Methods

Figure 1 shows the two phases of BAMBINO-LM. The learning phase involves continual pre-training a small-scale language model (*baby model* $\mathcal{B}$) whose initial pre-training was originally done on English data, while the feedback phase involves interactions with the Italian language model (*parent model* $\mathcal{P}$). During the learning phase, pre-training for $\mathcal{B}$ is continued by employing causal language modeling on Italian data. Causal language modeling (CLM), also known as next token prediction, is a standard technique to train a decoder-only model. The objective is defined as follows:

$$\mathcal{L}_{\text{CLM}} = -\frac{1}{|x|}\sum_{i=0}^{|x|}\log \mathbb{P}(x_t \mid x_0, \ldots, x_{t-1}).$$

There are two architectural innovations in BAMBINO-LM:

Feedback phase based on PPO We construct prompt x by selecting the first k tokens from the training example and solicit output $y_{\mathcal{B}} = \mathcal{B}(x)$ from the baby model. We then use Proximal Policy Optimization (PPO) where $\mathcal{B}$'s parameters are updated according to a clipped surrogate objective (Schulman et al., 2017). This objective moderates the updates to the policy, facilitating stable and efficient learning by incorporating a clipping mechanism. Its definition is given as follows:

$$\mathcal{L}_{\text{PPO}} = \hat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta);\epsilon)\hat{A}_t\right)\right],$$

with θ being the model parameters and $r_t(\theta)$ defined as:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}.$$

In the autoregressive setting of language modeling, θ controls the generation of tokens based on the given state or context s_t. The probability ratio $r_t(\theta)$ quantifies the change in the likelihood of selecting action a_t (the next token), under the updated policy parameters compared to the previous parameters θ_{old}. This ratio provides understanding on the impact of parameter updates on the policy's behavior, ensuring that changes do not excessively deviate from the previous policy, thereby maintaining training stability. The clipping mechanism, defined by $\text{clip}(r_t(\theta);\epsilon)$, restricts $r_t(\theta)$ within the bound $[1-\epsilon, 1-\epsilon]$, mitigating the risk of large policy updates that could lead to divergence.

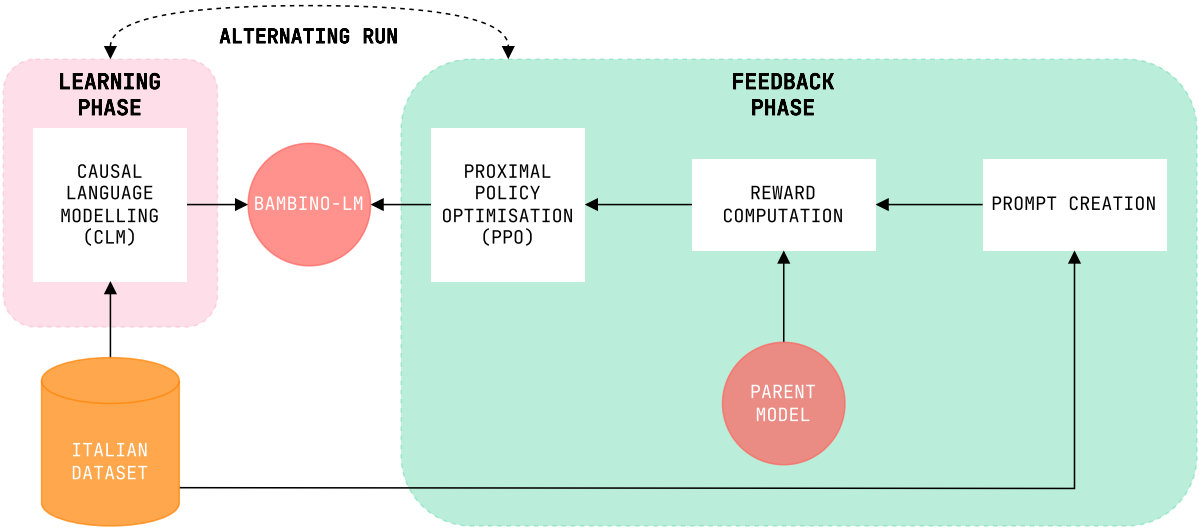

Figure 1: Architecture of BAMBINO-LM.

The advantage function, $\hat{A}_t = R + \gamma V(s_{t+1}) - V(s_t)$, which reflects the relative gains of selecting a_t given s_t. This function guides the optimization process by favoring actions that lead to better than expected outcomes.

The reward R for the advantage function is then calculated using the following function:

$$R(y_{\mathcal{B}}) = \frac{\alpha}{\beta(\text{PPL}_{\mathcal{P}}(y_{\mathcal{B}}) - \tau)}, \quad (1)$$

where α and β are parameters, $\text{PPL}_{\mathcal{P}}$ represents the perplexity of the parent model $\mathcal{P}$ for the sequence $y_{\mathcal{B}}$, and τ is a threshold value for perplexity. We use the following formulation of perplexity:

$$\text{PPL}(x) = \exp\left[\sum_{i=0}^{|x|} \log \mathbb{P}(x_t \mid x_0, \ldots, x_{t-1})\right].$$

Alternating run We adopt an alternating run strategy between the learning and feedback phases, which is summarized in Algorithm 1. The rationale behind this is two-fold: 1) this strategy simulates frequent interactions between a child and its parent through dialogues, which has been our main motivation behind this study; 2) using multiple rewards is shown beneficial for reinforcement learning (Dann et al., 2023). To expand on the second point, our findings further suggest that using perplexity as a reward can lead to exploitation when baby model $\mathcal{B}$ attempts to produce similar utterances to those coming from parent model $\mathcal{P}$. Without this strategic alternation between CLM and PPO, the pre-training tends to produce undesirable behaviours such as repeating words.

Algorithm 1 BAMBINO-LM Training.

procedure TRAIN($\mathcal{D}, \mathcal{B}, \mathcal{P}$)
  Input: pre-training dataset $\mathcal{D}$, baby model $\mathcal{B}$, and parent model $\mathcal{P}$.
  $r_{\text{CLM}}, r_{\text{PPO}} \leftarrow 10, 2$
  $r \leftarrow r_{\text{CLM}} + r_{\text{PPO}}$
  for $i, x \in$ enumerate($\mathcal{D}$) **do**
    if $i \% r < r_{\text{CLM}}$ **then**
      perform CLM step
    else
      $y_{\mathcal{B}} \leftarrow \mathcal{B}(x[1..k])$
      reward $\leftarrow R(y_{\mathcal{B}})$
      perform PPO step
    end if
  end for
end procedure

4 Experiment Setup

Mimicking their process to create the BabyLM challenge corpus (Warstadt et al., 2023), we create an Italian dataset that is comparable in size to the *strict-small* track of the challenge, and perform identical preprocessing[1]. Table 1 shows the statistics of the Italian language dataset. A cursory quality check was conducted to ensure that the dataset was in a readable format.

For the choice of the baby model, we use English baseline `OPT-125m` (Zhang et al., 2022) model for the *strict-small* track provided by the BabyLM organizers. For the parent model, we

[1] https://github.com/babylm/babylm_data_preprocessing; Accessed on 13th May, 2024.

use `gpt2-small-italian` model by de Vries and Nissim (2021). Using the Italian dataset described above, we conduct continual pretraining over 10 epochs, consisting of 10 learning phase steps followed by 2 feedback phase steps. We use $k = 5$ to solicit the first few tokens for prompting the baby model. All models are trained using HuggingFace's `transformer` (Wolf et al., 2020) and `trl` (von Werra et al., 2020) library.

Dataset	%
CHILDES (MacWhinney, 2000)	2.23
DailyDialog (Li et al., 2017)	4.45
QED (Abdelali et al., 2014)	11.86
OpenSubtitles (Lison and Tiedemann, 2016)	27.58
Standardised Project Gutenberg Corpus (Gerlach and Font-Clos, 2020)	16.19
Children's Story [2]	18.57
Wikipedia [3]	19.10

Table 1: Italian dataset used for continual pre-training.

For downstream tasks, we use four Italian language tasks in UINAUIL (Basile et al., 2023) and four English language tasks in GLUE (Wang et al., 2018). The tasks were selected primarily based on computational constraints for the project. We also include BLiMP (Warstadt et al., 2020) for that it is used in the original BabyLM challenge. All tasks are conducted in a *zero-shot classification* setting.

5 Results

Table 2 shows a significant improvement in Italian downstream tasks for BAMBINO-LM as compared with the BabyLM baseline. Specifically, we achieve an average improvement of 0.1197 $(0.3416 \rightarrow 0.4613)$ without substantial differences in English classification tasks. However, we notice an expected decrease of 0.0752 $(0.6255 \rightarrow 0.5503)$ in the English language BLiMP dataset. These observations are in line with Yadavalli et al. (2023) which show that native child-directed speech can lead to negative cross-lingual transfer and impede L2 acquisition depending on the choice of L1.

In Table 3 we examine two ablated versions of our model: (a) **w/o PPO:** Trained solely on the CLM objective with no feedback phase; (b) **w/o alternating:** Use BAMBINO-LM with no alternating runs. Instead, it trains with the CLM objective for the first 85% of each epoch and then switches to the PPO objective for the remaining 15%.

Task / Model	BabyLM	BAMBINO-LM
UINAUIL		
HaSpeeDe	**0.4774**	0.4592
IronITA	0.4966	**0.5516**
SENTIPOLC	0.1575	**0.4050**
Textual Entailment	0.4950	**0.5525**
Average	0.3416	**0.4613**
GLUE		
MNLI	0.3472	**0.3530**
MNLI-MM	0.3483	**0.3521**
RTE	**0.5271**	0.5199
SST2	0.5034	**0.5241**
Average	0.4315	**0.4373**
BLiMP		
Average	**0.6255**	0.5503

Table 2: Comparison of BAMBINO-LM with BabyLM.

Removing the interactive feedback mechanism (w/o PPO) and the alternating strategy (w/o alternating) significantly decreases Italian performance compared to our primary model. On UINAUIL tasks, the average score drops from 0.4613 to 0.4000 (w/o PPO) and 0.3513 (w/o alternating). However, we do not observe significant improvements in performance in both the English language task sets (GLUE and BLiMP). For GLUE tasks, the average scores remain consistent, with 0.4373 for BAMBINO-LM, 0.4375 for w/o PPO, and 0.4357 for w/o alternating. On the BLiMP dataset, the average scores are 0.5503 for BAMBINO-LM and 0.5554 for w/o PPO.

These results indicate that PPO modeling and alternating runs are both crucial for improving the bilingual ability of BAMBINO-LM without negatively impacting English performance. Furthermore, the lack of significant changes in English scores reinforces that these strategies enhance bilingual capabilities without compromising performance on existing benchmarks.

6 Conclusion

This paper introduces BAMBINO-LM, a continual pre-training strategy mimicking the process of sec-

[2] https://www.gutenberg.org/ebooks/bookshelf/353

[3] https://dumps.wikimedia.org/itwiki/

Task / Model	BAMBINO-LM w/o PPO	BAMBINO-LM w/o alternating
	UINAUIL	
HaSpeeDe	0.4798	0.4925
IronITA	0.4966	0.4989
SENTIPOLC	0.2775	0.1580
Textual Entailment	0.5500	0.5500
Average	0.4000	0.3513
	GLUE	
MNLI	0.3540	0.3522
MNLI-MM	0.3502	0.3545
RTE	0.5343	0.5271
SST2	0.5115	0.5092
Average	0.4375	0.4357
	BLiMP	
Average	0.5554	0.5268

Table 3: Results of the ablation experiments.

ond language acquisition in an interactive setting. BAMBINO-LM uses a two-phase approach: it incorporates reward from a parent Italian model into a PPO-based mechanism and alternates this procedure together with causal language modeling based on Italian language text. Our experiments demonstrate systematic improvement in Italian with a marginal but expected decrease in English, which echoes the past results in second language acquisition for large language models (Evanson et al., 2023). These findings highlight the efficacy of our approach in enhancing bilingual capabilities while maintaining performance in the original language.

In future work, we aim to explore the effect of alternative metrics and different reward learning mechanisms that better align with human feedback behaviors. This also includes exploring rewards that capture linguistic quality and provide direct, "constructive" corrections to the model output which is commonly known as an effective learning strategy for language development. Although BAMBINO-LM was applied for second language learning with Italian as an example, the method must be validated for other languages, especially languages that are distant from English or those that use a different set of tokens. The degradation in the performance of the first language, English, points to the potential of alternating with language modeling for the first language.

Limitations

The approach relies on the availability of a base model in a language, English in our case. Although we download Italian language datasets from known Italian sources, we do not explicitly validate the language of the text. We use the PPO model as is, and do not experimentally tune its parameters. Similarly, using perplexity as a metric for computing rewards may not be the optimal solution, as perplexity itself is influenced by many factors of the parent model.

Ethics Statement

The paper uses publicly available datasets for training and evaluation that do not possess known harms. The evaluative tasks are typical language learning tasks. However, the resultant models are not tested for harmful or biased content.

Acknowledgment

Zhewen Shen conducted this research as a part of UNSW Sydney's Taste of Research Program.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Valerio Basile, Livio Bioglio, Alessio Bosca, Cristina Bosco, and Viviana Patti. 2023. UINAUIL: A unified benchmark for Italian natural language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 348–356, Toronto, Canada. Association for Computational Linguistics.

Nitay Calderon, Subhabrata Mukherjee, Roi Reichart, and Amir Kantor. 2023. A systematic study of knowledge distillation for natural language generation with pseudo-target training. In *Proceedings of the 61st Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 14632–14659, Toronto, Canada. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Christoph Dann, Yishay Mansour, and Mehryar Mohri. 2023. Reinforcement learning can be more efficient with multiple rewards. In *International Conference on Machine Learning*, pages 6948–6967. PMLR.

Wietse de Vries and Malvina Nissim. 2021. As good as new. how to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.

Mehmet Demirezen. 1988. Behaviorist theory and language learning. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 3(3).

Linnea Evanson, Yair Lakretz, and Jean Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1).

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.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Terje Lohndal, Jason Rothman, Tanja Kupisch, and Marit Westergaard. 2019. Heritage language acquisition: What it reveals and why it is important for formal linguistic theories. *Language and Linguistics Compass*, 13(12):e12357.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Silvina Montrul. 2010. Current issues in heritage language acquisition. *Annual Review of Applied Linguistics*, 30:3–23.

David Samuel. 2023. Mean BERTs make erratic language teachers: the effectiveness of latent bootstrapping in low-resource settings. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 221–237.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. TRL: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.