Research Master Thesis

Grammaticality and LLMs: Evaluating the Potential of BabyLMs for Grammatical Error Detection in NLP

Malihehassadat Bani Fatemi

a thesis submitted in partial fulfilment of the requirements for the degree of

ReMA Humanities
(Human Language Technology)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Dr. Luis Morgado da Costa
2^{nd} reader: Sophie Arnoult

Submitted: June 27, 2025

Abstract

This thesis investigates the feasibility of using a smaller, more resource-efficient and environment-friendly RoBERTa-based Baby Language Model (BabyLM) for grammaticality assessment and grammatical error detection (GED), with a focus on three common grammatical error types (determiners, prepositions, and subject–verb-agreement) among English as Second Language (ESL) learners. While large language models (LLMs) such as RoBERTa have demonstrated good performance in GED tasks, their substantial computational and environmental costs motivate the exploration of small language models (SLMs). To this end, a few BabyLMs are trained using the training dataset from the strict-small track (9.96M words) of the first BabyLM Challenge, introduced in 2023. From these trained BabyLMs, the most promising model is selected based on the performance comparisons with the baseline RoBERTa BabyLM of the BabyLM Challenge on the BLiMP dataset. The best BabyLM is used for the following experiments of the thesis which involves two evaluation phases: a zero-shot grammaticality assessment and a fine-tuned GED classification task. For the zero-shot evaluation, the BLiMP dataset along with a BLiMP-style dataset for preposition errors, which I create from the BEA-2019 shared task dataset, is used to examine the BabyLM's sensitivity to the targeted error types over the its training stages (epochs) in comparison with the RoBERTa-base's performance. Results indicate that the BabyLM improves over time, showing the strongest performance in determiners and weakest in subject–verb-agreement. However, its overall zero-shot accuracy remains lower than the RoBERTa-base. In the fine-tuning phase, all models are fine-tuned on a sentence classification task with four classes including the three targeted error types and correct grammatical sentences. When fine-tuned, the BabyLM shows competitive performance, closely approaching that of RoBERTa-base, specially in detecting determiner errors. Despite a slight performance gap, the BabyLM achieves this with much fewer parameter numbers, far less training data, and reduced computational costs.

Declaration of Authorship

I, Malihehassadat Bani Fatemi, declare that this thesis, titled *Grammaticality and LLMs: Evaluating the Potential of BabyLMs for Grammatical Error Detection in NLP* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 27.06.2026

Signed: M. Bani Fatemi

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Luis Morgado da Costa, for his invaluable guidance, support, and patience throughout my thesis journey. Our weekly meetings provided structure, motivation, and critical feedback that were essential in shaping the experimental work of this thesis.

I am profoundly grateful to my family, whose unwavering support has encouraged me to continue trying despite the challenges of life. Thank you for standing by me during the most stressful moments of my studies, for always believing in my abilities, and for encouraging me to push forward even when things felt overwhelming. Your constant motivation and belief in my goals have meant more to me than words can fully express. I am also sincerely thankful to my friends, whose emotional support and encouragement carried me through the challenging moments. Your belief in me made all the difference.

Lastly, a special thank you to my cat, Dokme, who kept me company by sitting on my lap during long writing sessions. His quiet presence was a comforting reminder that I was never alone.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Large language models (LLMs) such as BERT, GPT, RoBERTa and have demonstrated remarkable capabilities in a wide range of natural language processing (NLP) tasks. Their good performance has often been attributed to their ability to learn deep contextual representations from large-scale datasets. However, the training and using such models come at a significant cost, both computationally and environmentally. The massive training data, computational resources, and energy consumption required to train and operate these models raises growing concerns in terms of sustainability and environment.

Researchers have shown an interest in working with smaller language models such as Baby Language Models (BabyLMs) (Warstadt et al., 2023a) that aim to retain as much of the performance of LLMs as possible while reducing resource requirements. Works such as Huebner et al. (2021) and the BabyLM Challenge (Warstadt et al., 2023a), exemplify this interest by encouraging the development of language models (LMs) trained on reduced datasets and with limited computational budgets. Huebner et al. (2021) train BabyBERTa with 8M parameters on a dataset comprising 5M words, which approximates to the linguistic input typically encountered by an average six-year-old English language learner. Their BabyBERTa achieved grammatical knowledge comparable to the pre-trained RoBERTa-base, but with ∼6000 times fewer words and ∼15 times fewer parameters. Similarly, the BabyLM Challenge (Warstadt et al., 2023a), introduced in 2023, encourages researchers to train LMs under a fixed data budget. At the time of writing this thesis, the third round of the BabyLM Challenge is underway[1].

Despite this recent interest to work with BabyLMs, their effectiveness in some language tasks like grammaticality assessment and Grammar Error Detection (GED) remains largely underexplored. This gap is also evident in the BabyLM Challenge, where GED has not been a primary focus. This gap provides an opportunity to investigate whether BabyLMs can meaningfully learn the notion of grammaticality assessment and effectively detect grammatical errors. Moreover, BabyLMs can act as a proxy to study in what order LLMs learn grammaticality judgment. The black box nature of LLMs makes it difficult to trace how linguistic abilities emerge during training. On the other hand, BabyLMs may be easier to use for studying how these models learn, for example by using interpretability methods. Their smaller scale makes it feasible to track the emergence of specific linguistic competencies over time, allowing researchers to use them as proxies for uncovering the internal mechanisms and learning trajectories that may also underlie larger models.

[1] The BabyLM Challenge website is available at: `https://babylm.github.io/index.html`.

1.1 Research Goals

My thesis aims to help to understand how the concept of grammaticality is learned by LLMs and to assess whether BabyLMs can be a suitable alternative to LLMs in GED. Specifically, it explores whether a BabyLM trained with significantly fewer resources can perform competitively with the RoBERTa-base model in this task. This is particularly relevant as it explores the potential to develop efficient, smaller language models that reduce computational costs and support environmental sustainability. The focus of my thesis is on three common grammatical errors including determiners, prepositions, and subject-verb-agreement among English as Second Language (ESL) learners. To this end, my thesis explores the following main research question, which is further explored through two sub-questions:

- **RQ1:** Can a RoBERTa-based BabyLM provide competitive grammaticality judgments when compared to the original RoBERTa model (based on three types of errors: determiner, preposition, subject-verb-agreement)?

- **RQ1-A:** How do these models compare in a zero-shot setting for the three targeted error types?

- **RQ1-B:** How do these models compare in a fine-tuned setting for the three targeted error types?

By addressing these questions, I aim to assess whether a BabyLM can serve as a feasible and environmentally friendly alternative to RoBERTa-base as an LLM for GED tasks while offering insights into its grammaticality assessment progression and performance trade-offs. The underlying hypothesis of the thesis is that, despite its smaller scale, the BabyLM's performance would be at least somewhat competitive with that of RoBERTa-base. If the BabyLM proves effective in the grammaticality assessment and the GED task, the findings could inform the broader development of resource-efficient small language models (SLMs), particularly for low-resource languages.

1.2 Structure

The background to this thesis is outlined in Chapter 2, which first describes the concept of grammaticality and, then, focuses on GED in NLP tasks. It goes on to discuss how BabyLMs relate to the current trends in LLMs, followed by the description of the BabyLM Challenge shared task. Finally, it surveys some of the available GED or Grammar Error Correction (GEC) datasets. Chapter 3 outlines the methodology of this thesis. It starts with the description of datasets, which are used in the experiments, and continues with describing the sets of experiments done in training the BabyLMs and two evaluation phases of zero-shot and fine-tuning for a GED task. Chapter 4 presents the results of the thesis experiments. It begins with the outcomes of training multiple BabyLMs. It then reports on the zero-shot evaluation results of the models and the results of fine-tuning the models for a GED classification task. Chapter 5 provides an error analysis based on the GED evaluation results. A reflection on the experimental findings, and some limitations of the study and a few suggestions for future research are provided in Chapter 6. Finally, Chapter 7 concludes the thesis with a summary of the main findings.

Chapter 2

Background and Related Works

This chapter presents a short review of GED task and its significance within NLP, followed by a discussion of BabyLMs as SLMs and their use in GED. Section 2.1 introduces the concept of grammaticality, the GED task, and three targeted grammatical error types, which are common among ESL learners, for this thesis. Section2.2 briefly describes LMs, LLMs, and SLMs. Section2.3 describes the BabyLM Challenge, its evaluation pipeline, and a summary of its top-performing systems. Section 2.4 introduces the GED classification task along with a review of some GED/GEC datasets. Section 2.5 discusses research intersecting GED and BabyLMs, followed by a chapter summary in Section 2.6.

2.1 Grammar Error Detection/Correction in NLP

Grammaticality refers to the conformity of a sentence to the syntactic and morphological rules of a given language, distinguishing correct constructions from ungrammatical ones (Chomsky, 2002). It plays an important role in linguistic theory and NLP and is a basis for syntactic parsing, language modeling, and grammatical error detection (Jurafsky and Martin, 2025). Grammaticality assessment involves evaluating whether a sentence follows the implicit structural rules of a language, which often relies on human judgments or computational benchmarks. In NLP, models are tested for their ability to differentiate between grammatical and ungrammatical sentences, providing insights into their grammaticality assessment (Warstadt et al., 2020).

Grammar error detection/correction in NLP can be categorized into three main sub-tasks: GED, GEC, and GEF (Grammatical Error Feedback). GED focuses on identifying grammatical errors in written text and is the first step in the error correction process and providing feedback. In contrast, GEC involves correcting the identified errors, while GEF aims to generate informative, often pedagogical, feedback to help users learn from their mistakes (Ng et al., 2014). GED is particularly important among these sub-tasks because it lays the foundation for the other two. Without accurately identifying errors, there would be no effective correction and feedback.

2.1.1 Grammar Error Detection

GED is an NLP task aimed at identifying grammatical errors within a given text. It generally focuses on detecting syntactic grammar errors such as subject-verb-agreement issues and inappropriate tense usage without providing corrections. GED is the initial step for more comprehensive grammar tasks, such as GEC (Napoles et al., 2017). GED systems typically use machine learning techniques and rely on annotated corpora to train models that are capable of distinguishing between grammatically correct and incorrect constructions (Bryant et al., 2023).

LLMs such as BERT, RoBERTa, XLNet, BART, and T5 have become the cornerstone of GED and GEC research (Bryant et al., 2023). These models, trained on extensive corpora and using Transformer architectures, excel in both classification and text-generation tasks. For GED, BERT-based models are commonly used to detect grammatical errors with high precision, while encoder-decoder models such as T5 and mT5 are used for generating corrected text in GEC tasks (Bryant et al., 2023). Evaluated on datasets such as BEA-2019 (Bryant et al., 2019), CoNLL-2014 (Ng et al., 2014), and CoNLL-2013 (Ng et al., 2013), these systems have shown a moderate performance, with top models like BART achieving a BEA-2019 ERRANT F0.5 score of 72.9 in GEC(Bryant et al., 2023).

Despite its importance, GED has often been overshadowed by GEC in research and shared tasks. Most shared tasks and benchmarks, such as CoNLL-2014, primarily emphasize correction, treating detection as a secondary component rather than a standalone task. However, accurate error detection is a critical prerequisite for effective correction, and improving GED systems can lead to more efficient and precise GEC models. Therefore, my research focuses on GED and studies the grammaticality judgment of the models over time. To this end, I use the Benchmark of Linguistic Minimal Pairs (BLiMP) dataset (Warstadt et al., 2020) (explained in Section 2.3) to check grammaticality assessments of the models. Additionally, I focus on three types of grammatical errors frequently encountered among ESL learners to study how well the models detect them.

2.1.2 Common Grammatical Error Types

The National Center for Education Statistics (2024) estimates that 10.6% of students in the US public school system speak a language other than English and have limited English proficiency. This shows the growing need for effective tools and strategies to support ESL instruction. I chose to focus on English because of its global significance as a lingua franca and the challenges it presents for non-native speakers, such as difficulties in everyday communication and engaging in academic discussions.

Studies on the most common grammatical errors made by ESL learners highlight the persistent challenges posed by certain aspects of English grammar. Bitchener et al. (2005) found that among 53 post-intermediate English for Speakers of Other Languages (ESOL) learners in New Zealand, the most frequent errors involved prepositions (29%), articles (20%), and verb tense (22%). Similarly, Dalgish (1991) analyzed errors made by ESL students at CUNY and identified articles (28%), vocabulary (20-25%), prepositions (18%), and subject-verb-agreement (15%) as the most common error types across learners of various first languages. Zhong and Yue (2022) also identify prepositions, determiners, and subject-verb-agreement among common grammar errors by ESL learners. For example, learners may incorrectly say “discuss about the issue” instead of “discuss the issue”. Or, a typical subject-verb-agreement mistake is “He go to school every day” instead of “He goes to school every day”. A determiner example is “This people are nice” instead of “These people are nice”. Notably, speakers of languages without article systems made significantly more article errors, though other error types were consistent regardless of linguistic backgrounds. For example, an article error example is "She bought car yesterday" instead of "She bought a car yesterday." These findings align with my teaching experience, where I have observed ESL learners frequently struggling with prepositions, determiners, and subject-verb-agreement. Even though these three grammatical error types do not cover the whole range of grammatical error types that ESL learners make, it covers a sufficient variation for the evaluation of the models' performance in a GED task. Moreover, there has been much effort on developing systems for detecting and correcting these error types such as BEA-2019 Shared Task (Bryant et al., 2019), CoNLL-2014 Shared Task (Ng et al., 2014), and HOO-2012 Shared Task (Dale et al., 2012) that shows their significance and makes them a suitable selection for this thesis study.

Previous research shows that the mastery of these grammatical error types occurs relatively late in early humans' language development. For instance, studies have shown that children do not fully understand subject-verb-agreement before the age of five (Johnson et al., 2005). The acquisition of prepositions in children follows a developmental trajectory that extends into the early school years. Children typically begin to produce basic locative prepositions like "in" and "on" around 27 to 30 months of age. More complex prepositions, such as "under", "back", "front", "beside", and "between", are generally learned between three and five years of age (Morgenstern and Sekali, 2009). Regarding determiners, studies have shown that children start using articles "a" and "the" between 18 and 61 months, with individual variation in the rate at which they omit or correctly use these determiners. When they are approximately 36 months old, the majority of children use determiners at a good level (Abu-Akel et al., 2004). It's important to note that while children begin using these grammatical elements at these ages, errors in their usage can persist as they continue to develop linguistic proficiency. Therefore, the mastery of these three grammatical error types is a gradual process that extends beyond the initial acquisition phase.

According to Dulay and Burt (1974), an order of acquisition may be found among child ESL learners who generally learn prepositions earlier than the determiners and subject-verb-agreement grammar. This can suggest that these learners first learn simpler syntactic elements before progressing to more complex grammatical structures. Moreover, Pienemann (2015) argues that language learners can produce only those linguistic forms they are cognitively ready to process. Therefore, it can be implied that ESL learners initially learn lexical morphemes like prepositions, which require less syntactic processing, before learning functional morphemes such as determiners and subject-verb-agreement grammar that need more advanced processing capabilities. Despite these arguments and the challenges of learning determiners for languages with different requirements concerning determiners, they may be learned earlier due to their high frequency in English language (Ellis and Collins, 2009). On the other hand, preposition error types tend to persist longer because they are semantically complex and often lack direct equivalents in the learners' native languages, which can result in persistent errors and confusion (Celce-Murcia and Larsen-Freeman, 1999). Subject-verb-agreement seems to be acquired over time and remains challenging for ESL learners even at intermediate and advanced levels, specially in contexts involving complex noun phrases (Killie, 2020; Stapa and Izahar, 2010).

2.2 Language Models

The current state-of-the-art in GED and GEC relies on LMs (Bryant et al., 2023). LMs are computational models trained to estimate the probability of word sequences, enabling tasks such as next-word prediction in models like GPT or masked token prediction in models like RoBERTa, depending on their training objectives. By analyzing patterns in its training dataset, a language model can evaluate the likelihood of encountering specific word sequences in an unseen text (Hiemstra, 2009). In modern NLP, deep learning techniques have revolutionized the field of language modeling. Advanced models such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and Generative Pretrained Transformer (GPT-3) (Brown et al., 2020) use transformer architectures to capture more complex and long-range dependencies between words, moving beyond simple n-gram models to encode richer contextual information. These models are trained on massive amounts of text data and are used for many NLP tasks, such as machine translation, sentiment analysis, and text generation.

2.2.1 Large Language Models

LLMs are advanced neural network architectures, typically based on transformers, which use attention mechanisms to capture complex linguistic patterns and contextual relationships across long text sequences. LLMs are pre-trained on massive, diverse textual sources and can be fine-tuned for specific NLP tasks. These models show remarkable performance, because of their scale, which often involves millions of parameters trained on a trillion of words (Jurafsky and Martin, 2025; Brown et al., 2020). Since 2020, they are among the top-performing models in GEC (Bryant et al., 2023). However, the computational and environmental costs of developing and operating LLMs are significant.

Ethical and societal concerns associated with LLMs have been highlighted by Bender et al. (2021). According to them, training LLMs demands substantial computational resources, leading to high energy consumption. For instance, training a single BERT-base model on GPUs is estimated to require as much energy as a trans-American flight. This energy usage contributes to environmental degradation that advocates for the development of smaller, more efficient models such as BabyLMs to mitigate such detrimental effects. Beyond environmental concerns, the reliance of LLMs on large-scale datasets also poses a limitation for many low-resource languages. These languages often lack sufficient annotated data, which would restrict the applicability of LLMs primarily to high-resource languages. As a result, many NLP tasks remain underexplored in low-resource language contexts, further reinforcing linguistic inequalities in access to advanced language technologies (Joshi et al., 2020). Developing smaller, data-efficient models can help bridge this gap and promote more equitable access to NLP technologies across languages.

2.2.2 Small Language Models

SLMs are small neural networks with significantly fewer parameters compared to LLMs. Research by Eldan and Li (2023) explored training smaller models on a simplified dataset, TinyStories, which uses child-like language constraints. Their findings suggest that SLMs, despite their reduced size, perform well on specific linguistic tasks such as basic grammar and sentence structure when trained on carefully created datasets.

Notably, SLMs show high competence in some language tasks relative to the limited and simplified data they receive, indicating that dataset quality and task-specific design can effectively compensate for smaller model capacity (Eldan and Li, 2023). Moreover, SLMs offer environmental benefits due to their reduced computational requirements and are more interpretable compared to their larger counterparts.

However, SLMs still exhibit limitations in some tasks like those that require nuanced reasoning or creative text generation, which shows the correlation between model size and complex linguistic capabilities. Research initiatives like those of Huebner et al. (2021) and Eldan and Li (2023) and challenges such as BabyLM Challenge (Warstadt et al., 2023a) provide valuable insights into the potential and limitations of SLMs such as BabyLMs. These studies highlight that while smaller LMs can achieve high performance on well-defined linguistic tasks with limited data, scaling up model size remains crucial for achieving higher-order language understanding and generation. By focusing on smaller-scale data and models, these efforts aim to better understand the mechanisms of language learning in LMs while addressing the computational and ethical limitations of LLMs.

2.3 The BabyLM Challenge

BabyLM (Baby Language Modeling) Challenge focuses on understanding and replicating the developmental trajectory of human language acquisition through computational models. The BabyLM Challenge, introduced in 2023, emphasizes the pretraining of LMs on developmentally plausible corpora, such as child-directed speech, children's literature, and simplified texts (Warstadt et al., 2023a). The BabyLM Challenge limits its datasets' size to approximately ∼10M and ∼100M words, simulating the restricted linguistic input available to children.

Unlike LLMs' pretraining approaches that rely on massive and diverse datasets (e.g., BERT and GPT-3), the BabyLM Challenge tries to optimize language model training under a fixed data budget, encouraging the development of models that learn linguistic competence from minimal and simplified input. This approach provides a foundation for exploring development of more human-like and efficient language models (Warstadt et al., 2023a). There are two available datasets for training a BabyLM including BabyLM Challenge dataset (Warstadt et al., 2023a) and TinyStories dataset (Eldan and Li, 2023). Since TinyStories dataset is created out of generated data, I will use the BabyLM Challenge dataset during this thesis. The following subsections briefly describe the BabyLM Challenge's training datasets and its zero-shot evaluation paradigm.

2.3.1 The BabyLM Challenge Datasets

The BabyLM Challenge training dataset is a pretraining corpus designed to approximate the linguistic input received by children. The dataset consists of two tracks: a strict track, which contains 98.04 million words, and a strict-small track, a scaled-down version with 9.96 million words, which is in alignment with the estimated linguistic exposure of children in their early developmental years (Warstadt et al., 2023a).

The first BabyLM Challenge corpus (Warstadt et al., 2023a), introduced in 2023, is predominantly composed of transcribed or scripted speech ($\approx 56\%$), reflecting the natural linguistic environment of young children, where spoken input is the primary source of language acquisition. However, it is important to note that while much of this data originates from spoken language, it has been transcribed into written form, making it different from written text. This distinction is particularly relevant when assessing GED, as GED or GEC datasets are typically composed of explicitly written texts, which may follow different grammatical norms and conventions compared to transcribed speech. Furthermore, around 40% of the dataset is taken from sources that are either intended or appropriate for children including children's books, dialogue, child-directed speech, and educational video subtitles. The first BabyLM Challenge pretraining corpus is composed of ten datasets. The largest dataset in the first BabyLM Challenge pretraining corpus is the OpenSubtitles dataset, which accounts for 31% of the total data. Following this, the Simple Wikipedia dataset makes up 15% of the corpus. The smallest dataset, the Switchboard Dialog Act Corpus, contributes just 1% of the total data (Warstadt et al., 2023a).

The second BabyLM Challenge (Hu et al., 2024), introduced in 2024, introduces some changes compared to its 2023 version, primarily in two areas: dataset flexibility and the inclusion of a multimodal track. The participants of the second BabyLM Challenge are allowed to bring their own datasets as long as they stay within the word limits (∼100M for the strict track and ∼10M for the strict-small track), enabling improvements beyond the baseline dataset. This decision was motivated by findings that link data quality to performance gains in large-scale language models (Hu et al., 2024). Moreover, the second BabyLM Challenge introduced a Multimodal track, incorporating an aligned text-image dataset and a new evaluation pipeline. The text-only dataset was also modified, with a notable increase in child-oriented content (70%, up from 39%) and transcribed speech (58%, up from 55%). Wikipedia (except for Simple English Wikipedia) and QED were removed, while reliance on OpenSubtitles was reduced due to concerns over scripted speech. CHILDES now constitutes a larger portion of the dataset, increasing child-oriented discourse from 5% to 29% (Hu et al., 2024). These refinements aim to better model child language acquisition and encourage more ecologically valid pretraining approaches. The third BabyLM Challenge (Charpentier et al., 2025) follows the same guideline for the training dataset.

Since my thesis is exploratory and focuses on understanding how models learn grammaticality assessment while also minimizing computational costs and reducing CO_2 emissions, I use the training dataset from the strict-small track of the first BabyLM Challenge (Warstadt et al., 2023a). Also, the first BabyLM Challenge has introduced three baseline models, with OPT-125M, RoBERTa-base, and T5-base architectures, that are trained on its training datasets for both the strict-small and strict tracks. Therefore, I work with this dataset because it provides a baseline RoBERTa model for comparison. The second and the third versions would invite other research questions, for example, what kind of data would be most suitable to pre-train models for GED.

2.3.2 The BabyLM Evaluation Paradigms

The BabyLM Challenge evaluation paradigms assess language models in data-efficient learning scenarios, simulating the limitations of early language acquisition. By training on limited datasets, these paradigms evaluate how well models generalize and adapt to linguistic tasks with minimal exposure. The BabyLM Challenge includes two main evaluation paradigms, fine-tuning and zero-shot. Its fine-tuning-based evaluations are done on a subsample of (Super)GLUE (General Language Understanding Evaluation) tasks, which include some text classification tasks, and Mixed Signals Generalization

Set (MSGS) tasks. The (Super)GLUE task types vary from paraphrase detection and sentiment classification to natural language inference, commonsense reasoning, question answering, and acceptability judgments (Warstadt et al., 2023a). In this thesis, I do not focus on this paradigm, but focus on the second one, which is a zero-shot evaluation of the BabyLMs. It is briefly described below.

Zero-Shot Evaluation on the BLiMP Dataset

Zero-shot evaluation is a technique used to assess a model's ability to generalize to tasks it was not explicitly trained for, without exposure to task-specific data (Palatucci et al., 2009). In the BabyLM Challenge, evaluation on the BLiMP dataset is performed in a zero-shot setting to check grammaticality assessments of models. The BLiMP dataset, introduced by Warstadt et al. (2020), is designed to evaluate how well LMs have learned the syntactic and semantic structures of their pretraining data. The BLiMP dataset consists of 67 datasets, which are grouped into categories such as subject-verb-agreement and ellipsis. Each dataset contains 1000 minimal pairs of sentences, which includes one sentence that is grammatical (sentence-good) and one that is ungrammatical (sentence-bad). For example, "Raymond is selling this sketch" is "sentence-good" while its pair, "Raymond is selling this sketches", is "sentence-bad" in the BLiMP's determiner-noun-agreement dataset.

Moreover, the BabyLM Challenge provides a filtered BLiMP dataset which is filtered based on the vocabulary of its training dataset. They keep examples in which each word has appeared at least twice in the training dataset (Warstadt et al., 2023a). In addition, the BabyLM Challenge introduces a supplementary set of test suites that extends the BLiMP dataset by covering linguistic phenomena not originally included, such as dialogue and questions (Warstadt et al., 2023a). These supplementary tests, created using semi-automatically generated templates, use the minimal-pair approach.

The BLiMP dataset provides sufficient data for two of the targeted error types in this thesis—determiners and subject-verb-agreement—but lacks data for prepositions. Therefore, a BLiMP-style dataset for prepositions is created and filtered using the BabyLM Challenge filtering approach, as described in Section 3.1.3. The BEA-2019 dataset (Bryant et al., 2019) is used to create the BLiMP-style dataset for preposition errors because it is the most recent dataset with prepositional error labels. Moreover, the BEA-2019 dataset is one of the available datasets that provides some subcategories for preposition error types and, unlike some other GED/GEC datasets, does not use a general label of "PREP" in its classification.

The zero-shot evaluation assesses whether models can identify the acceptable sentence in each pair by assigning it a higher probability than the unacceptable one. However, evaluating on the BLiMP dataset only indicates which sentence is more grammatical relative to its pair. Probability alone cannot determine grammaticality, which is a limitation of the zero-shot approach. Nonetheless, it remains useful for this thesis' experiments. While the BLiMP dataset provides a useful starting point for evaluating grammaticality assessments of models, fine-tuning on available GED/GEC datasets can better show the models' performance in detecting grammatical errors.

A Review of the BLiMP Scores in the First BabyLM Challenge

The performance of the top systems in the BabyLM Challenge on the BLiMP dataset shows some variations across the strict and strict-small tracks (Warstadt et al., 2023a).

In the strict track, BootBERT (Samuel, 2023) achieves the highest score (0.86). BootBERT uses a latent bootstrapping approach, where a student model learns not only to predict tokens but also to match contextualized embeddings from a teacher model, which is updated as a moving average of the student. In the strict-small track, which is related to my thesis topic, ELC-BERT (Charpentier and Samuel, 2023) achieves the highest BLiMP score (0.80), which is slightly below the RoBERTa-base model (0.87). ELC-BERT extends the LTG-BERT architecture by allowing each transformer layer to receive a learnable weighted combination of all previous layers' outputs, emphasizing recent and embedding-level representations. The second top-performing model in this track is Masked Latent Semantic Modeling (MLSM) (Berend, 2023) that achieves the BLiMP score of 0.79. MLSM modifies the output target from one-hot vocabulary labels to sparse semantic property vectors and applies a knowledge distillation-like training scheme. While MLSM alone lowers the BLiMP score, combining it with standard masked language modeling (MLM) improves its performance. McGill-BERT (Cheng et al., 2023), although competitive, achieves the BLiMP score of 0.75 and falls behind the other two systems in this track. McGill-BERT focuses on optimizing data formatting strategies—such as using full sentences instead of documents, avoiding sequence packing, and reducing maximum sequence length—which could boost the performance without architectural changes.

While the BLiMP datset provides a framework for evaluating the grammaticality of model predictions, it is important to note that the BabyLM Challenge does not explicitly emphasize GED as a focal task and does not investigate at which stage of training these models generalize grammaticality assessment. The BLiMP dataset primarily measures syntactic and grammatical acceptability by comparing the probabilities assigned to the minimal pairs of sentences, where one is grammatically correct and the other is incorrect. This setup evaluates whether models can generalize over the concept of grammaticality, but it does not reveal how or when this generalization emerges during the pretraining. Moreover, it does not directly address GED, which is a more complex task requiring nuanced grammatical understanding. This distinction highlights a potential gap in this evaluation. In addition, without studying the developmental trajectory of grammaticality assessment in BabyLMs, it remains unclear how early linguistic competence is formed.

2.4 GED Evaluation

As it was mentioned in Section 2.3.2, evaluating the models through fine-tuning on available GED/GEC datasets provides a better assessment of their ability to detect grammatical errors. In this thesis, the models are fine-tuned for a GED classification task (as presented in Section 3.2.3) to study their ability in detecting the selected grammatical error types. Fine-tuning involves taking a pre-trained language model and adapting it to a specific task by further training it on a task-specific dataset. This approach builds on the concept of transfer learning, where a model trained on a large, general corpus is then refined for specialized tasks, using knowledge learned during pre-training (Howard and Ruder, 2018). Fine-tuning pre-trained models offers several advantages, including reduced training time over training a model from scratch and the ability to achieve high performance with relatively small task-specific datasets. This efficiency is especially important in domains where annotated data is scarce or expensive to obtain (Devlin et al., 2019; Howard and Ruder, 2018).

2.4.1 Review of GED/GEC Shared Task Datasets

A variety of datasets have been developed to support research in GED and GEC. These datasets range from early efforts such as the HOO-2011 (Dale and Kilgarriff, 2011) and HOO-2012 (Dale et al., 2012) shared tasks, which focus on specific error types in academic writing, to larger and more comprehensive benchmarks like the CoNLL-2013 (Ng et al., 2013) and CoNLL-2014 (Ng et al., 2014) shared tasks, which expand error types coverage and emphasize essay-style learner texts. In addition, datasets like BEA-2019 (Bryant et al., 2019) provides a large-scale English learner corpus. In contrast, MultiGED-2023 (Volodina et al., 2023) introduce multilingual corpora to encourage the development of systems capable of handling multiple languages. Together, these datasets provide a foundation for advancing GED and GEC research. Below, I have shortly reviewed the available GED/GEC datasets, with a special focus on the availability of data concerning preposition error types.

HOO-2011 Shared Task

The Helping Our Own (HOO) 2011 Shared Task dataset was introduced as part of European Workshop on Natural Language Generation (ENLG) to promote research in GED/GEC. The HOO-2011 dataset was developed to support the development of automated writing assistance technologies (Dale and Kilgarriff, 2011). The dataset for the HOO-2011 Shared Task was composed of text fragments derived from 19 source documents. Each of these documents was originally a research paper that had been previously published in the proceedings of an Association for Computational Linguistics (ACL) conference or workshop. This dataset covers some error types such as article, preposition, and punctuation. However, the HOO-2011 dataset uses a single, general label—“PRE”——without distinction between subtypes such as incorrect choice, omission, or addition. This labeling limits the depth of analysis possible for prepositional errors. Therefore, despite the presence of prepositional instances, the HOO-2011 dataset is not a suitable resource for extracting preposition error data for the thesis.

HOO-2012 Shared Task

The HOO-2012 Shared Task dataset was created to advance research in GEC, particularly focusing on errors made by non-native English speakers. Organized as part of the BEA-2012 Workshop (Building Educational Applications), the HOO-2012 Shared Task concentrates primarily on determiner and preposition errors as the most common types of errors made by non-native speakers of English (Dale et al., 2012). The HOO-2012 dataset was constructed from the CLC FCE dataset, where “CLC” refers to the Cambridge Learner Corpus and “FCE” is a subset of CLC and stands for the First Certificate in English. The CLC FCE dataset contains exam scripts written by candidates who took the Cambridge ESOL FCE examination in 2000 and 2001. Moreover, the HOO-2012 dataset is built on the structure of the HOO-2011 dataset, but offers a detailed annotation and some subtypes for its targeted errors. Given its focus on preposition errors and targeted annotation, the HOO-2012 dataset is suitable for extracting preposition-specific grammatical errors.

CoNLL-2013 Shared Task

This dataset was introduced as part of the Conference on Computational Natural Language Learning (CoNLL) 2013 Shared Task on GEC, which aimed to evaluate systems capable of correcting a wide range of grammatical errors in non-native English writing (Ng et al., 2013). The CoNLL-2013 dataset is based on the National University of Singapore Corpus of Learner English (NUCLE) which consists of essays mainly written by Asian undergraduate students at the National University of Singapore. These essays represent authentic learner writing, containing a variety of grammatical errors that are typical of ESL learners (Ng et al., 2013). The CoNLL-2013 dataset targets five specific error types including article or determiner, preposition, noun number, verb form, and subject-verb-agreement. Similar to the HOO-2011 dataset, CoNLL-2013 dataset includes preposition errors, but they are annotated under a single general label—"Prep"—making it less suitable for the thesis.

CoNLL-2014 Shared Task

The CoNLL-2014 dataset (Ng et al., 2014) was introduced as part of the CoNLL-2013 Shared Task. The CoNLL-2014 shared task is built on the CoNLL-2013 shared task, but includes more error types, such as noun possessive and incorrect word order, and introduces some changes such as using F0.5-score as the evaluation metric. The dataset remains one of the most widely used resources for developing GEC models, providing a standardized framework for comparing system performance. The CoNLL-2014 dataset consists of essays written by ESL students. These essays were sourced from the NUCLE corpora. Each sentence in the dataset is annotated with grammatical errors and their corresponding corrections, allowing researchers to train and evaluate GEC models (Ng et al., 2014). The CoNLL-2014 dataset covers 28 error types. Although the preposition errors are included among these categories, they are annotated using a single general label—"Prep"—for all preposition-related errors. As a result, while it supports the identification of prepositional errors, it does not distinguish between subtypes, making it less suitable for the thesis.

The BEA-2019 Shared Task

The BEA-2019 dataset was introduced as part of the Building Educational Applications (BEA) 2019 Shared Task on GEC, which aimed to advance automated methods for improving learner writing. With a primary motivation of the need to reassess the field following a five-year hiatus, the BEA-2019 dataset continues the tradition of the previous HOO and CoNLL shared tasks. The dataset includes texts written by English language learners at different proficiency levels, ensuring a diverse range of grammatical errors (Bryant et al., 2019). Errors in the dataset are categorized into 25 main error types based on the ERRANT error type distributions. The corrections provided in the dataset are aligned using the M2 format, which allows for multiple correction possibilities for a given error. This flexibility is essential for GEC research, as many sentences can be corrected in different valid ways.

The BEA-2019 dataset consists of multiple learner corpora that were carefully selected to provide a broad and representative range of grammatical errors. The primary source of data is the CLC FCE dataset. In addition, the dataset includes data from the Lang-8 Corpus of Learner English, a collection of English-language writing samples produced by non-native speakers and corrected by native speakers. The other corpora is NUCLE, which contains academic essays written by university-level learners, covering a range of topics. (Bryant et al., 2019). The BEA-2019 shared task introduced a new annotated dataset (W&I+LOCNESS dataset) from the Cambridge English Write & Improve (W&I) and LOCNESS corpus. W&I+LOCNESS dataset comprises essays written by learners of English at different proficiency levels (A, B, and C) along with essays by native speakers (Bryant et al., 2019). The BEA-2019 dataset is suitable to create a BLiMP-style dataset for prepositions because it has three subcategories for it including “missing”, “replacement”, and “unnecessary”.

MultiGED-2023 Shared Task

The MultiGED-2023 shared task dataset is a resource designed to advance research in multilingual GED and to encourage interest in NLP for lower-resourced languages. It encompasses data from five languages: Czech, English, German, Italian, and Swedish, each derived from second language (L2) learner corpora (Volodina et al., 2023). Every entry consists of a token paired with an annotation indicating its correctness (‘c’ for correct, ‘i’ for incorrect). This structure supports binary classification at the token level, enabling precise error detection. While MultiGED-2023 provides binary error annotations, it does not include error type labels. As a result, it is not suitable for extracting preposition-specific errors without further manual annotation.

MultiGEC-2025 Shared Task

The MultiGEC-2025 dataset is a comprehensive resource designed for the Multilingual Grammatical Error Correction shared task, encompassing 12 European languages (Masciolini et al., 2025). It is compiled by the CompSLA working group in collaboration with over 20 external data providers. This dataset includes 17 subcorpora. Each subcorpus contains original learner texts accompanied by one or more correction hypotheses, that acts as the ’gold standard’ for evaluating system outputs. While the dataset is rich in multilingual learner errors, it is not specifically annotated for error types such as prepositions, making it less suitable for extracting preposition error data.

2.5 BabyLMs and GED

The environmental and computational costs of LLMs remain substantial due to the high energy consumption that is needed for pre-training. Training models with millions of parameters results in a significant carbon footprint and raises concerns about the environment. In contrast, BabyLMs seem to be a promising alternative to reduce computational costs while enhancing interpretability. These smaller models help run experiments in resource-constrained environments and show a potential to balance efficiency and accuracy in NLP tasks. Moreover, their lower data requirements make them more accessible for low-resource language contexts, where large annotated corpora are scarce. By reducing reliance on massive datasets, BabyLMs offer a more inclusive approach to NLP development, helping to mitigate the current bias toward high-resource languages. However, research on BabyLMs in GED remains limited, with little exploration of their grammaticality assessment and GED capabilities. Below, I review some of the relevant studies that have investigated BabyLMs in this context.

One of the earliest research on BabyLMs is the work of Huebner et al. (2021) that encouraged further research in this area of study. Huebner et al. (2021) explore the grammatical capabilities of transformer-based models when exposed to language datasets similar to child-directed input. The authors developed BabyBERTa, a scaled-down version of RoBERTa trained on only five million words, reflecting the linguistic environment available to children aged one to six years. Despite using 15 times fewer parameters and 6,000 times fewer words than RoBERTa, BabyBERTa acquired comparable grammatical knowledge, achieving an accuracy nearly identical (80.5%) to pre-trained RoBERTa (81.0%) on grammar tests. This study shows that child-directed speech, with characteristics such as shorter sentences and lower lexical diversity, may be helpful for grammar learning in LMs.

One key aspect of BabyLM research concerns their ability to generalize grammatical structures. Misra and Mahowald (2024) investigate how LMs can learn rare syntactic constructions by generalizing from more common ones, focusing on the English "Article + Adjective + Numeral + Noun" (AANN) construction (e.g., "a beautiful five days"). They trained transformer models on systematically altered corpora and compared learning outcomes between the models exposed to natural AANN structures and those where such constructions were removed or replaced with ungrammatical variants. They found that LMs can generalize to rare AANN patterns even when they have not seen any exact instances during training. This learning is facilitated by related constructions (such as "a few days") and generalization improves when models encounter greater variability in the input data.

Another critical question in BabyLM research involves the nature of training data and its influence on the model performance. Edman et al. (2024) present a novel approach to the BabyLM Challenge, where the authors explore L2 learning strategies for training LMs. Instead of mimicking first language (L1) acquisition, which focuses on massive data input, they adopt methods from L2 learning, which emphasizes explicit linguistic information such as grammar, word meanings, and paraphrasing. Their study experiments with four types of linguistic data: lexical information (from Wiktionary), grammatical examples (from grammar books), paraphrased sentences, and a mix of typical BabyLM data. Their results show that paraphrase data significantly improves model performance, while grammatical data provides only marginal improvements, and lexical information does not boost performance. Overall, the study shows that the selection of training data plays an important role in the model's effectiveness, suggesting that L2-like data structures may offer valuable insights into efficient language model training. Similarly, Chen and Portelance (2023) explore grammar induction through probabilistic context-free grammars (PCFGs), finding that syntactic embeddings improve model performance, but only to a degree comparable to models with randomly initialized embeddings. These findings suggest that while grammatical structuring may help learning, other factors such as tokenizer design and hyperparameters play a more significant role in performance improvements.

An alternative approach to optimizing BabyLMs involves refining model architecture and training paradigms. Bunzeck and Zarrieß (2023) present an alternative approach to language model development by focusing on smaller architectures and reduced training data rather than scaling up model size. The authors trained small GPT-wee models for the BabyLM Challenge, drawing on insights from usage-based linguistics, including factors such as word frequency, length, and lexical frames. They also implemented curriculum learning techniques to structure training data. Their findings suggests that small models can achieve notable proficiency in standard evaluation tasks, sometimes outperforming larger baseline models in zero-shot and fine-tuned settings.

Despite this, the curriculum learning strategy provided no clear improvements, except in specific cases, indicating complex interactions between model architecture, data characteristics, and learning processes that require further exploration. Timiryasov and Tastet (2023) take a different approach by using knowledge distillation, training an ensemble of larger teacher models (GPT-2 and LLaMA) on a small 10M-word dataset, then transferring the knowledge to a smaller 58M-parameter LLaMA student model. This distilled model not only matched but exceeded the performance of its teachers and a similarly-sized non-distilled model across linguistic benchmarks. Their findings highlight that distillation can enhance sample efficiency and outperform conventional training, even on restricted datasets. Collectively, these studies illustrate that reducing model size does not inherently degrade performance, provided that strategic training methodologies are employed.

2.6 Research Focus

In this thesis, I try to explore the potential of using smaller, more environmental friendly language models (BabyLMs) for the GED task. While LLMs have achieved state-of-the-art results in GED, they come with significant drawbacks, such as high computational costs and environmental impact. BabyLMs, which are smaller models trained with reduced data and computational resources, have not yet been thoroughly investigated for this task. This gap is also evident in the BabyLM Challenge, where GED has not been the focus of study.

Given the bidirectional architecture of BERT models and and their widespread use in GED research, I have decided to use the RoBERTa architecture to train my BabyLM. RoBERTa architecture is known for its effective approach to MLM without relying on next-sentence prediction tasks, enabling better generalization across different language patterns (Liu et al., 2019). A BabyLM with RoBERTa architecture seems to have the potential to acquire linguistic competence with fewer parameters and a limited vocabulary, as shown by Huebner et al. (2021). I train multiple BabyLMs, varying in vocabulary size and hidden layer numbers and compare their performance to select the one that shows a higher performance in comparison with RoBERTa baseline model of the BabyLM Challenge, that I call BabyLM Baseline RoBERTa (BB-RoBERTa) in the thesis. Based on the results of this comparison, I use the selected BabyLM to do further experiments. The first BabyLM Challenge strict-small training dataset is used for training my BabyLMs. The results of the BabyLM training experiment are shown in Section 4.1 of Chapter 4.

The second part of My thesis focuses on comparing the grammaticality assessment of my selected BabyLM and the original RoBERTa-base model (that I call “O-RoBERTa” in my thesis) (Liu et al., 2019), with an emphasis on the trade-offs in performance when using a more ecological model. Moreover, using BLiMP-based zero-shot experiments, I study the progression of my BabyLM’s grammaticality assessment by evaluating its grammaticality assessment at its various training stages. This evaluation can help me study how well and in what order my BabyLM detects the three types of common grammatical errors among ESL learners (prepositions, determiners, and subject-verb agreement) in a zero-shot setting. The results of the zero-shot experiments are shown in Section 4.2 of Chapter 4.

Then, I fine-tune my BabyLM and O-RoBERTa for a GED classification task where sentences are classified as either grammatical or as containing one fo the three classes of errors discussed above. The fine-tuning is evaluated over time, at multiple training stages of my BabyLM that have been saved during the training phase. This analysis can provide deeper insights into the strengths and weaknesses of the BabyLM over time in handling the targeted error types. The results of the GED evaluations are shown in Section 4.3 of Chapter 4. Finally, an error analysis is done on some of the models' misclassifications in the GED task (as reported in Chapter 5) to identify possible patterns.

Chapter 3

Methodology

This chapter presents the methodology of my thesis experiments. It introduces the datasets that are used throughout the experiments in Section 3.1. Then, Section 3.2 presents the overall experimental setup, including BabyLM training, zero-shot evaluation, and fine-tuning for the GED classification task.

3.1 Data

Three datasets are used for my experiments. The strict-small dataset of the first BabyLM Challenge (introduced in Section 3.1.1) is used to train the BabyLMs. The zero-shot evaluation data is provided by the filtered-BLiMP dataset that is described in Section 3.1.2. To fine-tune the models on the GED classification task, I use the GED classification dataset (explained in Section 3.1.3).

3.1.1 Training Data

The training dataset is provided by the first BabyLM Challenge, introduced in 2023. The strict-small dataset of the first BabyLM Challenge is chosen for this thesis because it is text-only and all the participants had to use it for their experiments without bringing data from other external corpora (Warstadt et al., 2023a).

The second (Hu et al., 2024) and the third (Charpentier et al., 2025) BabyLM Challenge, introduced in 2024 and 2025 respectively, allow their participants to bring their own dataset as long as they stay with the word limits of their challenge datasets ($\sim$10M for strict-small track and $\sim$100M for strict track). Moreover, as previously discussed in Section 2.3.1, the second BabyLM Challenge has increased the proportion of the child-directed data from 39% to 70% and the transcribed speech data from 55% to 58% in comparison to the previous year and removed the Wikipedia corpus (Hu et al., 2024). I use the strict-small dataset from the first BabyLM Challenge, because I want to compare my BabyLM's performance with the baseline RoBERTa model (BB-RoBERTa), introduced by the BabyLM Challenge, to ensure that my selected BabyLM is competitive enough to run the following experiments of this thesis. This dataset also is more suitable for my thesis goal as I need a more written-text dataset. Moreover, the strict-small dataset is used to lower the computational cost and its impacts on the environment.

Both the strict and strict-small datasets of the first BabyLM Challenge are mostly scripted or transcribed speech ($\approx$ 56%) and child-directed language ($\approx$ 40%). Warstadt

		Words		
Dataset	**Domain**	**Strict-Small**	**Strict**	**Proportion**
CHILDES (MacWhinney, 2000)	Child-directed speech	0.44M	4.21M	5%
British National Corpus (BNC), dialogue portion	Dialogue	0.86M	8.16M	8%
Children's Book Test (Hill et al., 2016a)	Children's books	0.57M	5.55M	6%
Children's Stories Text Corpus	Children's books	0.34M	3.22M	3%
Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2018)	Written English	0.99M	9.46M	10%
OpenSubtitles (Lison and Tiedemann, 2016)	Movie subtitles	3.09M	31.28M	31%
QCRI Educational Domain Corpus (QED; Abdelali et al. (2014))	Educational video subtitles	1.04M	10.24M	11%
Wikipedia	Wikipedia (English)	0.99M	10.08M	10%
Simple Wikipedia	Wikipedia (Simple English)	1.52M	14.66M	15%
Switchboard Dialog Act Corpus (Stolcke et al., 2000)	Dialogue	0.12M	1.18M	1%
Total	-	9.96M	98.04M	100%

Table 3.1: The table is taken from the first BabyLM Challenge, introduced in 2023, where they present their datasets for the Strict and Strict-Small tracks (Warstadt et al., 2023a).

et al. (2023a) control the size and domain of their datasets to have a plausible data in the sense of approximating the amount and type of linguistic input a child might realistically be exposed to during early language learning. A training dataset with only written-text corpus of ESLs could have been better to run my experiments with, because some grammatical constructions are more frequent in writing than speaking (Biber, 1991). Moreover, the grammar errors made in ESLs' writings are generally different from those of speech and children's mistakes. Despite these limitations, which will be discussed in Chapter 6, I use the strict-small datasets from the first BabyLM Challenge to have a comparative baseline for training my BabyLM model. Table 3.1 summarizes the contents and portions of the BabyLM Challenge pretraining dataset. The strict-small dataset is approximately 10% the size of the strict track corpus. Below, data source of each subset of this dataset is described.

The Child Language Data Exchange System (CHILDES) is a multilingual database that includes transcriptions of adult–child interactions in different contexts such as controlled experimental settings of laboratories or natural environments of homes that were gathered by various researchers (MacWhinney, 2000). Huebner and Willits (2021) created a refined subset of CHILDES by isolating interactions with only American English-speaking children aged between 0 and 6 years. They removed all child utterances and tokenized the data, resulting in a dataset that has about five million words. In the strict-small dataset, there are 0.44 million words from this corpus. Another corpus is that of the British National Corpus (BNC). It is a hundred million word corpus from different domains of spoken and written British English from the late twentieth century (Wynne, 2022). The BabyLM Challenge uses only 10% of this corpus which includes the transcribed dialogues (Warstadt et al., 2023a). In the strict-small dataset, there are 0.86 million words from this corpus. Another corpus with dialogue domain is that of the Switchboard Corpus which includes transcribed telephone conversations among 500 speakers, totaling about three million words (Stolcke et al., 2000). Each conversation is between two strangers who informally talk about general topics. In the strict-small dataset, there are 0.12 million words from this corpus.

For transcribed speech, the BabyLM Challenge pretraining dataset also includes two corpora that consist of subtitles. In the strict-small dataset, they use the English subset of the OpenSubtitles (3.09M words), which is a collection of publicly available TV and movie subtitles sourced from a third-party website (Lison and Tiedemann, 2016). Moreover, they use the English portion of the QCRI Educational Domain Corpus (QDC) (1.04M words), which contains subtitles created by volunteers for educational videos (Abdelali et al., 2014).

Two of the corpora are from the children's book domain. One is Children's Book Test (CBT) that is a collection of more than a hundred children's books done by Hill et al. (2016b) for Project Gutenberg (Hill et al., 2016a). The original dataset includes pairs of questions to test named entity recognition, but the BabyLM Challenge does not include it in their pretraining data. The other corpus is the Children's Stories Text Corpus, that was collected for the purpose of developing a story generation system, which includes manually selected children's stories from Project Gutenberg (Warstadt et al., 2023a). There are 0.57M words from CBT and 0.34M words from the Children's Stories Text Corpus in the strict-small dataset.

The BabyLM Challenge includes some data from the written domain and Wikipedia. The Standardized Project Gutenberg Corpus is a curated and preprocessed collection of more than fifty thousand public domain literary works from Project Gutenberg, totalizing over three billion tokens. This dataset includes detailed metadata that allows the filtering of texts by language and publication date (Gerlach and Font-Clos, 2018). In the strict-small dataset, there are 0.99M words from this corpus. They use the English portion of Wikipedia which is an online encyclopedia authored by volunteers and hosted by the Wikipedia Foundation. They also use Simple English Wikipedia whose texts have shorter sentences, no idioms, and high frequent vocabulary. Wikipedia treats Simple English as a distinct language, so its articles are separate from those found in the English Wikipedia (Warstadt et al., 2023a). There are 0.99M and 1.52M words from these two corpuses in the strict-small dataset, respectively.

3.1.2 Zero-shot Evaluation Dataset

The zero-shot evaluation of the thesis is done on the BLiMP dataset (Warstadt et al., 2020). The BLiMP dataset consists of minimal pairs in which each pair has one grammatically correct sentence (sentence_good) and one incorrect sentence (sentence_bad), where usually one word is changed. For example, in this minimal pair for determiner-noun-agreement, "Raymond is selling this sketch" is "sentence_good" and "Raymond is selling this sketches." is "sentence_bad". In a zero-shot setting, the models are evaluated by comparing the probabilities of both sentences in the minimal pair, under the assumption that the models give higher probability to a good sentence and a lower probability to a bad sentence. However, it is important to note that this is not the same as detecting errors or suggesting corrections. The evaluation with the BLiMP dataset measures only the preference of a model for a grammatical sentence over an ungrammatical one, which is a basic skill of a good LLM. In fact, both "Raymond is selling this sketch" and "Raymond is selling these sketches" should ideally receive higher probability scores than the ungrammatical version.

The BLiMP dataset consists of 67 datasets, which are grouped into categories such as subject-verb-agreement, ellipsis, anaphora, and island effects, each targeting a specific grammatical phenomenon. Each of its datasets contains 1,000 minimal pairs-pairs of sentences that differ minimally in structure but contrast in grammatical acceptability. These pairs are carefully designed to separate specific syntactic, morphological, or semantic phenomena. The data is generated using linguist-crafted grammar templates, with human agreement on the correctness of labels reaching 96.4% (Warstadt et al., 2020).

As previously discussed in Section2.3.2, the original BLiMP dataset provides sufficient data for determiners and subject-verb-agreement—two of the targeted error types in this thesis. However, it does not include data for preposition errors, which is the third targeted error type. To address this limitation, I create a BLiMP-style dataset for preposition errors (introduced in Section3.1.3).[1] The BabyLM's filtered BLiMP dataset is first used to evaluate the BabyLMs in a zero-shot setting in order to select the best BabyLM for the following experiments (as described in Section3.2.1). Then, for the two evaluation phases of the thesis (outlined in Section3.2), I use the BLiMP-style preposition dataset in combination with the filtered BLiMP dataset, enabling a more complete assessment across all three targeted error types.

3.1.3 GED Classification Dataset

As previously discussed in Section 3.1.2, the BLiMP dataset does not include preposition errors. To address this gap, I create a BLiMP-style dataset by incorporating preposition error data from the BEA-2019 Shared Task (Bryant et al., 2019). Among available GED/GEC datasets, I use the preposition subset of the BEA-2019 Shared Task dataset because it distinguishes three subtypes of prepositional errors: "missing", "replacement", and "unnecessary". Although the HOO-2012 dataset also has similar subcategories for prepositions, the BEA-2019 dataset is used because it is the most recent dataset with prepositional error labels. The other available GED/GEC datasets have only the general "PREP" label for this error type. In the BEA-2019 dataset, "missing" class indicates that a preposition is missed while it is necessary. The "replacement" suggests that the wrong choice of preposition which needs to be replaced by its correct one. The "unnecessary" indicates that the preposition should be removed because it is unnecessary. The preposition classifications of BEA-2019 dataset allows a more linguistic analysis of prepositional usage and model sensitivity to these three preposition error types. However, the dataset does not offer more specific preposition classifications such as spatial or temporal, which limits its coverage of deeper semantic distinctions. Despite this, it offers a balance between three broad preposition error types and having enough data to analyze preposition errors effectively.

Following the approach of the BabyLM Challenge for filtering the BLiMP dataset, I create my BLiMP-style preposition dataset. They filter their evaluation datasets such as BLiMP, GLUE, and MSGS and keep only those examples in which every word appears at least twice in the strict-small training dataset (Warstadt et al., 2023a). Following their approach, I examine each minimal pair in the BLiMP-style preposition dataset and keep any instance where either the grammatical (sentence-good) or ungrammatical (sentence-bad) version contains words that appear at least twice in the strict-small training dataset. This filtering ensures consistency with the BabyLM evaluation setup and guarantees that the BabyLMs are only tested on vocabulary they have encountered during their training. Table 3.2 shows the number of sentences in the BLiMP-style preposition dataset before and after the filtering. For example, the data for the unnecessary preposition error reduces from 1075 to 865 instances after filtering or the missing preposition error reduces from 3055 to 2374 instances.

To fine-tune and evaluate the models for the three targeted error types, I use the determiner-noun-agreement and subject-verb-agreement data from the filtered BLiMP dataset along with the created BLiMP-style dataset for preposition. In addition to these three error types, I have another classification for correct grammatical sentences. I split

[1] The created dataset is available at:
https://github.com/Farnaz-BNF/BabyLM_GED/blob/main/BLiMP_styled_prep/preposition.json.

	Before Filtering	**After Filtering**
Unnecessary Preposition	1075	865
Missing Preposition	3055	2374
Replacement Preposition	1414	1154
Total	5544	4393

Table 3.2: Sentence Distribution of Error Types in the BLiMP-Style Preposition Dataset Before and After Filtering.

Fine-Tuning Classification	**Data**	**Train**	**DEV**	**Test**
Subject-Verb-Agreement (SVA)	5535	3321	1107	1107
Determiner-Noun-Agreement (DET)	7542	4525	1509	1508
Preposition (PREP)	4393	2636	878	879
Grammatical (G)	17470	10482	3494	3494
Total	34940	20964	6988	6988
Tokens	325046	194474	65097	65475

Table 3.3: Sentence Distribution by Classes in the GED Classification Dataset. The total number of tokens for each training, development, and test sets are also shown.

this dataset into training (60%), development (20%), and test (20%) sets. Since my BabyLM is trained on the strict-small dataset (9.96M words), increasing the training size too much during fine-tuning could lead to overfitting or memorization of grammatical patterns beyond what was seen in pretraining (de Vries et al., 2022). This way of splitting provides sufficient learning from the training data while preserving separate sets for tuning and final evaluation. Table 3.3 shows the distribution of the GED classification dataset, which is used for fine-tuning on the GED task. As the table shows, in the training subset of the dataset, there are 3,321 sentences for the "SVA" class (for subject-verb-agreement error types), 4,525 sentences for the "DET" class (for determiner error types), 2,636 sentences for the "PREP" class (for preposition error types), and 10,482 sentences for the "G" class (for the grammatical sentences).

To increase structural diversity in the grammatical class ("G"), 70% of its samples come from the sentence-good subset of the three targeted error types. The remaining 30% are randomly sampled from the sentence-good portions of other grammatical error types in the filtered BLiMP dataset.[2] This approach helps to study if the models can distinguish correct grammatical sentences with some grammatical structures other than the three targeted error types. After combining these 70% and 30% of the class "G", the entire dataset is shuffled so that the model sees the grammatical sentences from all error classes without any order. To ensure that the models encounter data across classes for the class "G", I first split each error type individually, then combine them into train, development, and test sets.

Overall, the training set of the GED classification dataset has 194474 tokens which is about 0.02% of the strict-small dataset (9.96M) that is used in training the BabyLMs. The training set (60% of the GED classification dataset) will be used to fine-tune the models for a sentence-level GED classification with the four possible classes.

The GED classification dataset is used to fine-tune the models for a GED clas-

[2] The distribution of sentences for the 30% part of the G class is shown in Table A.1 in Appendix A.

sification task. I spit this dataset into training (60%), development (20%), and test (20%) sets. The training set is used to fine-tune the models for a GED classification task (described in Section 3.2.3). The test set is used to evaluate the BabyLM and RoBERTa-base for the classification task of detecting the three targeted grammatical error types and grammatical sentences.

3.2 Setup Experiment

My experimental setup[3] consists of three phases, which are all done on an L4 GPU using Colab Pro. First, I train six BabyLMs (explained in Section 3.2.1) and select the most competitive one among them to run the follow up experiments. I use the selected BabyLM's zero-shot evaluation (introduced in Section 3.2.2) at each of its training stages to study its grammaticality assessment over time. Moreover, a zero-shot evaluation of O-RoBERTa is done to compare its performance with the BabyLM. The same evaluation is also performed for BB-RoBERTa. Then, the models are fine-tuned for the GED classification task (presented in Section 3.2.3) with the training set of the GED classification dataset (explained in Section 3.1.3). I do the fine-tune my BabyLM at its pretraining stages to study its performance over time. Finally, I evaluate the performance of the fine-tuned models on the GED task, with a primary focus on the BabyLM and O-RoBERTa, in classifying the three targeted error types and grammatically correct sentences.

3.2.1 Training BabyLM

In the first phase of my experiment, I train a BabyLM for my experiments. Given the bidirectional architecture of BERT-based models (Liu et al., 2019) and their established effectiveness in GED research (Bryant et al., 2023), I have chosen to use the RoBERTa architecture to train my BabyLM. Since one of my primary reasons for working with BabyLMs is their lower impact on the environment, I train a BabyLM with the strict-small (9.96M words) dataset of the BabyLM Challenge. Rather than optimizing for the best-performing BabyLM architecture, the focus of this thesis is to compare a standard BabyLM with RoBERTa. Therefore, I do not focus on implementing any special designs to train the best BabyLM, but rather on comparing the performance of the BabyLM with RoBERTa. Despite this, I run some experiments to train a competitive BabyLM, which I describe in the following paragraphs.

One challenge in working with BabyLMs is the absence of a universally accepted definition. While the dataset size is restricted in the BabyLM Challenge, there are no fixed constraints regarding the model's size, number of model parameters, or training epochs. Therefore, to guide the development of a BabyLM that aligns with the goals of the BabyLM Challenge, I first reviewed the configurations, that used RoBERTa architecture and competed in the strict-small track, reported in the Proceedings of the BabyLM Challenge (Warstadt et al., 2023b). The number of parameters among the submitted models for the challenge are between 0.75M to 125M. Since the winner of the strict-small track of the BabyLM Challenge shared task has 24M parameters, I set the maximum limit of number of model parameters for my BabyLM at 24M (Charpentier and Samuel, 2023).

[3] My setup experiment code is available at https://github.com/Farnaz-BNF/BabyLM_GED.

The BabyLM Challenge has introduced some baseline models trained on the datasets provided for the strict-small and strict tracks (Warstadt et al., 2023a). The baselines have three architectures including OPT-125M, RoBERTa-base, and T5-base. Each model has the original design and training objectives outlined in their original papers (OPT; Zhang et al., 2022, RoBERTa; Liu et al., 2019, T5; Raffel et al., 2020). The selection includes both decoder-only architectures, as represented by OPT-125M, and encoder-only and encoder-decoder configurations, as seen respectively in RoBERTa-base and T5-base. These models use different training objectives: OPT-125M uses next-token prediction, RoBERTa-base applies masked-token prediction, and T5-base relies on sequence-to-sequence matching losses. All baseline models are trained with the hyperparameter settings, including a fixed context length of 128 tokens, a learning rate of 1e-4, and a batch size of 128. Each baseline model is trained for 20 epochs, with the dataset shuffled randomly and independently before each epoch (Warstadt et al., 2023a). Since my BabyLM follows a RoBERTa architecture, I compare its zero-shot performance with that of BB-RoBERTa, which is a baseline for my thesis comparison, to select the best BabyLM for my experiments.

To ensure the competitiveness of my BabyLM in relation to BB-RoBERTa, I begin by training six BabyLMs, each varying in vocabulary size and hidden layers. I keep the other parameters such as RoBERTa architecture and training objectives fixed. According to Edman and Bylinina (2023), the vocabulary size of 40K provides the best performance on the BLiMP dataset. Therefore, I experiment with vocaulary size of 40K and lower (20K) and bigger (50K) vocabularies to check the validity of their results for my BabyLM. I also vary between four and six hidden layers to train my BabyLMs to ensure a balance between model complexity and performance, especially because of the limited size of the strict-small (9.96M words) dataset. According to Jiao et al. (2020), models with four and six hidden layers show a good performance in low-resource language understanding tasks. Moreover, this configuration is used in some of the submitted papers of the BabyLM Challenge, such as the works by Proskurina et al. (2023) or Veysel Çağatan (2023). Since this configuration still gives me about 80M parameter BabyLMs, I lower the number of the parameters by using one third of the hidden size of O-RoBERTa (256) and limit the model to eight attention heads. This hidden size is commonly used in the BabyLM Challenge and is employed in some submissions to the challenge, including those by Yang et al. (2023), Veysel Çağatan (2023), and Huebner et al. (2021). With this setting, I am able to keep my BabyLMs under my defined parameter threshold of 24M.

Table 3.4 shows RoBERTa's and the BabyLMs' configurations and number of parameters. The first BabyLM Challenge trains its baseline models for 20 epochs and does not set a boundary for the number of epochs. To keep the computational costs low and avoid overfitting, I train my BabyLMs for five and ten epochs, which is done by the majority of the submitted models in the strict-small track for RoBERTa-architecture BabyLMs (Warstadt et al., 2023b). Moreover, the third BabyLM Challenge allows its participant to do no more than ten epochs over their training data (Charpentier et al., 2025). Evaluating at both five and ten epochs allows for a better understanding of how the BabyLMs performance evolve during training and may help identify whether their performance plateau early or continue with more training. Since this evaluation results may be affected by random seeds, I run an additional experiment with some different random seeds (explained in Section 4.1 of Chapter 4) to better choose the best BabyLM for the following experiments.

BabyLM	**Vocabulary_Size**	**Hidden_Layers**	**Hidden_Size**	**Attention_Head**	**Epochs**	**Parameters_Number**
BB-RoBERTa	50K	12	768	12	20	125M
BabyLM-V20-L4	20k	4	256	8	5, 10	13M
BabyLM-V20-L6	20k	6	256	8	5, 10	16.5M
BabyLM-V40-L4	40k	4	256	8	5, 10	18M
BabyLM-V40-L6	40k	6	256	8	5, 10	22M
BabyLM-V50-L4	50k	4	256	8	5, 10	20.5M
BabyLM-V50-L6	50k	6	256	8	5, 10	24M

Table 3.4: The Configurations of BB-RoBERTa's and the Six BabyLMs.

For training the BabyLMs, I use the training BabyLM code made available by Rozema (2024)[4] as the basis of my BabyLM training setup since it also works with RoBERTa architecture. This code uses Huggingface transformers (Wolf et al., 2020) Trainer class to train a BabyLM. Moreover, I use its custom dataset which tokenizes sequences during training (Rozema, 2024). Since Rozema's thesis works with curriculum learning, I made some adjustments on its code to design my training setup. My experimental setup saves the BabyLMs at each of their training stages. The BabyLMs' training stages represent the number of epochs. In other words, the stages that I saved the model is at the end of each epoch. For example, the first stage is the BabyLM pretrained at the first epoch. So, when I talk about the BabyLM at an epoch in the thesis, it is the same as its training stage. This approach of saving the BabyLMs at each epoch allows me to study the development of the selected BabyLM.

To explore different vocabulary sizes, I create a ByteLevelBPETokenizer (Sennrich et al., 2016) based on the BabyLM training dataset. For each vocabulary size, I create one tokenizer and use it for training my BabyLMs with the same vocabulary size. After evaluating the six BabyLMs on the BLiMP evaluation pipeline of the BabyLM Challenge (which is shown in Section 4.1), I select the best BabyLM, which shows the closest performance to that of BB-RoBERTa (69.5%).

3.2.2 Zero-Shot Evaluation

In zero-shot evaluation, the models are tested on tasks they have not been explicitly trained on. The BabyLM Challenge evaluation pipeline, introduced in 2023, is used for this evaluation (Warstadt et al., 2023a). Using the filtered BLiMP dataset in combination with my filtered BLiMP-style preposition dataset, I compare the grammaticality assessment of the selected BabyLM with those of BB-RoBERTa and O-RoBERTa by evaluating their performance. As discussed in Section 3.1.2, the BLiMP dataset consists of minimal pairs in which each pair has one grammatically correct sentence (sentence_good) and one incorrect sentence (sentence_bad). In a zero-shot setting, the models are evaluated by comparing the probabilities of both sentences in the minimal pair, under the assumption that the models give higher probability to a good sentence and a lower probability to a bad sentence. However, it is important to note that this is not the same as detecting errors or suggesting corrections. The evaluation with the BLiMP dataset measures only the preference of a model for a grammatical sentence over an ungrammatical one. The BLiMP dataset contains many error types, but I focus on three error types: determiners, subject-verb-agreement, and prepositions, whose distribution is shown in Table 3.3. In my thesis experiments, I use the BLiMP dataset to assess whether sensitivity to the three targeted error types is comparable and to investigate how LLMs develop grammaticality assessment. In other words, I use the

[4] Rozema's thesis code is available at: https://github.com/CRozema22/BabyLM-SLA.

BabyLM as a proxy to do an interpretability study of O-RoBERTa.

The metric used for the zero-shot evaluation is accuracy which is used in the pipeline of the BabyLM Challenge for the BLiMP dataset. The zero-shot evaluation is done on the BabyLM, BB-RoBERTa, and O-RoBERTa and provides the score for the grammaticality assessment of the models. This evaluation is done at each of the BabyLM's epochs to study in what order it learns the three targeted grammatical error types.

3.2.3 Fine-Tuning the Models for the GED Task

As it was stated in Section 3.1.3, 60% of the available data is used as the training set. I use this training dataset to fine-tune the models for the sentence-level classification with four possible classes: "G" (for the grammatical sentences), "DET" (for determiner error types), "SVA" (for subject-verb-agreement error types), and "PREP" (for preposition error types). The sentence-level classification avoids ambiguity in pinpointing specific error locations, which often depends on the intended correction. For example, for identifying the location of the subject-verb-agreement error in "Student on Tuesdays like to meet the teacher", the error could be either in "student" or "like", depending on whether the intended meaning is singular or plural. Since the BLiMP dataset does not provide the best correction, but one of the possible grammatical forms of a sentence, I do a sentence-level classification. This approach aligns well with the bidirectional architecture of the models, which are designed to evaluate the acceptability of entire sentences rather than predict specific corrections.

I fine-tune the best BabyLM at its pretraining epochs to study its performance over time. In addition, I fine-tune BB-RoBERTa and O-RoBERTa on the GED task to compare their performance against that of the fine-tuned BabyLM. To choose the number of epochs for fine-tuning the models, I run some experiments using the BabyLM. I limit the upper bound of epochs for fine-tuning to ten because of computational cost, environmental impact, and limitation of time. This approach also allows monitoring for signs of overfitting. In this thesis, overfitting is defined as a performance decline occurring over two consecutive evaluation steps. When this pattern is observed, fine-tuning is stopped, and the epoch preceding the decline is selected for the fine-tuning phase of my thesis experiment. The results of this epoch fine-tuning experiment is shown in Section 4.3. It would have been ideal to also run this experiment for BB-RoBERTa and O-RoBERTa to choose the epoch numbers for their fine-tuning phase, but it was not conducted due to the limitation of time and computational cost.

After fine-tuning the models for the GED classification task, I use the standard classification metrics including precision, recall, and F0.5-score to assess the models' performance. The F0.5-score is conventionally used in the GED literature, since the CoNLL-2014 Shared Task (Ng et al., 2014). The F0.5-score is particularly relevant for GED tasks because it weights precision twice as much as recall, reflecting the greater importance of high precision over recall in detecting grammatical errors in educational applications (Volodina et al., 2023). I use the classificationreport from scikit-learn and modify it for F0.5-score. These metrics are computed for the four classes including three error types (determiners, prepositions, and subject-verb-agreement errors) as well as correct grammatical sentences. This evaluation helps to show how competitive the BabyLM is in comparison with RoBERTa-base, after both models have been fine-tuned.

For the GED classification task, I use the macro average instead of the micro average. They are two approaches to aggregate the classification performance across multiple classes. Macro average computes metrics independently for each class and then averages them. It treats all classes equally regardless of their frequency. In contrast, micro average aggregates the contributions of all classes to compute a global metric, which inherently gives more weight to frequent classes (Sokolova and Lapalme, 2009). Therefore, the macro average score is more appropriate in tasks with class imbalance, because it ensures fair evaluation across all classes. In GED tasks, macro average scores are generally preferred because error types are often imbalanced, with some errors being more frequent than the others (Ng et al., 2014). In my thesis, the macro average score can ensure a fair evaluation of performance across the targeted error types.

Conclusion of the Chapter

The chapter presented the methodologies used to train the BabyLM and the evaluation phases of the thesis. The results of training the BabyLMs is shown in Section 4.1. Moreover, the results of the zero-shot evaluation and the GED classification evaluations of the models can be found in Sections 4.2 and 4.3 in Chapter 4. Then, an error analysis (presented in Chapter 5) is done on some of the misclassifications in the models' GED predictions across the four classes to identify which types of grammatical errors pose the greatest challenge for the models, especially for the BabyLM.

Chapter 4

Results

This chapter presents the results of my thesis. The results of training and evaluating multiple BabyLMs in a zero-shot setting on the BLiMP dataset is shown in Section 4.1. From the trained BabyLMs, the model with the best BLiMP-scores is selected to use for the following experiments. As previously mentioned in Section 3.2.1, the BabyLM is saved at each training stage (from epoch one to ten) for evaluation in both the zero-shot and GED settings. The results of the models' zero-shot evaluations are presented in Section 4.2. Finally, Section 4.3 shows the results of fine-tuning the models for the GED classification task (as described in Section 3.2.3).

4.1 BabyLM Training Results

As described in Section 3.2.1, I train six BabyLMs[1] to ensure their competitiveness in relation to BB-RoBERTa. I evaluate my BabyLMs' performance in a zero-shot setting using the BLiMP dataset to choose the best performing model. BB-RoBERTa's performance is used as a baseline for comparison.

As previously discussed in Section 3.2.1, the BabyLMs have a RoBERTa architecture, but vary in vocabulary size (20K, 40K, and 50K) and hidden layers (four and six). To keep their number of parameter below the defined threshold of 24M, the BabyLMs are trained for five and ten epochs with the hidden size of 256 and eight attention heads.[2] Except these changes, the BabyLMs' other training hyperparameters are kept the same as Rozema's thesis code, which is used as the basis of my BabyLM training setup. The changes in the BabyLMs' configurations are shown in Table 3.4 in Chapter 3. While the performance of the BabyLMs could be further improved by modifying architectures or hyperparameters, this is not the focus of my thesis.

I train and evaluate my BabyLMs on an L4 GPU using Colab Pro. Due to the BabyLMs' different number of parameters, the duration of each models' training per epoch varies between 40-60 minutes. The BLiMP evaluation takes about 15 minutes.

The zero-shot evaluation of the trained BabyLMs and that of BB-RoBERTa are displayed in Table 4.1, Table 4.2, and Table 4.3. Table 4.1 presents the zero-shot evaluation of the BabyLM with the vocabulary size of 20K and either four or six hidden layers, referred to as "BabyLM-V20-L4" and "BabyLM-V20-L6", respectively. Table 4.2 presents the zero-shot evaluation of the BabyLM with the vocabulary size of 40K and

[1] The code for training the BabyLMs is available at: https://github.com/Farnaz-BNF/BabyLM_GED/tree/main/Training_BabyLMs.

[2] The configuration of one of the BabyLMs can be found on my GitHub.

	Vocab-Size=20k				**BB-RoBERTa**
	Layer=4		**Layer=6**		
	Epoch=5	**Epoch=10**	**Epoch=5**	**Epoch=10**	**Epoch=20**
anaphor_agreement	64.62	64.98	69.68	66.31	81.54
argument_structure	61.47	62.43	59.88	59.69	67.12
binding	60.57	61.59	61.95	62.67	67.26
control_raising	57.78	59.77	59.37	59.63	67.85
determiner_noun_agreement	75.83	77.29	75.83	77.14	90.75
ellipsis	45.55	52.66	50.69	57.51	76.44
filler_gap	60.86	60.26	63.01	61.95	63.48
irregular_forms	86.72	84.68	84.58	82.60	87.43
island_effects	39.24	44.51	41.44	44.36	39.87
npi_licensing	56.71	59.52	54.57	51.69	55.92
quantifiers	65.71	67.62	65.89	67.90	70.53
subject_verb_agreement	54.42	55.01	54.60	55.34	65.42
Mean (All BLiMP)	60.79	**62.52**	61.80	62.23	**69.46**
Mean (DET and SVA)	65.12	66.15	65.21	66.24	78.08

Table 4.1: Evaluation Results of BabyLM-V20-L4 and BabyLM-V20-L6 (Vocabulary Size of 20K and 4 and 6 Hidden Layers) on the Filtered BLiMP Dataset in Comparison with that of BB-RoBERTa. The table also shows the overall mean BLiMP scores of all the errors and the mean scores of "SVA" and "DET" error types.

either four or six hidden layers, referred to as "BabyLM-V40-L4" and "BabyLM-V40-L6", respectively. Table 4.3 presents the zero-shot evaluation of the BabyLM with the vocabulary size of 50K and either four or six hidden layers, referred to as "BabyLM-V50-L4" and "BabyLM-V50-L6", respectively. The tables show the BLiMP scores, which is accuracy, for the twelve error classes including anaphor-agreement, argument-structure, and binding in the BLiMP dataset. As previously discussed in Section 3.2.2, the zero-shot evaluation pipeline of the BabyLM Challenge uses the BLiMP dataset. The table also reports the overall mean BLiMP-scores of all the errors and the mean scores of subject-verb-agreement (SVA) and determiner-noun-agreement (DET), which are two of the three targeted error types in this thesis. These averages are used to compare the six trained BabyLMs and to select the best BabyLM for the subsequent experiments. Based on the results, the BabyLMs generally show higher performance when they are trained for ten epochs.

As shown in the results presented in Table 4.1, Table 4.2, and Table 4.3, the BabyLM with a 20K vocabulary size, four hidden layers, and ten training epochs (BabyLM-V20-L4), and the BabyLM with a 40K vocabulary size, four hidden layers, and five training epochs (BabyLM-V40-L4), achieve the highest performance among the BabyLMs, with overall mean BLiMP scores of 62.52% and 63.18%, respectively. Although their performances are the closet to BB-RoBERTa (69.46%), they still show a slightly lower performance. This performance gap may be attributed to the chosen training configurations of the BabyLMs. As previously stated, optimizing the BabyLM architecture is not the primary focus of this thesis. Future research may explore more effective training designs.

To ensure BabyLM-V20-L4's and BabyLM-V40-L4's performance differences are not because of random initialization, I train and evaluate these two BabyLMs across five extra random seeds, in addition to the default seed, for a more reliable comparison. The default seed in TrainingArguments[3] is set to 42, which is used as the main seed in

[3] The link of TrainingArguments class, which offers a wide range of options to customize how a

	Vocab-Size=40k				**BB-RoBERTa**
	Layer=4		**Layer=6**		
	Epoch=5	**Epoch=10**	**Epoch=5**	**Epoch=10**	**Epoch=20**
anaphor_agreement	69.43	69.38	66.72	62.12	81.54
argument_structure	62.48	63.54	61.93	62.38	67.12
binding	61.46	62.72	64.14	62.94	67.26
control_raising	60.65	60.30	61.20	61.27	67.85
determiner_noun_agreement	75.36	74.87	76.62	76.32	90.75
ellipsis	46.13	52.02	46.82	49.19	76.44
filler_gap	62.14	61.55	63.62	63.54	63.48
irregular_forms	88.70	84.07	83.72	79.24	87.43
island_effects	39.65	43.24	38.75	40.88	39.87
npi_licensing	46.05	48.79	48.33	47.98	55.92
quantifiers	75.04	71.92	73.80	73.96	70.53
subject_verb_agreement	54.07	55.18	54.72	57.27	65.42
Mean (All BLiMP)	**63.18**	62.30	61.70	61.42	**69.46**
Mean (DET and SVA)	64.71	65.02	65.67	66.79	78.08

Table 4.2: Evaluation Results of BabyLM-V40-L4 and BabyLM-V40-L6 (Vocabulary Size of 40K and 4 and 6 Hidden Layers) on the Filtered BLiMP Dataset in Comparison with that of BB-RoBERTa. The table also shows the overall mean BLiMP scores of all the errors and the mean scores of "SVA" and "DET" error types.

	Vocab-Size=50k				**BB-RoBERTa**
	Layer=4		**Layer=6**		
	Epoch=5	**Epoch=10**	**Epoch=5**	**Epoch=10**	**Epoch=20**
anaphor_agreement	72.39	69.99	70.91	63.55	81.54
argument_structure	62.75	62.57	63.07	62.91	67.12
binding	63.28	61.92	61.83	62.10	67.26
control_raising	62.64	62.44	61.18	61.60	67.85
determiner_noun_agreement	74.24	74.25	76.52	76.50	90.75
ellipsis	47.58	52.71	45.73	53.70	76.44
filler_gap	62.54	61.83	62.08	63.38	63.48
irregular_forms	83.21	82.44	85.04	80.76	87.43
island_effects	41.93	39.24	42.56	44.06	39.87
npi_licensing	50.00	43.99	51.18	55.04	55.92
quantifiers	71.05	70.66	63.68	62.83	70.53
subject_verb_agreement	53.80	55.23	56.62	56.60	65.42
Mean (All BLiMP)	62.12	61.44	61.70	61.91	**69.46**
Mean (DET and SVA)	64.02	64.74	66.57	66.55	78.08

Table 4.3: Evaluation Results of BabyLM-V50-L4 and BabyLM-V50-L6 (Vocabulary Size of 50K and 4 and 6 Hidden Layers) on the Filtered BLiMP Dataset in Comparison with that of BB-RoBERTa. The table also shows the overall mean BLiMP scores of all the errors and the mean scores of "SVA" and "DET" error types.

this thesis. Table 4.5 and Table 4.4 show the average performance of the five random seeds (56, 162, 354, 670, 1278) throughout the BLiMP dataset and two of my targeted error types (determiner and subject-verb-agreement) in BabyLM-V20-L4 (trained for ten epochs) and BabyLM-V40-L4 (trained for five epochs). To better compare the average performance of the multiple random seeds and that of the default seed (42), Table 4.5 and Table 4.4 also include the results of the default seed (42). The "Seed Average (Total)" refers to the mean of the overall BLiMP scores—calculated across all

model is trained, is available at: https://huggingface.co/docs/transformers/en/main_classes/trainer error types—for the five different random seeds. For example, in Table 4.4, this value (62.00%) represents the average of the following scores for BabyLM-V20-L4: 60.12% (random seed 56), 61.75% (random seed 162), 63.24% (random seed 354), 61.25% (random seed 670), and 63.61% (random seed 1278). On the other hand, "Seed Average (DET and SVA)" represents the average of the combined "DET" and "SVA" scores across all five random seeds. These two error types are among the targeted errors of this thesis, making this average particularly relevant for the comparison. As the table shows, "Seed Average (DET and SVA)" for BabyLM-V20-L4 is 64.60%.

	BabyLM-V20-L4, Epoch = 10						**BB-RoBERTa**
	Seed=42	**Seed=56**	**Seed=162**	**Seed=354**	**Seed=670**	**Seed=1278**	**Epoch=20**
anaphor_agreement	64.98	66.16	62.78	61.55	68.35	65.29	81.54
argument_structure	62.43	60.67	59.53	61.14	61.30	60.92	67.12
binding	61.59	60.27	60.92	64.38	62.50	63.73	67.26
control_raising	59.77	56.94	59.66	59.30	60.08	60.83	67.85
determiner_noun_agreement	77.29	79.16	74.54	77.57	74.85	78.22	90.75
ellipsis	52.66	50.87	56.47	53.75	51.10	50.17	76.44
filler_gap	60.26	60.82	63.60	65.22	60.36	63.73	63.48
irregular_forms	84.68	78.02	81.93	83.41	80.10	82.95	87.43
island_effects	44.51	41.70	46.15	44.28	40.36	42.26	39.87
npi_licensing	59.52	46.46	50.15	54.11	58.94	62.25	55.92
quantifiers	67.62	64.68	71.35	79.16	60.54	74.57	70.53
subject_verb_agreement	55.01	55.68	53.86	55.00	56.57	58.43	65.42
Mean (All BLiMP)	62.52	60.12	61.75	63.24	61.25	63.61	**69.46**
Mean (DET and SVA)	66.15	67.42	64.20	66.29	65.71	68.33	**78.08**
Seed Average (Total)	-			**62.00**			-
Seed Average (DET and SVA)	-			**64.60**			-

Table 4.4: The BLiMP-Score Performance Across Multiple Random Seeds for BabyLM-V20-L4 in Comparison with BB-RoBERTa. The table also reports the overall mean BLiMP scores across all error types, as well as the mean scores for the "SVA" and "DET" error types across five random seeds. It also includes the BLiMP score for the default seed (42) and its averages, facilitating comparison with the multi-seed averages.

	BabyLM-V40-L4, Epoch = 5						**BB-RoBERTa**
	Seed=42	**Seed=56**	**Seed=162**	**Seed=354**	**Seed=670**	**Seed=1278**	**Epoch=20**
anaphor_agreement	69.43	69.84	69.02	62.63	77.75	62.68	81.54
argument_structure	62.48	62.60	62.80	61.83	62.84	61.92	67.12
binding	61.46	63.43	63.42	63.92	61.53	63.24	67.26
control_raising	60.65	60.05	61.31	59.94	61.31	61.07	67.85
determiner_noun_agreement	75.36	74.65	74.53	75.54	76.07	74.09	90.75
ellipsis	46.13	49.83	45.67	45.21	49.65	45.50	76.44
filler_gap	62.14	63.48	62.96	65.53	61.75	62.76	63.48
irregular_forms	88.70	85.55	82.54	72.82	84.68	69.97	87.43
island_effects	39.65	44.77	42.86	41.78	44.06	39.57	39.87
npi_licensing	46.05	55.30	57.01	50.94	56.35	53.81	55.92
quantifiers	75.04	70.17	72.020	63.60	72.75	62.73	70.53
subject_verb_agreement	54.07	54.54	54.54	53.19	52.18	54.47	65.42
Mean (All BLiMP)	63.18	62.85	62.39	59.74	63.41	59.32	**69.46**
Mean (DET and SVA)	64.71	64.60	64.54	64.37	64.13	64.28	**78.08**
Seed Average (Total)	-			**61.54**			-
Seed Average (DET and SVA)	-			**64.38**			-

Table 4.5: The BLiMP-Score Performance Across Multiple Random Seeds for BabyLM-V40-L4 in Comparison with BB-RoBERTa. The table also reports the overall mean BLiMP scores across all error types, as well as the mean scores for the "SVA" and "DET" error types across five random seeds. It also includes the BLiMP score for the default seed (42) and its averages, facilitating comparison with the multi-seed averages.

As shown in both Table 4.5 and Table 4.4, the average BLiMP scores across all error types over five random seeds for BabyLM-V20-L4 is 62.00%, which is close to the model's BLiMP score using the default seed (62.52%). Although the "Seed Average (Total)" and "Seed Average (DET and SVA)" scores are slightly close between BabyLM-V20-L4 (62.00% and 64.60%, respectively) and BabyLM-V40-L4 (61.54% and 64.38%, respectively), the BabyLM-V20-L4 model trained for ten epochs achieves slightly higher performance in both averages. Based on these results, I use BabyLM-V20-L4 for the following thesis experiments. This BabyLM also provides a better study of the BabyLM's grammaticality assessment throughout its training epochs. I call "BabyLM-V20-L4" as "the BabyLM" or "my BabyLM" in the remaining parts of the thesis.

4.2 Zero-Shot Evaluation

As previously discussed in Section 3.2.2, the grammaticality assessments of the models are evaluated in a zero-shot setting on the filtered BLiMP dataset, along with the BLiMP-style preposition data, which I created for this thesis from the preposition data of the BEA-2019 dataset (as described in Section 3.1.3). Although the zero-shot evaluation provides BLiMP-scores, which is accuracy, on 13 error types, I focus on determiners, prepositions, and subject-verb-agreement errors. The BabyLM's zero-shot evaluation throughout its ten epochs is shown in Table A.4 in Appendix A. A summary including the zero-shot results for the BabyLM's epochs one, three, five, eight, nine, and ten are shown in Table 4.6. This selection is sufficient to study of the model at the beginning, middle, and late stages of its training. Since the performance increases the most at the BabyLM's epochs eight and nine, they are also included. Moreover, the zero-shot evaluation is also done on O-RoBERTa[4], whose performance is included in the table. Although the BabyLM's zero-shot performance is primarily compared to that of O-RoBERTa, it is also compared with that of BB-RoBERTa. Therefore, a zero-shot evaluation is done on BB-RoBERTa[5] to compare the performance of my BabyLM (13M parameters) with BB-RoBERTa—a baseline RoBERTa BabyLM from the first BabyLM Challenge, which has 125M number of parameters and was trained for 20 epochs.

	The BabyLM						**BB-RoBERTa**	**O-RoBERTa**
Category	**E=1**	**E=3**	**E=5**	**E=8**	**E=9**	**E=10**		
anaphor_agreement	44.63	60.48	63.85	67.59	66.56	64.98	81.54	90.70
argument_structure	60.81	60.18	61.60	62.00	62.08	62.43	77.12	83.04
binding	61.29	60.15	60.79	61.07	61.25	61.59	67.26	79.18
control_raising	58.44	58.24	58.59	59.32	59.99	59.77	67.85	81.95
determiner_noun_agreement	51.99	70.95	75.93	**78.37**	78.16	77.29	**90.75**	**97.28**
ellipsis	23.61	42.21	48.04	51.50	52.08	52.66	76.44	92.15
filler_gap	62.81	60.83	60.96	60.04	59.77	60.26	63.48	89.39
irregular_forms	61.27	85.80	88.09	86.11	85.29	84.68	87.43	95.67
island_effects	49.63	43.31	41.59	43.98	43.80	44.51	39.87	79.71
npi_licensing	42.03	47.07	56.98	58.43	58.96	59.52	55.92	82.61
quantifiers	44.28	61.39	67.34	69.78	68.26	67.62	70.53	70.79
subject_verb_agreement	50.46	52.72	53.93	**55.52**	**55.52**	55.01	**65.42**	**91.47**
preposition	49.75	57.85	60.63	61.65	**62.24**	61.99	**73.08**	**91.28**
Mean (All BLiMP)	50.85	58.55	61.41	**62.72**	62.61	62.49	**73.59**	**86.55**
Mean ("DET", "SVA", PREP)	50.73	60.51	63.50	65.18	**65.30**	64.76	**76.42**	**93.34**

Table 4.6: Zero-Shot Evaluation of the BabyLM's Epochs One, Three, Five, Eight, Nine, and Ten in Comparison with those of BB-RoBERTa and O-RoBERTa. "E" represents the BabyLM's epoch number.

[4] O-RoBERTa model is available in https://huggingface.co/FacebookAI/roberta-base.

[5] BB-RoBERTa model is available in https://huggingface.co/babylm/roberta-base-strict-small-2023.

According to Table 4.6, among the three targeted error types, the class of determiner shows the highest performance. The BabyLM's performance for determiner errors peaks at epoch eight (78.37%) before plateauing. Similarly, BB-RoBERTa (90.75%) and O-RoBERTa (97.28%) achieve higher zero-shot evaluation score for this error type. BB-RoBERTa's higher score may be related to its higher number of epochs, which is twice the epoch number of the BabyLM (with ten epochs), and its higher number of parameters (about ten times more than the BabyLM). The configurations of the selected BabyLM (BabyLM-V20-L4) and BB-RoBERTa can be seen in Table 3.4 in Chapter 3, which presents the configurations of six trained BabyLMs for this thesis (explained in Section 3.2.1). On the other hand, O-RoBERTa is trained with 160GB uncompressed text, and 125M parameters (Liu et al., 2019). For the preposition category, the BabyLM shows steady gains and peaks at the ninth epoch (62.24%). While BB-RoBERTa (73.08%) performs better than the BabyLM for this error type, it falls behind O-RoBERTa (91.28%) for 18.20%. The BabyLM's performance in the subject-verb-agreement class falls behind the other two error types with more gradual improvements and an earlier plateau. When detecting this error type, the BabyLM's performance peaks at the eighth epoch (55.52%), increasing only 5.06% from the first epoch. O-RoBERTa still detects this error type with a high score (91.47%), which is very close to its score for the preposition class (91.28%). Similar to the BabyLM (55.52%), BB-RoBERTa (65.42%) also seems to have difficulty in detecting the subject-verb-agreement class.

I focus on the epochs one, three, five, eight, nine, and ten of the BabyLM because the distinct phases in the BabyLM's development can be observed there. At the first epoch, the BabyLM exhibits its initial zero-shot performance with scores of 51.99% for determiner errors, 50.46% for subject-verb-agreement, and 49.75% for preposition errors. This reflects the BabyLM's early learning stage, where it has had minimal exposure to the training dataset. It seems that these scores are random scores, since the zero-shot evaluation is a binary choice. The random scores (close to 50%) is expected in a binary classification task when the model has not yet meaningfully learned from the data. In other words, scores around 50% show that the model has learned almost nothing. The same is true for the scores across other error types, except for ellipsis errors with the score of 23.61%.[6] Moreover, while the score for the preposition errors is lower than that of subject-verb-agreement at this training stage, it moves ahead from the second epoch (shown in Table A.4).

By the third epoch, the BabyLM shows an increase of performance, particularly in the determiner category where the performance increases for $\Delta = +18.96\%$. This improvement suggests that the model learns the determiner errors better than the other two error types. In contrast, the performance for the preposition ($\Delta = +2.26\%$) and subject-verb-agreement ($\Delta = +8.1\%$) are modest, indicating that the BabyLM may require more training exposure to learn these grammatical error types. At epoch five, the BabyLM achieves 75.93% for the determiner category, while its score for the preposition errors increases to 60.63%, indicating continued improvement in its grammaticality judgment. On the other hand, its progress for the subject-verb-agreement errors is small, rising for $\Delta = +1.21\%$. These results indicate that the subject-verb-agreement

[6] Ellipsis refers to grammatical constructions where part of a sentence is omitted because it is inferable from context. For example, “John can play the guitar, and Mary can too” omits the repeated verb phrase, unlike the ungrammatical form “John can play the guitar, and Mary can play the guitar too.”

error type is the most challenging among the three error types, possibly because of its reliance on longer-range dependencies and more abstract syntactic features.

The BabyLM peaks its performance at epochs eight and nine. Epoch eight marks the highest performance for the determiner (78.37%) and subject-verb-agreement (55.52%) error types, while the BabyLM shows its highest score for the preposition errors at epoch nine (62.24%). At epochs nine and ten, the BabyLM's performance remains relatively stable. Its score for the determiner error types drops slightly to 78.16% and then 77.29% ($\Delta = -0.87\%$), while its subject-verb-agreement score holds at 55.52% before decreasing slightly to 55.01% ($\Delta = -0.51\%$). The score for the preposition errors also drops slightly in epoch ten ($\Delta = -0.25\%$). Overall, Table 4.6 shows that the BabyLM's performance peaks in grammaticality assessment at the eighth epoch. The following epochs show marginal fluctuations rather than a continued improvement, which could indicate a plateauing effect in the learning process.

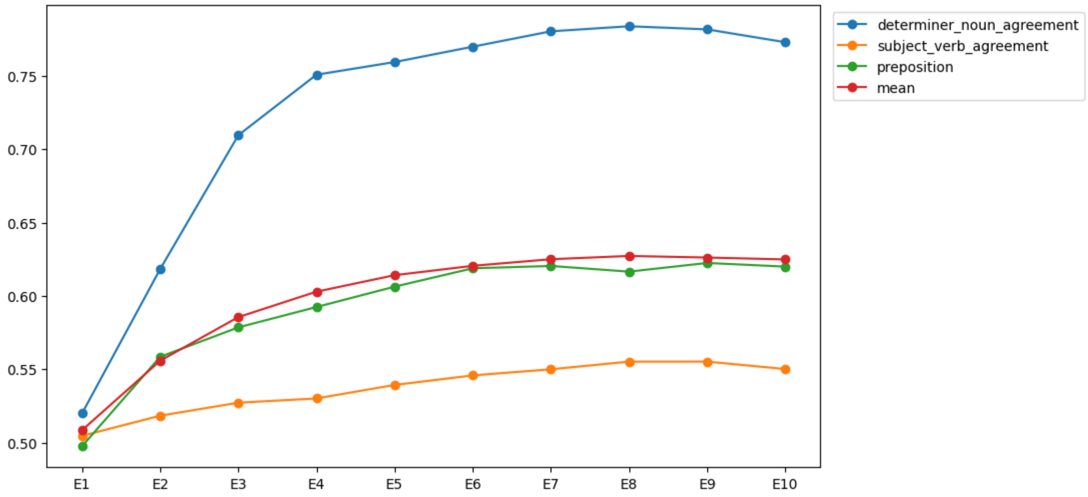

Figure 4.1: Zero-Shot Evaluation of The BabyLM Throughout Its Training Stages (Epochs One to Ten). The figure displays the results of Table 4.6 for determiner, preposition, and subject-verb-agreement errors. It also shows the BabyLM's overall average performance for the three targeted error types across its ten epochs.

The development of the BabyLM's grammaticality assessment throughout its ten epochs, as discussed in previous paragraphs in relation to the results in Table 4.6, is displayed in Figure 4.1. The figure presents the zero-shot evaluation results of the BabyLM across its ten epochs, focusing on the three targeted grammatical error types: determiners, subject-verb-agreement, and prepositions. In addition, the mean performance across these three error types (shown with the red line in Figure 4.1) provides a holistic view of the model's grammaticality assessment. Overall, the figure shows an increasing trajectory in the BabyLM's performance during the initial training stages (epochs one through four) for the three targeted error types.

Among the three targeted error types, displayed in Table 4.6, the BabyLM shows the highest zero-shot performance for the determiner errors. Similarly, BB-RoBERTa and O-RoBERTa also show their highest scores not only among these three error types, but also among all the error categories. Based on these results, it seems that determiners are easier to learn for LMs. For the preposition and subject-verb-agreement errors, while O-RoBERTa's scores are close, BB-RoBERTa's scores differ about 7.66%. In the

BabyLM's performance, the score difference between these two error types is 6.98% at its last training stage, which is close to BB-RoBERTa's scores difference.

4.3 GED Evaluation

As previously discussed in Section 3.2.3, I use 60% of the GED classification data (as explained in Section 3.1.3) to fine-tune my models for the sentence-level classification with four classes: "G" (for the grammatical sentences), "DET" (for determiner error types), "SVA" (for subject-verb-agreement error types), and "PREP" (for preposition error types). To examine the performance progression of the BabyLM over time, similar to what I have done with the zero-shot evaluation, I fine-tune the BabyLM across its training stages, which are represented as "epoch". Although the BabyLM is fine-tuned throughout its ten epochs, some of them (epochs one, three, five, nine, and ten) are selected for detailed analysis.[7] This selection helps to study the development of fine-tuning the BabyLM's epochs on the GED task. The ninth epoch is included in the analysis because the BabyLM shows its highest performance for the average scores of the targeted three error types in the zero-shot evaluation (reported in Table 4.6). In addition, I fine-tune BB-RoBERTa and O-RoBERTA to compare their performance against the BabyLM's fine-tuned epochs.

Research shows that the amount of data and the number of epochs used during fine-tuning has a big impact on the performance. As such, I start with an additional experiment to see the impact of the training data and choose the best number of epochs to use during the fine-tuning phase of this thesis. To explore the impact of pretraining progress on fine-tuning performance, I select two checkpoints from my BabyLM training: one from the initial stage (epoch one) and one from the final stage (epoch ten). I fine-tune each of these for up to ten epochs, using the development set to evaluate their performance. These two points are not necessarily the ideal candidates for fine-tuning, but they serve to contrast early and late representations learned during pretraining. As previously discussed in Section 3.2.3 in Chapter 3, the upper bound for the number of epochs is set to ten because of the computational cost, environmental considerations, and time limitations. This approach also serves to monitor overfitting. In the context of my experiment, I define overfitting as a decline in performance across two consecutive epochs. When this occurs, the most recent stable epoch is selected as the final epoch for the fine-tuning phase of my thesis.

To choose the number of epochs for fine-tuning, I use a fixed seed (42) to control the random initialization effect. To assess how much training data is necessary to reach performance saturation, I conduct two sets of experiments: one using the full training set of the GED classification dataset (referred to as "100%" in Table 4.7), and one using only half of the training set (referred to as "50%" in the table). This experimental design is motivated by concerns that, in some scenarios, particularly low-resource settings, using too much fine-tuning data can lead to memorization rather than generalization (de Vries et al., 2022). This concern was especially relevant in the setup experiments of this thesis, given that the task involves only four classes, raising the possibility that performance could saturate quickly. Since the complexity of the task for the BabyLM was not known in advance, it was important to empirically assess whether full supervision was necessary. The results, shown in Table 4.7, suggest that

[7] The classification reports and confusion matrices of fine-tuning the BabyLM across its remaining pretraining epochs (two, four, six, seven, and eight) are shown in Appendix B.2.

	BabyLM at Epoch 1		**BabyLM at Epoch 10**	
Train Data	**50%**	**100%**	**50%**	**100%**
1 Epoch	0.59	0.59	0.71	0.74
2 Epochs	0.66	0.80	0.82	0.89
3 Epochs	0.69	0.87	0.81	0.91
4 Epochs	0.76	0.88	0.84	0.89
5 Epochs	0.75	0.86	0.84	**0.90**
6 Epochs	0.79	**0.88**	0.78	0.89
7 Epochs	**0.80**	0.87	**0.85**	0.88
8 Epochs	0.76	0.83	0.82	0.84
9 Epochs	0.79	0.83	0.82	0.85
10 Epochs	0.78	0.84	0.82	0.84

Table 4.7: The Results of the Epoch Experiment to Choose the Number of Epochs for Fine-Tuning the Models. The table shows the macro-average (F0.5-score) on the DEV dataset for the BabyLM pretrained for one and ten epochs across 50% and 100% of the training set of the GED classification dataset (presented in Section 3.1.3).

100% training setup provides an adequate amount of data for effective generalization without signs of early performance saturation. Moreover, as it can be seen in Table 4.7, training with only half of the training set of the GED evaluation data requires more training steps to reach similar performance with that of the whole training set, making it a less efficient setup. In contrast, the results of the full dataset suggests that the model converges slightly more quickly and effectively. This not only makes the 100% setup more practical but also environmentally favorable, as fewer training steps mean lower energy consumption and reduced CO_2 emissions.

As shown in Table 4.7, when fine-tuned on the whole training set, the BabyLM pretrained for one epoch reaches its peak performance after six epochs (88.00%) on the GED task. On the other hand, the BabyLM pretrained for ten epochs achieves its highest performance (90.00%) after five epochs. The performance of the BabyLM pretrained for one and ten epochs both decline after being fine-tuned for five and six epochs on the GED task. It indicates overfitting, based on the decline in performance for two consecutive epochs that I defined for my experiment. Since the BabyLM pretrained for ten epochs is the last stage of my BabyLM and its performance acts like a proxy, I use five epochs to fine-tune the models. It would have been ideal to run the same fine-tuning epoch number experiment for both O-RoBERTa and BB-RoBERTa, but it was not done due to the limitations of time and computational costs and, rather, the same five epoch number is used to fine-tune them. Future research can further explore this experiment.

I use simpletransformers[8] to fine-tune my models with the training set of the GED classificatio dataset for five epochs. Since the default manual-seed is set to "None" in simpletransformers, a random number (464) is generated and used as the main seed for the fine-tuning phase. To enable a more reliable comparison of the fine-tuned models' performance on the GED task, I use a fixed seed (464). Moreover, using a fixed seed allows comparability across fine-tuning the BabyLM's epochs. This also supports the replicability of my thesis experiments in future research.

I fine-tune the BabyLM, BB-RoBERTa, and O-RoBERTa for five epochs on the

[8] simpletransformers is available in https://simpletransformers.ai.

	precision	recall	F0.5-score
BabyLM at Epoch 1	0.831	0.870	0.834
BabyLM at Epoch 2	0.848	0.878	0.851
BabyLM at Epoch 3	0.855	0.885	0.858
BabyLM at Epoch 4	0.855	0.888	0.858
BabyLM at Epoch 5	0.857	0.889	0.860
BabyLM at Epoch 6	0.857	0.888	0.860
BabyLM at Epoch 7	0.856	0.888	0.860
BabyLM at Epoch 8	0.857	0.888	0.859
BabyLM at Epoch 9	**0.858**	**0.890**	**0.861**
BabyLM at Epoch 10	0.855	0.885	0.858
BB-RoBERTa	**0.859**	**0.883**	**0.862**
O-RoBERTa	**0.905**	**0.887**	**0.901**

Table 4.8: Macro Average Scores Across the Fine-Tuned Models for the GED classification task of Four Classes (described in Section 3.2.3). All BabyLM stages, BB-RoBERTa, and O-RoBERTa are fine-tuned for five epochs.

GED task (described in Section 3.2.3). Table 4.8 shows the macro average scores of the fine-tuned models. The BabyLM is fine-tuned across its training stages, which are represented as "epoch" in the table. For example, "BabyLM at Epoch 1" refers to the BabyLM fine-tuned at its first pretraining epoch. Research shows that random seeds can affect LLMs' performance (Dodge et al., 2020). To assess the robustness of the results of Table 4.8 to random initialization, I conduct an additional experiment using five extra random seeds (20, 58, 150, 342, 613). In consideration of computational and environmental costs, these experiments are limited to fine-tuning the BabyLM at three pretraining stages. I fine-tune the BabyLM at three pretraining checkpoints—epoch one, five, and ten—which correspond to the BabyLM's early, middle, and last stages of training. These checkpoints allow me to observe how the model's performance evolves with increased exposure to the training data. The macro-average F0.5-scores for these epochs using the main seed are 83.40% (epoch 1), 86.00% (epoch 5), and 85.80% (epoch 10), as shown in the column "main seed = 464" in Table 4.9. The mean macro average F0.5-scores across these seeds are 83.30% (epoch 1), 86.30% (epoch 5), and 85.40% (epoch 10), as shown in Table 4.9.[9] These results are closely aligned with those of the main seed, suggesting that the selected seed produces representative performance. Although the macro-average F0.5-scores of the fune-tuned BabyLM pretrained for one and ten epochs (83.40% and 85.80% respectively) are slightly higher with the main seed than the mean macro-average F0.5-scores across these seeds (83.30% and 85.40% respectively), the comparison indicates that the seed is not performing either too high or too low and can be used for my experiments. Therefore, I use the same seed (464) to fine-tune all the models. Ideally, same experiment would have been conducted for both O-RoBERTa and BB-RoBERTa. However, this was not done due to limitation of time and computational and environmental costs. Future research can further explore this experiment.

According to Table 4.8, even the BabyLM fine-tuned on the GED task at epoch one achieves relatively high performance, which indicates the benefits of pre-trained LMs.

[9] The classification reports and confusion matrices of the five extra random seeds experiment are shown in Appendix B.

	Seed=20	Seed=58	Seed=150	Seed=342	Seed=613	Mean	Main Seed=464
BabyLM at Epoch 1	0.833	0.832	0.835	0.844	0.820	**0.833**	**0.834**
BabyLM at Epoch 5	0.863	0.866	0.858	0.861	0.866	**0.863**	**0.860**
BabyLM at Epoch 10	0.851	0.855	0.852	0.856	0.856	**0.854**	**0.858**

Table 4.9: Mean of the Macro Avg F0.5-Scores of the Three Fine-Tuned BabyLM's Training Stages (Epochs One, Five, and Ten) Across Multiple Random Seeds. The macro avg F0.5-scores of the selected seed (464) is shown for better comparison.

The table shows that the BabyLM's performance on the GED task improves consistently when fine-tuned at successive stages of its pretraining, particularly from epoch one to five. As previously discussed in Section 3.2.3, F0.5-score is often used in educational applications and GED tasks because it weights precision twice as much as recall. Fine-tuning the BabyLM at different stages of pretraining (epochs one through five) shows substantial gains across all three metrics of precision, recall, and F0.5-score. After fine-tuning the BabyLM pretrained for five epochs on the GED task, the performance seems to plateau, with only minor improvements or fluctuations. The highest performance belongs to the BabyLM fine-tuned at epoch nine, with a precision of 0.858, recall of 0.890, and F0.5-score of 0.861. It is interesting to note that the performance of the BabyLM fine-tuned at epoch ten is the same as that of the performance of the BabyLM fine-tuned at epoch three. Moreover, as Table 4.8 shows, the difference between the F0.5 macro scores of the performance of the BabyLM fine-tuned at epoch five and ten is quite quite small ($\Delta = 0.2\%$). Similarly, according to Table 4.9, this difference is also small ($\Delta = 0.9\%$) for the mean of the multiple seeds experiment. These results suggest that, on average, little performance gain is achieved by fine-tuning the BabyLM beyond epoch five.

	DET			**PREP**			SVA			**G**		
	Precision	**Recall**	**F0.5-Score**	**Precision**	**Recall**	**F0.5-Score**	**Precision**	**Recall**	**F0.5-Score**	**Precision**	**Recall**	**F0.5-Score**
BabyLM at Epoch 1	0.983	0.992	0.985	0.561	0.832	0.600	0.879	0.864	0.876	0.901	0.793	0.877
BabyLM at Epoch 2	0.988	1.000	0.990	0.556	0.778	0.590	0.940	0.912	0.935	0.908	0.821	0.889
BabyLM at Epoch 3	0.991	1.000	0.993	0.553	0.783	0.587	0.962	0.931	0.956	0.916	0.826	0.896
BabyLM at Epoch 4	0.991	1.000	0.993	0.553	0.782	0.587	0.955	0.948	0.954	0.922	0.824	0.900
BabyLM at Epoch 5	0.993	1.000	0.995	0.556	0.785	0.591	0.956	0.947	0.954	0.921	0.826	0.900
BabyLM at Epoch 6	0.995	1.000	0.996	0.553	0.776	0.587	0.960	0.948	0.957	0.920	0.828	0.900
BabyLM at Epoch 7	0.994	1.000	0.995	0.549	0.776	0.583	0.962	0.950	0.960	0.921	0.826	0.900
BabyLM at Epoch 8	0.995	1.000	0.996	0.551	0.793	0.586	0.960	0.937	0.955	0.920	0.823	0.899
BabyLM at Epoch 9	0.993	1.000	0.994	0.557	0.787	0.591	0.960	0.946	0.957	0.922	0.827	0.901
BabyLM at Epoch 10	0.995	1.000	0.996	0.548	0.782	0.583	0.960	0.936	0.955	0.917	0.824	0.896
BB-RoBERTa	0.999	1.000	0.999	0.538	0.702	0.565	0.985	0.986	0.985	0.915	0.845	0.900
O-RoBERTa	0.999	1.000	0.999	0.722	0.617	0.698	0.995	0.992	0.995	0.905	0.939	0.911

Table 4.10: Precision, Recall, and F0.5 Scores for the Four Defined Error Classes Across the Fine-Tuned Models on the GED Classification Task (described in Section 3.2.3). The BabyLM, BB-RoBERTa, and O-RoBERTa are fine-tuned for five epochs on the GED task. The BabyLM is fine-tuned across its training stages.

Table 4.10 presents the fine-tuned models' precision, recall, and F0.5-scores for the GED classification task across four classes. As noted earlier, the fine-tuned BabyLM shows performance gains in the early epochs, especially between epochs one and five. The "DET" scores remain consistently high across all fine-tuned models, while the "PREP" class is the most challenging. The fine-tuned O-RoBERTa performs well on the "G" and "SVA" classes but underperforms on the "PREP" class. Overall, the fine-tuned BabyLM achieves competitive results across all four classes. A broader analysis of the selected BabyLM's fine-tuned epochs, along with the fine-tuned BB-RoBERTa and the fine-tuned O-RoBERTa, is provided in the following paragraphs.

GED Evaluation of the BabyLM Pretrained for One Epoch:

Table 4.11 presents the classification report for the BabyLM fine-tuned at epoch one. It shows the BabyLM's performance fine-tuned at its first stage of learning, with clear disparities in performance across error categories. The model shows high precision and recall for "DET" and "SVA" error types, achieving F0.5-scores of 98.50% and 87.60%, respectively. This is further supported by its confusion matrix (shown in Figure 4.2), where "DET" predictions are nearly perfect (1496 correct, 12 misclassifications). Similarly, "SVA" class is well predicted (957 correct predictions), but with more mislabels (150 instances). However, performance is low for the "G" class and much lower for the preposition error types. The "G" class has a notable number of misclassifications, especially with 572 sentences mislabeled as "PREP" and 131 as "SVA", leading to a relatively low recall of 79.30% and an F0.5-score of 87.70%. The "PREP" class is especially challenging, with many true preposition errors (147 instances) being misclassified as grammatical and only 731 correctly identified. This results in high recall (83.20%) but low precision (56.10%) for the "PREP" class, with an F0.5-score of 60.00%.

	precision	**recall**	**F0.5-score**	**support**
DET	0.983	0.992	0.985	1508
G	0.901	0.793	0.877	3494
PREP	0.561	0.832	0.600	879
SVA	0.879	0.864	0.876	1107
accuracy			0.852	6988
macro avg	0.831	0.870	0.834	6988
weighted avg	0.872	0.852	0.865	6988

Table 4.11: The Classification Report of the Results for Fine-tuning the BabyLM Pretrained for One Epoch on the GED Classification Task.

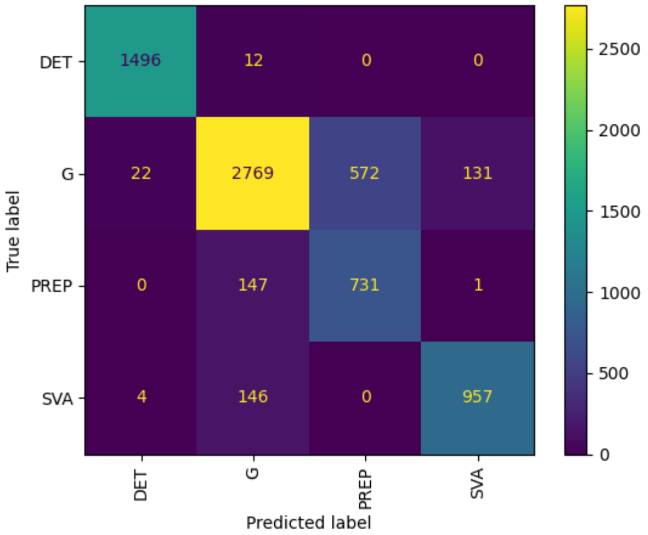

Figure 4.2: The Confusion Matrix of Fine-tuning the BabyLM Pretrained for One Epoch on the GED Classification Task.

Overall, the BabyLM fine-tuned at epoch one, shows a good performance in detecting “DET” and “SVA” error types, but it struggles with classifying “G” and “PREP” classes. Although the macro average F0.5-score of 83.40% indicates a moderately effective model, it still requires refinement. In other words, the results suggest that more training is needed to improve the class balance, especially in recognizing grammatical sentences and preposition error types.

GED Evaluation of the BabyLM Pretrained for Three Epochs:

According to Table 4.12, the fine-tuning at epoch three of the BabyLM shows improvement in overall performance. The “DET” class has a F0.5-score of 99.30%, achieving 100% recall and zero false negatives. The “SVA” class prediction also improves, with a higher F0.5-score of 95.60% (up from 87.60% at epoch one), supported by high precision (96.20%) and improved recall (93.10%), which is evident in its confusion matrix (displayed in Figure 4.3) with 1031 correct classifications and only minimal missclassifications into the “G” class (40 instances). The “G” class shows improved performance, especially in precision (91.60%). The recall of the “G” class remains slightly lower (82.60%) due to confusion with “PREP” (557 instances) and “SVA” (40 instances). The “PREP” class continues to struggle, with a modest F0.5-score of 58.70%, largely because of frequent misclassification as “G” (190 instances). The recall of the “PREP” class has decreased to 78.30%.

	precision	**recall**	**F0.5-score**	**support**
DET	0.991	1.000	0.993	1508
G	0.916	0.826	0.896	3494
PREP	0.553	0.783	0.587	879
SVA	0.962	0.931	0.956	1107
accuracy			0.875	6988
macro avg	0.855	0.885	0.858	6988
weighted avg	0.894	0.875	0.888	6988

Table 4.12: The Classification Report of the Results for Fine-tuning the BabyLM Pretrained for Three Epochs on the GED Classification Task.

As Figure 4.3 shows, the reduction in misclassifying “SVA” as the “G” class and “G” as the “SVA” class is a notable improvement over the performance of the BabyLM fine-tuned at epoch one. The “PREP” class remains challenging with 190 misclassifications into “G” (close to the 147 in the performance of the BabyLM fine-tuned at epoch one). The model captures lower true positives of “PREP” class (688 instances), in comparison with 731 correct labels of the performance of the BabyLM fine-tuned at epoch one. The “G” class is frequently mistaken for the class of “PREP”. However, improvements in correctly identifying 2,885 “G” instances (up from 2,769 in the BabyLM fine-tuned at epoch one) show that the model is improving in distinguishing grammatical sentences from the three error types. On the other hand, it confuses the “G” and “PREP” classes more (747 instances).

Overall, fine-tuning the BabyLM from the first to the third pretraining epoch results in performance gains, with the model pretrained for three epochs achieving a higher macro-average F0.5-score (85.80%) compared to the one pretrained for one epoch (83.40%). Although the F0.5-scores of the “G”, “DET”, and “SVA” classes have increased, especially the “SVA” class ($\Delta = 8\%$), the performance of the “PREP” class has lowered ($\Delta = -1.3\%$).

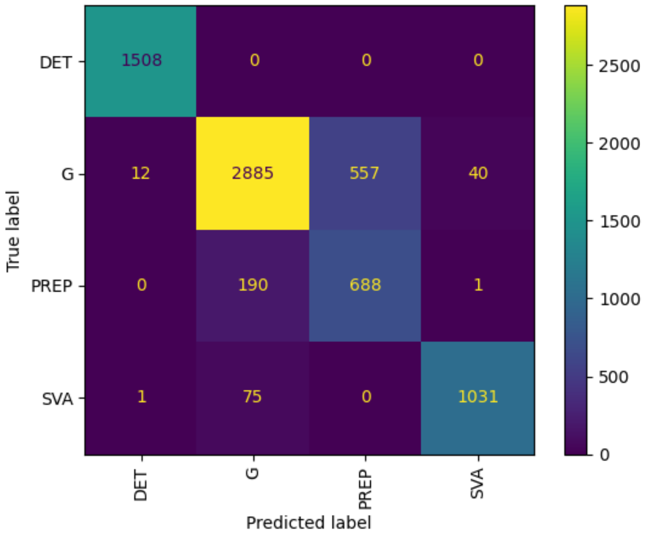

Figure 4.3: The Confusion Matrix of Fine-tuning the BabyLM Pretrained for Three Epochs on the GED Classification Task.

GED Evaluation of the BabyLM Pretrained for Five Epochs:

The results for the BabyLM fine-tuned at epoch five, displayed in Table 4.13 show a continued trend of improvement in the classification performance. The "DET" class maintains perfect recall (100%) and near-perfect precision (99.30%), achieving an F0.5-score of 99.50%, confirming the model's mastery of determiner error detection. The performance for the "SVA" class is also strong, with an F0.5-score of 95.40%, high precision (95.60%) and improved recall (94.70%). The "G" class shows marginal gains in precision (92.10%) in comparison with the BabyLM fine-tuned at epoch three (91.60%), though recall remains fixed at 82.60%, reflecting ongoing misclassifications, especially into the "PREP" class (550 instances), shown in Figure 4.4. The model still struggles with "PREP" class, with an F0.5-score of 59.10%. It shows a low precision (55.60%) due to consistent misclassification into "G" (189 instances), though recall (78.50%) remains relatively stable.

	precision	recall	F0.5-score	support
DET	0.993	1.000	0.995	1508
G	0.921	0.826	0.900	3494
PREP	0.556	0.785	0.591	879
SVA	0.956	0.947	0.954	1107
accuracy			0.878	6988
macro avg	0.857	0.889	0.860	6988
weighted avg	0.896	0.878	0.890	6988

Table 4.13: The Classification Report of the Results for Fine-tuning the BabyLM Pretrained for Five Epochs on the GED Classification Task.

As shown in Figure 4.4, the model demonstrates perfect recall for the "DET" class, correctly identifying all 1508 determiner errors. However, its precision for "DET" is not perfect, as ten instances labeled with "G" were incorrectly predicted as "DET".

The model also performs well on the “SVA” class (1048 correct predictions) and shows a modest improvement in reducing misclassifications into the “G” class (59 errors), compared to 75 when the BabyLM, pretrained for three epochs, was fine-tuned. The “G” class still suffers from confusion with the class of “PREP”. The misclassification of “PREP” into “G” (189 instances) remains nearly unchanged, suggesting the model’s inability to fully distinguish these two classes. However, true positive “PREP” predictions increase slightly to 690, contributing to stable recall. Overall, fine-tuning from epoch three to five, the BabyLM fine-tuned at epoch five shows slight performance gains (Δ = 0.2% for the macro average F0.5-score).

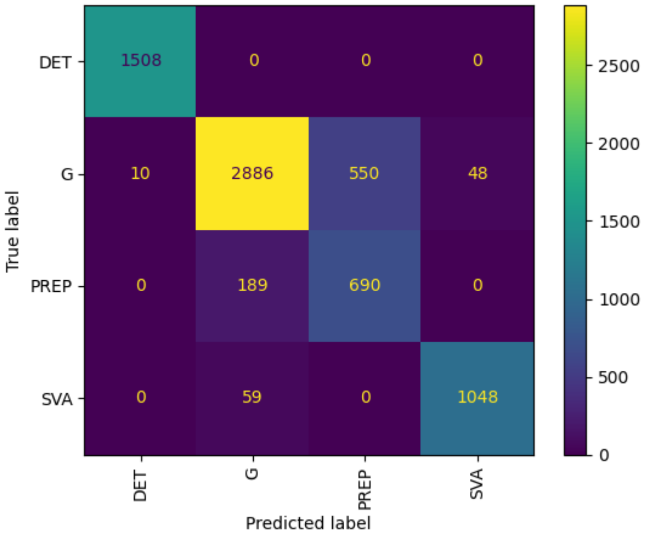

Figure 4.4: The Confusion Matrix of Fine-tuning the BabyLM Pretrained for Five Epochs on the GED Classification Task.

GED Evaluation of the BabyLM Pretrained for Nine Epochs:

The BabyLM fine-tuned at epoch nine achieves its highest overall performance, with a macro-average F0.5-score of 86.10% (see Table 4.14). Similar to the earlier fine-tuned training stages of the BabyLM, performance on “DET” and “SVA” classes remains strong, with near-perfect scores. The “SVA” class shows a slight improvement in precision (Δ = 0.4%) compared to the BabyLM fine-tuned at epoch five. In contrast, performance on the “G” class remains lower, with a precision of 92.20% and a recall of 82.70%. The “PREP” class also continues to be challenging, with an F0.5-score of 59.10%. In comparison with the BabyLM fine-tuned at epochs three and five, although the BabyLM fine-tuned at epoch nine shows slightly higher precision (55.70%) and recall (78.70%) for “PREP” class, overall performance on this class remains consistently low across all fine-tuned pretraining epochs.

Although the misclassification counts for the “SVA” class vary slightly (e.g., fewer “SVA” errors are now misclassified, and one error goes to “DET”), the overall misclassification trends mirror those from the earlier BabyLM’s fine-tuned epochs. As seen in Figure 4.5, the error patterns largely reflect the same challenges noted previously—particularly the frequent “G” class misclassified as “PREP” and the “PREP”

	precision	recall	F0.5-score	support
DET	0.993	1.000	0.994	1508
G	0.922	0.827	0.901	3494
PREP	0.557	0.787	0.591	879
SVA	0.960	0.946	0.957	1107
accuracy			0.878	6988
macro avg	0.858	0.890	0.861	6988
weighted avg	0.897	0.878	0.891	6988

Table 4.14: The Classification Report of the Results for Fine-tuning the BabyLM Pretrained for Nine Epochs on the GED Classification Task.

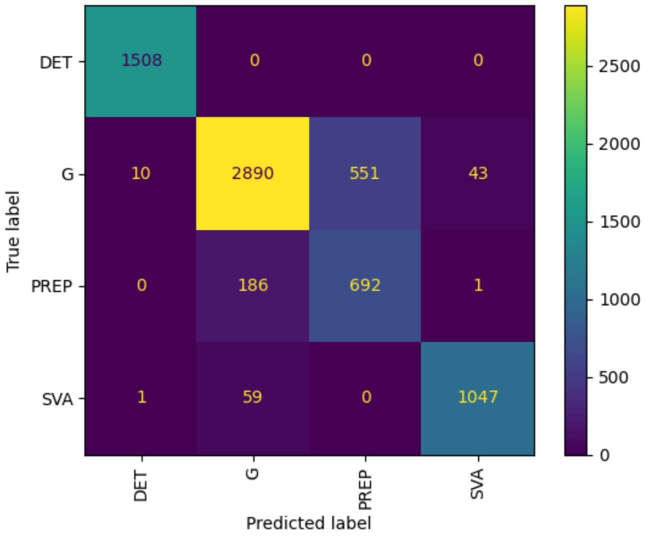

Figure 4.5: The Confusion Matrix of Fine-tuning the BabyLM Pretrained for Nine Epochs on the GED Classification Task.

class misclassified as “G”. This suggests a persistent difficulty in distinguishing grammatical sentences from preposition errors, likely due to the nuanced nature of prepositional usage. Overall, fine-tuning from epoch five to nine, the BabyLM fine-tuned at epoch nine shows marginal performance gains ($\Delta = 0.1\%$ for the macro average F0.5-score), but it still struggles with the “PREP” class.

GED Evaluation of the BabyLM Pretrained for Ten Epochs:

The macro average precision (85.50%), recall (88.50%) and F0.5-score (85.80%) of the BabyLM fine-tuned at epochs ten and three are the same (shown in Table 4.15 and Table 4.12). Despite these same average scores, they perform slightly differently in classifying the classes. Based on these results and those of Table 4.8, there is not much performance gains after the BabyLM fine-tuned at epoch five, except at the epoch nine. Therefore, it may suggest that it is enough to fine-tune the BabyLM up to its fifth pretraining epoch.

According to Table 4.15 and Figure 4.6, the model continues to perfectly identify 1508 instances of the “DET” class, achieving a 99.60% F0.5-score. This consistency

	precision	recall	F0.5-score	support
DET	0.995	1.000	0.996	1508
G	0.917	0.824	0.896	3494
PREP	0.548	0.782	0.583	879
SVA	0.960	0.936	0.955	1107
accuracy			0.874	6988
macro avg	0.855	0.885	0.858	6988
weighted avg	0.894	0.874	0.888	6988

Table 4.15: The Classification Report of the Results for Fine-tuning the BabyLM Pretrained for Ten Epochs on the GED Classification Task.

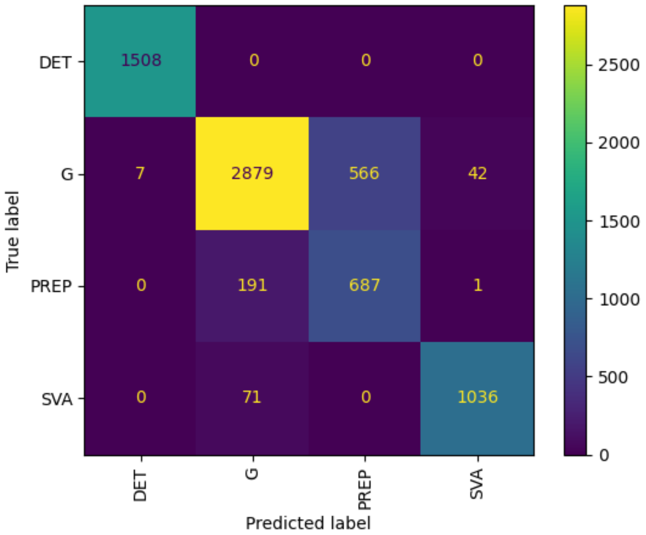

Figure 4.6: The Confusion Matrix of Fine-tuning the BabyLM Pretrained for Ten Epochs on the GED Classification Task.

aligns with the earlier BabyLM's fine-tuned training stages (epochs three, five, and nine), where the "DET" class appears to be the most easily learned error type. For the "SVA" class, the model achieves strong performance, correctly classifying 1036 instances. However, it mislabels 71 "SVA" instances as "G", an increase of 12 misclassifications compared to the model fine-tuned at epoch nine (59 misclassifications; see Figure 4.5). This misclassification shift likely contributes to the slight drop in overall performance at epoch ten. The "PREP" class remains the most challenging error type. The model correctly identifies 687 out of 879 instances, but misclassifies 191 as the "G" class. Despite a relatively high recall (78.20%), the model's precision is low (54.80%), leading to an F0.5-score of 58.30%. This again highlights the difficulty of accurately detecting prepositional errors, which often require nuanced syntactic and semantic understanding. Lastly, for the "G" class, the model achieves an 89.60% F0.5-score, correctly classifying 2879 instances while misclassifying 566 as "PREP". As shown in Table 4.15, although the overall macro performance remains stable, minor shifts in class predictions contribute to a modest performance drop ($\Delta = -0.3\%$) compared to epoch nine (as shown in Table 4.14).

GED Evaluation of BB-RoBERTa:

As discussed above (according to Table 4.8), the BabyLM fine-tuned from its third to last pretraining epochs, especially at epoch nine, shows a competitive performance to the fine-tuned BB-RoBERTa. As shown in this table, the fine-tuned BB-RoBERTa achieves a macro average F0.5-score of 86.20%, slightly outperforming the BabyLM fine-tuned at epoch nine by 0.1%.

	precision	**recall**	**F0.5-score**	**support**
DET	0.999	1.000	0.999	1508
G	0.915	0.845	0.900	3494
PREP	0.538	0.702	0.565	879
SVA	0.985	0.986	0.985	1107
accuracy			0.883	6988
macro avg	0.859	0.883	0.862	6988
weighted avg	0.897	0.883	0.893	6988

Table 4.16: The Classification Report of the Results for BB-RoBERTa Fine-Tuned on the GED Classification Task.

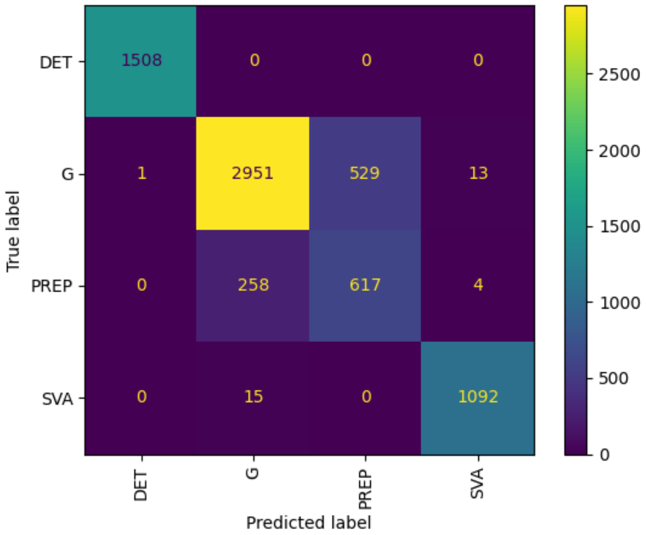

Figure 4.7: The Confusion Matrix of BB-RoBERTa Fine-Tuned on the GED Classification Task.

Similar to the BabyLM's fine-tuned epochs, the fine-tuned BB-RoBERTa performs well in classifying the "DET" class, with F0.5-score of 99.90% (as shown in Table 4.16). In contrast, it classifies the "SVA" class with higher precision (98.50%), recall (98.60%), F0.5-score (98.50%) scores. Close to the performance of some of the BabyLM's fine-tuned epochs, the fine-tuned BB-RoBERTa achieves precision of 91.50%, recall of 84.50%, and F0.5-score of 90.00% for the "G" class. While it correctly distinguishes 2951 instances, it confuses 529 instances of the "G" class with the class of "PREP" (shown in Figure 4.7). Similar to the BabyLM's fine-tuned training stages, the fine-tuned BB-RoBERTa struggles to distinguish the "PREP" class. In comparison with other error types, it shows lower precision (53.80%), recall (70.20%), F0.5-score

(56.50%) scores for the "PREP" class. The fine-tuned BB-RoBERTa correctly classifies 617 instances of the "PREP" class, but mislabels 258 instances as the "G" class and four instances as the "SVA" class.

GED Evaluation of O-RoBERTa:

Among all the fine-tuned models, the fine-tuned O-RoBERTa shows the highest performance, with the macro average F0.5-score of 90.10%, as shown in Table 4.17. Similar to the fine-tuned BB-RoBERTa and the BabyLM's fine-tuned epochs, the fine-tuned O-RoBERTa achieves high performance in detecting the "DET" class, with precision of 99.90%, recall of 100%, and F0.5-score of 99.90%. Overall, the "DET" class is the easiest to learn for all the fine-tuned models. After the "DET" class, the fine-tuned O-RoBERTa shows high performance for the class of "SVA". It achieves precision of 99.50%, recall of 99.20%, and F0.5-score of 99.50% for "SVA" error type.

	precision	**recall**	**F0.5-score**	**support**
DET	0.999	1.000	0.999	1508
G	0.905	0.939	0.911	3494
PREP	0.722	0.617	0.698	879
SVA	0.995	0.992	0.995	1107
accuracy			0.920	6988
macro avg	0.905	0.887	0.901	6988
weighted avg	0.917	0.920	0.917	6988

Table 4.17: The Classification Report of the Results for O-RoBERTa Fine-Tuned on the GED Classification Task.

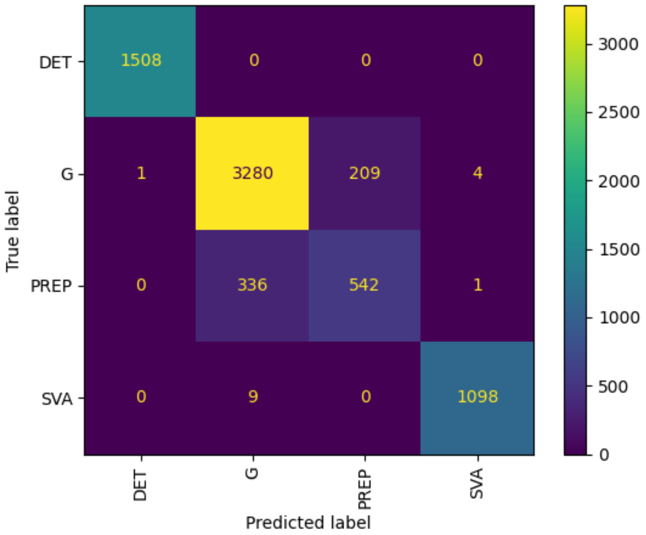

Figure 4.8: The Confusion Matrix of O-RoBERTa Fine-Tuned on the GED Classification Task.

As it is displayed in Figure 4.8, the fine-tuned O-RoBERTa correctly classifies 1098 instances of the "SVA" class and misclassifies only nine instances as the class of "G".

Similar to the the BabyLM's fine-tuned epochs and the fine-tuned BB-RoBERTa, most of its misclassifications happen with the "PREP" class, where it mainly mislabels them as the "G" class. It shows its lowest performance for the "PREP" class, with the macro average precision of 72.20%, recall of 61.70%, and F0.5-score of 69.80%. This challenge with the "PREP" class can be observed in Figure 4.8 where the fine-tuned O-RoBERTa incorrectly classifies 336 instances as the class of "G". This persistent challenge, even in O-RoBERTa models, likely stems from the inherent ambiguity of prepositional usage in English, where context often plays a key role and labeled examples may lack sufficient variability. Among all the fine-tuned models, the fine-tuned O-RoBERTa shows the highest misclassification for the preposition error types. Although it shows high precision (90.50%), recall (93.90%), and F0.5-score (91.10%) for the "G" class, it misclassifies 209 instances as the "PREP" class. Yet, it shows the highest correct classification for the "G" class (3280 instances).

Summary of the Results:

As the results show, unlike the zero-shot evaluation results (reported in Section 4.2) where the performance difference between the BabyLM, BB-RoBERTa, and O-RoBERTa was large, the BabyLM shows a competitive performance when it is fine-tuned for the GED task (reported in Section 4.3). Most of the performance gap disappears when the BabyLM is fine-tuned even for one epoch on the task. This suggests that the BabyLM is a competitive alternative to its larger LLM counterpart for similar GED tasks. Although there is $\Delta = -4.3\%$ F0.5-score between the BabyLM fine-tuned at epoch ten (the last pretraining epoch) and and the fine-tuned O-RoBERTa (according to Table 4.8), the BabyLM has ∼10 times lower parameters (13M vs 125M), ∼3000 times lower training data size (52MB vs 160GB), and much lower environmental impacts and computational costs. However, this finding is limited to the three targeted error types, and the BabyLM may not scale well to other error types.

As the results of the zero-shot evaluation shows, the subject-verb-agreement errors seem to be the most problematic error type, among the three targeted errors, for the models. In contrast, when the models are fine-tuned on the GED task, they struggle with preposition errors. Even the fine-tuned O-RoBERTa faces much challenge in detecting the "PREP" class.

An interesting point is also observed in the zero-shot and GED evaluation results. Based on Table 4.6, the zero-shot performance of the BabyLM at its tenth epoch lowers both in average BLiMP score of the all error types and the average BLiMP score of the three targeted error types. This may suggest that the BabyLM is overfitting from its tenth epoch on its training dataset. Also, Figure 4.1 shows that the BabyLM's performance plateaus from the fifth epoch in its zero-shot evaluation which is in line with the fine-tuning results of the BabyLM. In other words, there is not much performance gain after fine-tuning the BabyLM pretrained for five epochs on the GED task.

Chapter 5

Error Analysis

This section presents an error analysis of the GED classification task of detecting the three targeted error types as well as correctly formed grammatical sentences. Among the BabyLM's fine-tuned training stages, the fine-tuned BabyLM pretrained for nine epochs was selected due to its highest performance, which was also closest that of the fine-tuned O-RoBERTa (as described in Section 4.3). In this chapter, I refer to the BabyLM fine-tuned after being pretrained for nine epochs as "the fine-tuned BabyLM". While the primary focus is on comparing the GED performance of this model with that of the fine-tuned O-RoBERTa, the performance of the fine-tuned BB-RoBERTa is also examined in relevant sections. The chapter is divided into two sections. First, Section 5.1 presents an overview of the models' misclassifications in comparison with each other and then, Section 5.2 analyzes some of the misclassified instances.

5.1 Comparing the Models' Predictions

Table 5.1 summarizes a comparative analysis of the models' misclassifications across the three targeted error types of prepositions ("PREP"), determiners ("DET"), and subject-verb-agreement ("SVA"), as well as the correct grammatical sentences ("G"). In this table, the performance of one or two models is contrasted with that of the remaining model(s). Types of misclassifications are represented using the formats such as "SVA-G", "PREP-SVA", "G-SVA", and "DET-G", where the first label indicates the gold label of an instance and the second denotes the incorrect predicted class by one or

			O and BB (Error)	Baby and O (Error)	Baby and BB (Error)	O (Error)	BB (Error)	Baby (Error)
	All Errors	Unique Errors	Baby (Ok)	BB (Ok)	O (Ok)	Baby and BB (Ok)	Baby and O (Ok)	O and BB (Ok)
SVA-G	62	62	1	0	5	3	4	49
SVA-DET	1	1	0	0	0	0	0	1
SVA-PREP	0	0	0	0	0	0	0	0
PREP-SVA	2	2	1	0	1	0	0	0
PREP-G	420	420	63	23	73	182	57	22
PREP-DET	0	0	0	0	0	0	0	0
G-SVA	51	51	1	0	5	2	6	37
G-PREP	411	405	7	14	308	5	31	46
G-DET	12	12	0	0	0	1	1	10
DET-G	0	0	0	0	0	0	0	0
DET-PREP	0	0	0	0	0	0	0	0
DET-SVA	0	0	0	0	0	0	0	0

Table 5.1: The Models' GED Classification Performance in Comparison with Each Other. "O", "BB", and "Baby" represent O-RoBERTa, BB-RoBERTa, and the BabyLM respectively. "Errors" refers to "misclassifications". In the last six columns, "Error" indicates the model(s) misclassified, but "OK" means the model(s) classified correctly.

two models. For example, “SVA-G” refers to instances where the gold label is “SVA”, but one or two models incorrectly predicted as “G”, while the remaining model(s) correctly identified the instance as “SVA”. The “All Errors” column includes the total number of instances where one or two models misclassified between two error types, while the remaining model(s) correctly classified it. It may include repeated instances. “Unique Errors”, by contrast, displays the unique instances of “All Errors”. The remaining columns show the models’ comparative predictions. Each column represents instances where only one or two models correctly identified the error while the other model(s) did not. For example, “O and BB (Error); Baby (OK)” counts instances correctly classified by the BabyLM but not by the other two models, and “Baby (Error); O and BB (OK)” includes those instances where both O-RoBERTa and BB-RoBERTa correctly detected but the BabyLM failed. This structure provides a better comparative view of the models’ GED capabilities. For instance, high counts in the “Baby and BB (Error); O (OK)” set for “G” misclassified as the “PREP” class suggest that O-RoBERTa performs better in detecting the grammatical instances from the preposition error types compared to these two BabyLMs.

Table 5.1 highlights some differences in how the three models perform on the GED classification task. The table shows that the highest number of misclassifications occurs in the categories involving prepositions and grammatical instances, with 420 samples in “PREP-G” and 411 samples in “G-PREP” categories. It indicates that that preposition errors pose the greatest challenge across all the models. In contrast, there are no misclassifications observed between the “PREP” and “DET” error types, and all instances from the “DET” class are correctly classified. Among the three models, the fine-tuned O-RoBERTa consistently shows better performance, particularly in distinguishing grammatical sentences from preposition errors. The fine-tuned BabyLM shows a good performance overall but tends to misclassify more grammatical sentences, especially in the “G-PREP” category. On the other hand, the fine-tuned BB-RoBERTa demonstrates a slightly weaker performance compared to the fine-tuned O-RoBERTa.

5.2 Analysis of the Models’ Misclassifications

Out of the 12 possible misclassification types that could occur in the GED task, only seven types are observed in the models’ predictions. The remaining five possible error types never occurred in any of the models. The following subsections present a manual error analysis of these seven misclassification types, focusing on some mislabeling patterns and what they might reveal about the models’ grammatical sensitivity. For five of the categories (“SVA-G”, “SVA-DET”, “PREP-SVA”, “G-SVA”, and “G-DET”), all misclassified instances by each model are analyzed. In these categories, the first part denotes the gold label, and the second indicates the incorrect prediction made by a model. For example, “SVA-G” refers to instances that a model incorrectly classified “SVA” errors as “G”. However, since misclassifications involving “PREP” and “G” classes had too many instances, up to 30 random samples are taken from each set based on Table 5.2. This table displays instances where one or two models misclassified between “PREP” and “G” classes, but the remaining model(s) correctly classified them. Therefore, a total of 281 samples were randomly selected for these misclassifications: 165 for “PREP-G” and 116 for “G-PREP,” as detailed in Table 5.2. These randomly selected samples are analyzed in subsections 5.2.6 and 5.2.7. For example, 30 random samples from a total of 182 instances are analyzed in which

	All G+PREP	G-PREP	PREP-G	G-PREP Sampling Number	PREP-G Sampling Number
O (Error) Baby and BB (OK)	187	5	182 ⋆	5	30 ⋆
BB (Error) Baby and O (OK)	88	31	57	30	30
O and BB (Error) Baby (OK)	70	7	63	7	30
Baby and O (Error) BB (OK)	37	14	23	14	23
Baby (Error) O and BB (OK)	68	46	22	30	22
BB and Baby (Error) O (OK)	381	308	73 ⋄	30	30 ⋄
Total	831	411	420	116	165

Table 5.2: The Models' GED Performance in Distinguishing Preposition and Grammatical Instances in Comparison to Each Other. In the first column, "Error" indicates the model(s) misclassified, but "OK" means the model(s) classified correctly. The number of random samples extracted for the manual error analysis is also shown.

the fine-tuned O-RoBERTa misclassified "PREP" errors as "G", while the fine-tuned BabyLM and fine-tuned BB-RoBERTa classified them correctly (see ⋆ in Table 5.2). Similarly, 30 samples are randomly selected from a total of 73 instances where both the fine-tuned BabyLM and the fine-tuned O-RoBERTa misclassified "PREP" errors as "G", but the fine-tuned O-RoBERTa classified them correctly (see ⋄ in Table 5.2). On the other hand, as shown in Table 5.2 shows, for sets with fewer than 30 instances, all samples are analyzed. For example, there are only five cases where the fine-tuned O-RoBERTa misclassified "G" errors as "PREP", while the fine-tuned BabyLM and fine-tuned BB-RoBERTa classified correctly. Likewise, there are only seven cases where both the fine-tuned O-RoBERTa and fine-tuned BB-RoBERTa misclassified "G" errors as "PREP", but the fine-tuned BabyLM classified them correctly. All such instances are included in the analysis.

5.2.1 "SVA" Misclassified As "DET"

This section analyzes all the instances in which the models incorrectly labeled "SVA" errors as "DET". While the fine-tuned O-RoBERTa and fine-tuned BB-RoBERTa have no misclassifications for this category, the fine-tuned BabyLM has one misclassified instance. Despite this one misclassification, the fine-tuned BabyLM shows 95.70% F0.5-score for "SVA" (reported in Table 4.10). It incorrectly classifies "The sketch of these pictures alarm April." as "DET" instead of "SVA". This sentence potentially represents a syntactically ambiguous sample. While the intended subject–verb-agreement error lies between the singular subject "sketch" and the plural verb "alarm," the model may have been misled by an alternative parse. One possibility is that it interpreted "pictures" as a verb, followed by "alarm" as a noun and "April" as a proper noun in apposition (e.g., "an alarm called April"). Under this reading, the model may have perceived a missing determiner before "alarm," leading it to classify the sentence as a determiner error rather than an "SVA" error. Given the ambiguous nature of this case and the fact that the model has successfully classified similar structures elsewhere, this error type is not much problematic.

5.2.2 “G” Misclassified As “DET”

This section studies all the instances in which the models incorrectly classifies “G” errors as “DET”. There are ten misclassifications by the fine-tuned BabyLM for this category. A close examination of the misclassified examples indicates that the model struggles particularly with complex syntactic constructions, including object and subject relative clauses and long-distance dependencies. Sentences such as “Some children question every child that the movie embarrasses” and “Travis should notice this hospital that all mouths annoy that hasn’t upset every teacher” suggest that the model may be misled by the presence of nested clauses. Moreover, both sentences contain the quantifier “every”, a type of determiner that may be contributing to the confusion. The model might misinterpret the structure, wrongly attributing the syntactic complexity or quantifier usage to a determiner-related error.

There is only one misclassified sentence by the fine-tuned O-RoBERTa. It classifies “Matt can reveal one school and Lawrence can reveal a few brown school.” as “DET”. This sentence is from the BLiMP dataset and is wrongly labeled as grammatical because “school” should be “schools”, as imposed by “a few”. I checked the fine-tuned BabyLM’s and the fine-tuned BB-RoBERTa’s predictions for the same sentence to see how they performed. They classify it as “G”. Since the sentence has determiner-noun-agreement error type, the fine-tuned O-RoBERTa has performed better.

There is also one misclassified sentence for this form by the fine-tuned BB-RoBERTa. It labels “Irene knew senators that Pamela hugs.” as “DET”. The model likely misclassified this due to the absence of a determiner before “senators” while it is grammatically acceptable in English when referring to plural count nouns in a generic or indefinite context. This sentence appears to be a corner case that may challenge even robust models due to its syntactic subtlety.

5.2.3 “PREP” Misclassified As “SVA”

This section studies all the instances in which the models incorrectly classify “PREP” errors as “SVA”. Both the fine-tuned BabyLM and the fine-tuned O-RoBERTa incorrectly classify one sentence, “The truth of Shopping”, as “SVA”. This sentence is also misclassified as “SVA” by the fine-tuned BB-RoBERTa. The correct form (gold version) in the BLiMP-style preposition dataset is “The truth about Shopping”. The intended error type here is “PREP”, which may not be easily detected due to the fragmentary form of the instance. Although the BLiMP-style preposition dataset was randomly checked, it still needs to be validated by human annotators or through crowd-sourced judgments to ensure high-quality instances. There are three more similar misclassifications by the fine-tuned BB-RoBERTa. As stated earlier, the BLiMP-style preposition dataset is created from BEA-2019 shared task dataset. As previously discussed in Section 2.4.1 in Chapter 2, the BEA-2019 task builds upon CoNLL-2014 shared task and introduces a new annotated dataset, the Cambridge English Write & Improve and LOCNESS (Write&Improve+LOCNESS) corpus (Bryant et al., 2019). Although Bryant et al. (2019) state that the new corpus and CoNLL-2014 dataset are filtered and approved by human annotators, there are still a few problematic samples.

The other mislabeled instances by the fine-tuned BB-RoBERTa are “First tell to your parents.” and “A lot of money is involved in research to stop the increase levels of pollution.”. In the first instance, the preposition “to” is incorrectly inserted after the verb “tell,” which does not require a preposition when followed by an indirect object ("tell your parents"). In the second instance, there is a missing preposition error, as the correct form would be: "A lot of money is involved in research to stop the increase in levels of pollution". However, the sentence could also plausibly be corrected as "to stop the increasing level of pollution," which involves changing the noun phrase and verb form rather than addressing a preposition error. This ambiguity may partly explain why the model misclassified it as an "SVA" error. Misclassifying these sentences as "SVA" may suggest that the fine-tuned BB-RoBERTa may be misled by samples' complexity or lexical cues. Overall, the fine-tuned BabyLM detects this category better than the fine-tuned BB-RoBERTa and shows similar performance to the fine-tuned O-RoBERTa.

5.2.4 "G" Misclassified As "SVA"

This section analyzes all the instances in which the models incorrectly classify "G" as "SVA" errors. There are only four misclassified samples by the fine-tuned O-RoBERTa for this category, three of which are incorrectly labeled as "G" by the BLiMP dataset. Once again the BLiMP dataset included problematic instances with wrong gold labels. For example, there is an "SVA" error in "Most ladies confuses that person.", "Some actresses distracts Michael.", and "All drivers confuses those ill actresses." due to the plural subjects "Most ladies", "Some actresses", and "All drivers", which require plural verbs. However these sentences are incorrectly labeled as "G" in the dataset. Both the fine-tuned BabyLM and the fine-tuned BB-RoBERTa correctly identify just one of these sentences ("Some actresses distracts Michael.") having an "SVA" error, in contrast to its incorrect "G" gold label. Both the fine-tuned BabyLM and the fine-tuned BB-RoBERTa fail to detect the errors in the other two samples and, similar to the incorrect gold labels in the BLiMP dataset, label them as grammatical sentences. The other misclassified sample by the fine-tuned O-RoBERTa is "Only steak that a lot of guests notice ever embarrassed Danielle.". While the fine-tuned BB-RoBERTa also misclassifies this sentence, the fine-tuned BabyLM correctly classifies it despite its complex structure. This sentence contains an embedded subject relative clause ("that a lot of guests notice") modifying the singular noun "steak," which serves as the subject of the main clause. The misclassification by the fine-tuned O-RoBERTa and the fine-tuned BB-RoBERTa may suggest difficulty in identifying the correct head of the subject noun phrase and maintaining subject-verb-agreement across embedded structure of this sentence.

There are twelve more misclassifications by the fine-tuned BB-RoBERTa for this category. Two of them are due to the wrong gold label or low-quality samples of the datasets. The BLiMP dataset incorrectly assigns the gold label "G" to the sentence "A lot of alumni bothers some hospital", even though it contains an "SVA" error. "Alumni" is the plural form of "alumnus" or "alumna" and requires a plural verb. While the fine-tuned O-RoBERTa and the fine-tuned BB-RoBERTa correctly detect the "SVA" error in this sample, the fine-tuned BabyLM classifies it as "G", the same as the wrong gold label of the BLiMP dataset. The other problematic instance is ".", which is just a punctuation and should have been removed from the BLiMP-style preposition dataset. The other ten mislabeled sentences by the fine-tuned BB-RoBERTa have different grammatical structures, ranging from simple (e.g., "The child hunts.") to complex (e.g., "A lot of senators who had disturbed Ellen profit.").

There are 43 misclassified instances by the fine-tuned BabyLM for this category, six of which are similarly mislabeled by the fine-tuned BB-RoBERTa. Two of these six cases involve problems with the gold labels themselves, as discussed in earlier paragraphs. For example, the sentence “Some actresses distracts Michael.” is incorrectly labeled as grammatical in the dataset despite having a clear “SVA” error; the fine-tuned O-RoBERTa successfully detects the error here. However, for another problematic sentence—“The senators embarrasses Martin.”—even the fine-tuned O-RoBERTa fails to detect the “SVA” error and instead follows the incorrect gold label by classifying it as grammatical. The remaining four shared misclassifications include sentences such as “The Borgias shocks Elaine.” and “The truth about Shopping,” which may seem grammatical on first glance but carry subtle syntactic ambiguities. “The Borgias shock Elaine.”, is ambiguous. If “The Borgias” refers to the TV show, the singular verb shocks” is correct; if it refers to the Borgia family, the plural verb “shock” is appropriate. Others, like “The child hunts.” and “First tell your parents.” are grammatically correct and are wrongly classified as having “SVA” errors by both the fine-tuned BabyLM and fine-tuned BB-RoBERTa. These misclassifications by the fine-tuned BabyLM and the fine-tuned BB-RoBERTa may be related to the limited exposure to some lexical or syntactic patterns during pretraining, as both models are trained on the strict-small dataset provided by the BabyLM Challenge. The remaining 37 misclassifications by the fine-tuned BabyLM have different grammatical structures, ranging from simple (e.g., “Theodore cures Christina.”) to complex (e.g., “A niece of a lot of cashiers implores Matt to ascend a lot of steps.”).

One possible pattern that can be observed among the fine-tuned BabyLM’s misclassifications is related to the verb tenses. One possible issue is that the fine-tuned BabyLM incorrectly labels sentences as “SVA” class when they have tenses such as past, perfect, or passive forms, rather than the present simple. For example, it misclassifies “Sharon froze the ice cream.”, “Many adults went fishing.”, and “Companies were researched by the patient.” with “SVA” errors. Moreover, it struggles with the irregular plural noun “paralyses” in “The paralyses disturb Frank.” and “Many paralyses annoy Tracy.” and incorrectly classifies them as “SVA” class instead of “G”, even though this noun appears in the training subset of the fine-tuning dataset. It also faces some challenges in detecting sentences that have nested or complex structures. For example, “These waitresses’ associate wasn’t boycotting few concealed lakes and Maria wasn’t boycotting many.” has complex subject structure and negative format, which may mislead the model. Similarly, “These actors that haven’t investigated Cheryl paint.” has a subject-relative clause, which may challenge the model to correctly identify the main clause subject (“actors”) and its verb (“paint”) for subject-verb-agreement. Additionally, the sentence structure is somewhat ambiguous and may be misinterpreted as a noun phrase fragment—especially in the absence of punctuation—making it harder for the model to robustly parse the clause boundaries.

5.2.5 “SVA” Misclassified As “G”

This section examines all the instances in which the models incorrectly classify “SVA” errors as “G”. There are nine misclassifications for this category by the fine-tuned O-RoBERTa. It aligns with its strong overall performance on “SVA” with an F0.5-score of 99.50% (reported in Table 4.10). Three of these misclassifications result from the incorrect gold labels in the BLiMP dataset: “The person bores Kevin.”, “Some actresses distract Michael.”, and “A lot of alumni bother some hospital.” In contrast to their incorrect gold labels (“SVA”), the fine-tuned O-RoBERTa correctly classifies them as “G”. As previously discussed in Section 5.2.4, the BLiMP dataset has incorrectly labeled sentences such as “Some actresses distracts Michael.” and “A lot of alumni bothers some hospital." as "G" instead of "SVA". It seems that there are some confusions in the BLiMP dataset between grammatical and ungrammatical constructions.

The remaining five misclassified instances by the fine-tuned O-RoBERTa are "The noses bothers Alicia.", "The oases hurts this doctor.", "The glasses bothers those senators.", "A lot of men insults Wayne." and "Those men obliges every bank to cooperate." In all of these sentences, the subject is clearly plural, yet the verb is incorrectly used as singular, resulting in an "SVA" error. The failure of the fine-tuned O-RoBERTa to detect these violations may reflect difficulty in consistently identifying the plurality of certain noun forms—particularly irregular plurals (e.g., "oases")—or in handling complex quantifier phrases (e.g., "a lot of men"). The last mislabeled instance by the fine-tuned O-RoBERTa, "The Borgias shock Elaine.", is ambiguous, as discussed in Section 5.2.4. Therefore, this misclassification appears to be a corner case.

Similarly, both the fine-tuned BabyLM and the fine-tuned BB-RoBERTa misclassify six instances, which are the same samples that are mislabeled by the fine-tuned O-RoBERTa. Three of them ("The person bores Kevin.", "Some actresses distract Michael.", "A lot of alumni bother some hospital.") suffer from the aforementioned gold label issues. All three models, therefore, correctly predict against the incorrect gold labels. Similar to the fine-tuned O-RoBERTa, they struggle to classify the ambiguous instance "The Borgias shock Elaine.", which appears to be a corner case. Similar to the fine-tuned O-RoBERTa, the fine-tuned BB-RoBERTa misclassifies "The oases hurts this doctor." and "The glasses bothers those senators." as correct grammatical sentences. While the fine-tuned BabyLM fails to detect the "SVA" error in "The oases hurts this doctor.", it manages to correctly classify "The glasses bothers those senators." as the "SVA" class.

Despite the fine-tuned BB-RoBERTa's near-perfect F0.5-score (98.50%) for the "SVA" class, if misclassifies fifteen instances for this category, six of which were discussed in the previous paragraph. The remaining nine misclassifications by this model have simple syntactic structure of regular plural subject-verb-agreement. For example, "Senators has that carriage." and "Schools criticizes these glaciers." are incorrectly detected as correct grammatical sentences while they have clear "SVA" errors.

Comparing the three fine-tuned models together, the fine-tuned BabyLM exhibits the highest number of misclassifications for this category. It misclassifies 59 total instances. Five of these overlap with the fine-tuned O-RoBERTa cases already discussed. While this suggests room for improvement, it's important to note that despite this count, the fine-tuned BabyLM maintains a high F0.5-score of 95.70% for the "SVA" class (reported in Table 4.10). This means that its performance remains strong in practice, and the 59 misclassifications happen in particularly challenging or ambiguous constructions. These include errors with regular plural subject–verb-agreement (e.g., "Senators listens to Janet."), irregular plurals (e.g., "Stimuli was astounding Katherine."), and long-range dependencies (e.g., "Most waitresses who concealed Laura spurs Martin to aggravate Kevin.")

5.2.6 "PREP" Misclassified As "G"

This section analyzes instances where the models misclassify "PREP" errors as "G". As shown in Table 5.2, I analyze 165 random samples out of 420 misclassifications for this error type. Some of the misclassified instances have preposition selection errors typical of ESL learners, such as "opposite of the hotel" (instead of "opposite the hotel"), "keen at sports" (rather than "keen on sports"), or "thank you very much about that" (instead of "for that"). These misuses often reflect non-idiomatic yet semantically plausible constructions, which may explain why models sometimes classify them as grammatical. The fine-tuned O-RoBERTa and BB-RoBERTa often misclassify such sentences as "G", while the fine-tuned BabyLM correctly identifies the prepositional errors. For instance, the fine-tuned BabyLM correctly classifies samples like "provide me all these emotions" (missing "with") and "arrived to the shopping centre" (instead of "at"), which are misclassified by the other two models. This may suggest that the fine-tuned BabyLM may be more sensitive to verb-preposition collocations. However, this advantage is not consistent across all samples. In contrast, the fine-tuned BabyLM sometimes misses clear prepositional errors that are correctly identified by both the fine-tuned O-RoBERTa and the fine-tuned BB-RoBERTa, such as "write you" (instead of "write to you"), "enter in my grandmother's room" (rather than "enter my grandmother's room"), or "reply your letter" (instead of "reply to your letter"). These instances may suggest the fine-tuned BabyLM's limitations in modeling less explicit syntactic constructions or more idiomatic usages.

In other cases, the fine-tuned BB-RoBERTa correctly identifies subtle errors such as "addicted by using their car" (instead of "to using"), "bad experiences about tents" (rather than "with tents"), or "good talent of golf" (instead of "for golf"), while both the fine-tuned BabyLM and the fine-tuned O-RoBERTa fail to correctly classify them. These instances may suggest that the fine-tuned BB-RoBERTa may have learned some fine-grained idiomatic patterns. Similarly, the fine-tuned O-RoBERTa shows strength in detecting prepositional collocation violations in samples such as "invest enough money on it" (instead of "in"), "tell to everyone" (rather than "tell everyone"), or "discussed about it" (instead of "discussed it"), which the other two models misclassified. In these instances, the fine-tuned O-RoBERTa's performance may reflect more accurate modeling of syntactic dependencies or exposure to relevant training data.

Another consistent pattern in the error patterns is that all models—especially the fine-tuned O-RoBERTa and the fine-tuned BB-RoBERTa—struggle with long, informal, or conversational sentences that embed the preposition error in otherwise fluent contexts. For example, "for" is unnecessary in "I'm looking forward to joining you in July and hoping for this trip is an amazing experience." Similarly, "they wouldn't give me it" is ungrammatical in "And as lastly that was the worst night I had ever had so I asked for my money back but they wouldn't give me it so I had an argument with the people who worked there but I couldn't get my money back." In such cases, the syntactic noise and overall fluency seem to distract the models from recognizing missing or incorrect prepositions, especially when these are not central to the sentence's main clause.

Overall, while all models struggle with certain types of "PREP" errors—especially those that are semantically transparent or embedded in fluent, complex sentences—some differences emerge. The fine-tuned BabyLM appears more sensitive to overt learner-like preposition misuse, while the fine-tuned O-RoBERTa and the fine-tuned BB-RoBERTa occasionally excel in idiomatic or collocational contexts. However, these strengths are distributed inconsistently, and no single model dominates across all subtypes.

5.2.7 "G" Misclassified As "PREP"

This section analyzes samples where the models incorrectly label grammatically correct sentences as "PREP" errors. As shown in Table 5.2, this type of misclassification occurs in 411 instances across all six sets. Of these 411 instances, the largest group

(308 instances) consists of instances that the fine-tuned O-RoBERTa correctly classifies as grammatical, while the fine-tuned BabyLM and the fine-tuned BB-RoBERTa misclassify them. Across the 116 randomly selected samples from the table, a few broad patterns emerge. One of the most common involves idiomatic or phrasal verb constructions, such as "write to you", "ask you to", or "good at swimming", which are standard English collocations that require lexical and syntactic knowledge for correct detection. Both the fine-tuned BabyLM and the fine-tuned BB-RoBERTa tend to misclassify these as prepositional errors. This tendency may be attributed to the fine-tuned BabyLM's smaller scale and likely more limited exposure to diverse syntactic structures during training, which leads it to overdetect prepositional misuse where none exists.

Another pattern involves sentences with embedded syntactic structures. Examples like "He had been replaced by a really bad actor, of whom I don't even know the name," or "Finally, I would like to ask some questions such as about clothes and money," feature non-canonical word orders. Although grammatical, such sentences may contain unexpected surface forms (e.g., "of whom", "such as about") that trigger false error detections by the fine-tuned BabyLM and the fine-tuned BB-RoBERTa. In particular, sequences like "such as about" are syntactically difficult to interpret without broader context, which motivates using sequence-labeling approaches in GED tasks. In contrast, the fine-tuned O-RoBERTa appears more robust to these deviations, often classifying them correctly. Moreover, fragmentary sentences such as "Specially at weekends." or "Concerning accommodation," also pose challenges, primarily for the fine-tuned BabyLM and the fine-tuned BB-RoBERTa. These structures are pragmatically acceptable in discourse contexts but may be classified as ungrammatical if models prioritize complete syntactic forms.

Overall, the fine-tuned O-RoBERTa shows relatively consistent ability to identify grammatical constructions, even in idiomatic or structurally complex contexts. Both the fine-tuned BabyLM and the fine-tuned BB-RoBERTa, however, tend to over-predict preposition errors, especially in cases where syntactic complexity or lexical collocations may resemble typical ESL misuse. This may suggest that these models require improved modeling of diverse syntactic structures and idiomatic expressions.

Conclusion of the Chapter:

This chapter presented a manual error analysis of misclassified instances across several error categories, with particular focus on the "PREP" and "G" classes due to their high frequency of confusion. These two categories appear to pose the greatest challenge for the models, often being influenced by subtleties in collocation, idiomatic usage, or syntactic structure. Among the three models, the fine-tuned BabyLM exhibited greater difficulty in handling longer, structurally complex sentences. It frequently failed to detect clear prepositional misuse or incorrectly labeled grammatical sentences, suggesting limitations in its syntactic generalization ability and its sensitivity to idiomatic expressions.

Moreover, the manual inspection also revealed issues in the dataset itself. Some sentences were found to have incorrect gold labels, which may have affected the models' training or evaluation. In other cases, sentences were genuinely ambiguous—either between being grammatical or ungrammatical, or between different error types. These ambiguities reflect real-world variation in language use and make certain classifications inherently subjective, even for human annotators.

Chapter 6

Discussion

This chapter interprets the findings of the experiments conducted in this thesis. By comparing the performance of a RoBERTa-based BabyLM to the RoBERTa-base model across both zero-shot and fine-tuned GED tasks, this study aims to assess the potential of environmentally and computationally efficient BabyLMs for grammaticality assessment and GED task, focusing on three common error types among ESL learners (determiners, prepositions, and subject-verb-agreement). The discussion integrates the results within the broader context of LM research, highlights the strengths and limitations of using BabyLMs for the similar tasks, and goes through some implications for model design and training or evaluation datasets.

6.1 Training BabyLMs

One of the challenges in this thesis was engaging with the concept of a "BabyLM"—a term that, while widely used in the BabyLM Challenge, lacks a precise and widely accepted definition. It is not clear whether it is defined by the numbers of parameters, training data size, or combination of these. In the BabyLM Challenge, it often refers to models trained under resource constraints that are intended to simulate early language learning in humans. However, this definition leaves considerable room for interpretation, particularly when it comes to balancing model complexity and training data size.

In my thesis, I aimed to train BabyLMs that are resource-efficient yet representative of scalable language learning. To this end, I imposed a parameter limit of 24M, informed by the strict-small track of the BabyLM Challenge (described in Section 3.2.1). While this cap was effective in constraining model size, it also limited architectural choices. The final configurations (reported in Table 3.4 in Section 3.2.1) were not chosen for optimal performance but rather to remain in the defined threshold and train a competitive BabyLM for the thesis. Another critical design decision involved the training dataset. I used the strict-small BabyLM dataset (9.96M words), which is significantly smaller than what LLMs are typically trained on. This data limitation likely constrained the BabyLM's generalization ability, especially in capturing subtle grammatical patterns such as prepositional usage or subject–verb-agreement errors. The training dataset will be raised as a potential limitation below. As the results in Chapter 4 show, the BabyLM consistently underperforms in the zero-shot setting—especially for error types that likely require more extensive exposure or deeper representational depth. However, despite the constraints mentioned above, the BabyLM, when fine-tuned on the GED classification task, achieves competitive results approaching those of O-RoBERTa.

Overall, the decision to prioritize comparison over optimization is methodological. Rather than focusing on marginal improvements, I sought to understand how the BabyLM develops grammatical knowledge over time and how effectively it can assess grammaticality across the training stages. This focus offers insight into the development of SLMs and invites future research into the relative contributions of exploring BabyLMs for different NLP tasks.

6.2 Zero-Shot Evaluation: Grammaticality Assessment

For the zero-shot evaluation, the BLiMP dataset is used along with the BLiMP-style preposition dataset that I created for my thesis experiments. This evaluation serves two key purposes: first, to track how the BabyLM develops grammaticality assessment over training epochs, and second, to select the best BabyLM for the following experiments. The zero-shot evaluation, assesses how well a model generalizes to grammaticality judgments without task-specific fine-tuning. In this sense, it offers a lens into the BabyLM's internal linguistic representations and how they evolve through training. However, using the zero-shot evaluation to select the best BabyLM for this thesis and focusing on three error types will be raised as potential limitations below.

The results (presented in Section 4.2) reveal an upward trajectory in the BabyLM's performance during its early training stages (epochs one to four), followed by a performance plateau beginning around epoch five. Its best BLiMP scores are observed at epochs eight and nine, after which no substantial improvement is detected. The BabyLM performed best on determiner errors, reaching a peak of 78.37% at epoch eight. However, it is still substantially outperformed by BB-RoBERTa (90.75%) and O-RoBERTa (97.28%). For preposition errors, the BabyLM reached the accuracy of 62.24% at epoch nine, while BB-RoBERTa and O-RoBERTa achieved 73.08% and 91.28%, respectively. Notably, the BabyLM struggled most with "SVA" errors, improving only gradually and peaking with the accuracy of 55.52% at epoch eight. In contrast, O-RoBERTa excelled at "SVA" errors (91.47%), while BB-RoBERTa achieved the score of 65.42%.

The zero-shot evaluation results indicate that the BabyLM learns determiners better than prepositions and subject-verb-agreement. Among the three targeted error types, subject-verb-agreement seems more challenging and comes in the last position based on its lower BLiMP-score. This raises an intriguing question: why is "SVA"—which is typically considered a fundamental syntactic relation—so difficult for the BabyLM in the zero-shot setting? One possibility is that subject–verb-agreement errors involve long-distance dependencies and hierarchical structures that require a deeper level of syntactic abstraction than determiners or even prepositions. While determiners often occur in relatively fixed positions, "SVA" errors may depend on correctly identifying grammatical subjects and verbs across clauses, something particularly demanding for SLMs with limited training data. Interestingly, the pattern reverses after fine-tuning: the BabyLM performs much better on "SVA" errors than on preposition errors. This shift suggests that fine-tuning plays a critical role in specializing the model's grammatical representations. In contrast, the high zero-shot performance on preposition errors may reflect surface-level patterns learned during pretraining—patterns that do not generalize well under the stricter labeling conditions of the GED classification task (as reported in Section 4.3). This discrepancy reinforces the idea that zero-shot performance reflects implicit knowledge, whereas fine-tuning reflects task-specific adaptation.

The zero-shot evaluation also provides insight into the relative learnability of grammatical structures, paralleling findings in second-language acquisition. Like ESL learners, the BabyLM shows faster improvement in more surface-level grammatical structures such as determiners and slower gains in areas that generally require deeper syntactic understanding, like that of subject-verb-agreement (reported in Table 4.6). This alignment reinforces the argument that BabyLMs can offer insight into more human-like syntactic development, that may be useful in modeling low-resource language scenarios. However, these parallels should be interpreted cautiously, as language models learn through statistical optimization rather than cognitive processes.

6.3 Fine-Tuned Evaluation: GED Performance

Another set of insights can be learned from the GED classification performance of fine-tuned models on the GED task. The task involves sentence-level classification into four classes: grammatical ("G"), determiner ("DET"), subject-verb-agreement ("SVA"), and preposition ("PREP") error types. The BabyLM is fine-tuned for five epochs at at different stages of pretraining, using 60% of the GED classification dataset (described in Section 3.1.3). As previously discussed in Section 4.3, although the BabyLM is fine-tuned at all ten epochs, only some of them (epochs one, three, five, nine, and ten) are analyzed in detail. For comparison, BB-RoBERTa and O-RoBERTa are also fine-tuned for five epochs on the same task/dataset. An additional experiment with multiple random seeds was conducted for fine-tuning the BabyLM to confirm the consistency of the results of the main seed, which was randomly selected for the fine-tuning phase.

A key finding from this phase is that fine-tuning substantially reduces the performance gap observed in the zero-shot setting. After fine-tuning, the BabyLM's performance becomes competitive to that of the fine-tuned O-RoBERTa, despite its limited pretraining dataset and drastically fewer parameters. This result reinforces the importance of task-specific fine-tuning for compensating for pretraining limitations, especially in BabyLMs.

Interestingly, the BabyLM, which had the greatest difficulty with subject–verb-agreement errors in the zero-shot setting, detected this category more effectively after fine-tuning. Instead, preposition errors became the most challenging error type after fine-tuning the models on the GED task. This shift in difficulty may suggest that while "SVA" requires deep syntactic representations that may not readily emerge from pretraining alone, those representations can be learned through task-specific supervision. To further investigate this hypothesis, future work could evaluate the BabyLM's sensitivity to subject–verb-agreement using an MLM task, isolating agreement phenomena to better understand how syntactic representations develop across training stages. In contrast, preposition errors may require a more nuanced and extensive distributional exposure, which was likely underrepresented in both the pretraining and GED classification datasets. This highlights the limitation of training on relatively small datasets, particularly for capturing fine-grained grammatical categories like prepositional usage.

In terms of model development, the BabyLM shows the most substantial gains in GED performance during its earlier epochs—particularly from epochs one to five (displayed in Section 4.3). After this point, performance plateaus. While the highest performance is observed at epoch nine, the improvements beyond epoch five are minimal. The F0.5-score difference between fifth and tenth epochs is only 0.2%, which seems to indicate diminishing returns with further training. Moreover, the performance of the BabyLM fine-tuned at epoch ten and three are identical in their macro-average scores, reinforcing the lack of much gains beyond the fifth pretraining epoch (as shown in Table 4.8). More broadly, this plateau aligns with trends observed in the zero-shot evaluation, where improvements also flatten after epoch five. Overall, the results may imply that training BabyLM for five epochs followed by fine-tuning may be optimal, offering a balance between performance, efficiency, and reduced computational and environmental costs.

Another important outcome is the efficiency of the fine-tuned BabyLM compared to the fine-tuned O-RoBERTa. Even though the BabyLM falls behind O-RoBERTa by 4.3% in macro-average F0.5-score after fine-tuning (as reported in Table 4.8), it achieves this performance with significantly fewer computational and parameter resources. It has about ten times fewer number of parameters (13 million vs. 125 million), uses around 3000 times less training data (52MB vs. 160GB), and has substantially lower computational and environmental costs. Moreover, training BabyLM variants specifically for the GED task using tailored datasets may further narrow the performance gap. These comparisons underline the potential of SLMs for low-resource or environmentally conscious applications, especially when fine-tuned effectively.

However, several constraints must be acknowledged, some of which will be raised as potential limitations in this chapter. The GED task in this thesis involved only three error types, and extending this framework to cover a broader set of grammatical categories would have required a more extensive dataset and perhaps a more expressive model. It remains an open question whether similar levels of performance could be achieved with more subtle error types. Moreover, the GED evaluation was framed as a sentence-level classification task, which lacks the granularity of sequence labeling tasks. Extending this framework to identify not just the type of error but also its position (e.g., via sequence labeling or token-level classification) would complicate the task. For BabyLMs, such a shift would likely require architectural adjustments or more targeted training strategies.

6.4 Using BabyLMs As an Interpretability Proxy for LLMs

Due to factors such as their reduced size and constrained training data, BabyLMs may offer an interpretability advantage over LLMs. The results from both zero-shot and fine-tuned evaluations indicate that BabyLMs can show grammatical competence in a measurable way. Their smaller architecture allows for better tracking of performance development across training stages, providing some possible insights into how LMs might learn syntactic structures over time. Moreover, this approach not only reduces computational costs during the exploratory phases of research but also contributes to promoting more sustainable experiments.

6.5 Limitations

While my thesis tries to investigate the development and evaluation of BabyLMs for grammaticality assessment and a GED task, some limitations inevitably shaped the scope and outcomes of the research. In this section, I address some of these limitations, which arose from factors such as methodological decisions or dataset constraints. Acknowledging these constraints is essential for interpreting the results accurately and understanding the broader implications of the findings. The following sections discuss some of these limitations, which may have influenced the generalizability and interpretability of the results.

6.5.1 Computational Limitations

Computational limitations played a significant role in shaping the design and results of this thesis. Due to the limitation of time and computing resources, I was unable to train a larger number of BabyLM variants or conduct exhaustive hyperparameter searches that might have led to a more optimal architecture for my experiments. For instance, I trained only six BabyLMs with variations only in vocabulary size and number of hidden layers, while keeping other critical components such as hidden size and attention heads fixed. This constrained setup may have overlooked potentially better-performing configurations that could have emerged from a broader exploration of the training design.

Furthermore, the limitation of time and computing resources also affected my approach to fine-tuning. While I conducted experiments to determine an appropriate number of epochs for fine-tuning the selected BabyLM, I did not repeat this process individually for BB-RoBERTa and O-RoBERTa. Instead, I applied the same epoch setting derived from the BabyLM tuning experiment to all models. Ideally, separate tuning for each model would have allowed for a more fair comparison, as the optimal number of fine-tuning epochs may vary depending on the model's architecture and pretraining.

Moreover, as stated in Section 4.3, I carried out an additional experiment with multiple random seeds for fine-tuning the BabyLM, as a form of sanity-check to assess the reliability of the selected random seed. This experiment was not replicated for BB-RoBERTa and O-RoBERTa due to time and resource constraints. Running multiple seed experiments across all models would have provided a more robust evaluation by accounting for performance variability introduced by random seeds. This inconsistency limits the comparability across the models and may affect the reliability of conclusions drawn from their fine-tuned performance.

6.5.2 The Theoretical Limitations

The present thesis experiments also face theoretical limitations, some of which are briefly discussed in the following paragraphs.

Three Error Types

One limitation of this study is the focus on only three grammatical error types in both the zero-shot and the fine-tuning for the GED classification task evaluations. While narrowing the scope allowed for a more controlled and manageable experimental setup, it also constrained the comprehensiveness of the evaluations. Grammatical error correction and detection in real-world learner language includes a wide range of error categories, each of which may engage different aspects of a model's linguistic competence. By restricting the evaluation to just three error types, the findings may not generalize to broader GED tasks, potentially overlooking important differences in how the BabyLM handle other types of grammatical errors. Moreover, this narrow focus limits the interpretability of model performance in relation to the broader goals of grammatical error detection in second language acquisition, where learners often produce complex, overlapping error patterns. Future work would benefit from expanding the range of error types evaluated to obtain a better understanding of BabyLMs' capabilities in GED tasks.

Training the BabyLMs

While my training approach enabled the development and selection of a competitive BabyLM—one that approximates the performance of the BB-RoBERTa—it also introduces certain methodological trade-offs that may be seen as limitations. Importantly, this thesis did not aim to optimize for the best-performing BabyLM but rather to design a resource-efficient RoBERTa-based model suitable for comparative analysis with its larger counterpart, RoBERTa-base. To that end, I limited the training configurations to a limited range of vocabulary sizes (20K, 40K, 50K), two hidden layers (four and six), hidden size of 256, and eight attention heads (reported in Table 3.4), while keeping other parameters fixed. This constrained design ensured a controlled environment for comparison but may have limited the potential performance gains that more extensive hyperparameter exploration could offer. Similarly, setting a maximum parameter threshold of 24M—based on the review of the BabyLM Challenge's strict-small track (as discussed in Section 3.2.1)—may be seen as a limitation to the experiments. While this choice aligns with the goals of environmental efficiency and computational costs, it also reflects one interpretation of what a "BabyLM" should be.

Another factor is the issue of variance in results, which has been a critical topic in machine learning and particularly in deep learning, where even small changes such as different random seeds can lead to significantly different outcomes. While this issue is beyond the scope of this thesis, understanding them could help determine what level of difference is meaningful when only one source of variance is changed. However, despite accounting for variance through multiple random seed experiments, it would have been better to also include some other systematic approaches.

Using the BLiMP-Score to Choose the Competitive BabyLM

Another potential limitation of my BabyLM selection strategy lies in its reliance on zero-shot evaluation using the BLiMP dataset, following the evaluation pipeline introduced by the BabyLM Challenge (as described in Section 3.2.1). While this approach provides a standardized and interpretable benchmark for assessing the BabyLMs' grammaticality judgment, it may not reflect their potential performance in GED tasks. In fact, my experimental results revealed that the selected BabyLM, despite its relatively low BLiMP-score, performed substantially better after fine-tuning on the GED task. This discrepancy suggests that zero-shot BLiMP-scores may not reliably predict fine-tuned task performance. Consequently, it is possible that one of the other five BabyLMs I trained, which showed lower performance in the zero-shot evaluation, might have achieved stronger results on the GED task if selected and fine-tuned. This highlights the limitation of my BabyLM selection approach and points to the value of incorporating other factors into future BabyLM evaluation strategies.

Using BabyLMs as a Proxy for LLMs

One of the primary goals of this thesis was to use a BabyLM as a proxy for its LLM counterpart to study its performance in grammaticality assessment and a GED classification task over time. While this approach offers practical advantages such as reduced computational cost and interpretability, it presents some limitations. BabyLMs are trained on significantly smaller datasets and have fewer parameters than LLMs, which implies that they may not acquire the same linguistic representations. A particularly noteworthy factor is the limited vocabulary size, which may have influenced the model's ability to detect certain grammatical errors. For instance, if key function words or infrequent syntactic constructions appear too sparsely—or not at all—in the training data, the BabyLM may fail to learn the cues necessary for identifying errors related to those items. This is particularly relevant for tasks like subject–verb-agreement or preposition errors, which often rely on subtle lexical and syntactic patterns. Moreover, while my evaluation datasets were filtered to ensure all words appeared in the BabyLM's training dataset, this filtering also narrows the scope of grammatical generalization being tested. Future work can investigate whether the frequency of specific words—or broader linguistic categories such as function words or syntactic constructions—in the pretraining data can be directly linked to downstream performance on GED tasks. Such an analysis could offer better guidance for dataset design and data selection in low-resource pretraining.

Overall, insights drawn from BabyLMs may not fully generalize to LLMs, especially in terms of subtle grammatical distinctions or complex error patterns. Therefore, while BabyLMs are explanatory, any conclusions drawn from them must be interpreted with caution on their LLM counterparts.

6.5.3 Quality of the Datasets

Another limitation that affected the training phase of the BabyLMs and the two evaluation phases is the quality of the training, fine-tuning, and evaluation datasets. They are briefly discussed in the following paragraphs.

The Strict-Small Training Dataset of the BabyLM Challenge

As it was stated, the strict-small dataset (9.96M words) of the first BabyLM Challenge was used to train multiple BabyLMs. Although one of the primary reasons to use this dataset was to lower the environmental impacts and computational costs, it may restrict the model's ability to capture deeper syntactic and semantic nuances, especially for the grammaticality assessment and GED tasks. Moreover, the training dataset itself introduces additional constraints. The strict-small dataset used in the first BabyLM Challenge consists primarily of transcribed speech and child-directed language, which differs markedly from the written texts typically produced by ESL learners. Since some grammatical constructions are more prevalent in written language and that the nature of grammatical errors in ESL writing differs from those in spoken or child-directed contexts (Biber, 1991), this domain mismatch could limit the relevance of my BabyLM for the tasks in second language acquisition. Despite these drawbacks, I decided to use the strict-small dataset to maintain consistency with the BabyLM Challenge baseline and to facilitate fair comparative analysis. However, future work could benefit from training on more domain-specific corpora, such as ESL writing samples, to improve task relevance and model applicability.

The BLiMP Dataset

Although the fine-tuning and evaluation datasets are taken from reliable datasets such as the BLiMP dataset (Warstadt et al., 2020) and the BEA-2019 shared task dataset (Bryant et al., 2019), they have some problematic samples or wrong gold labels. The BLiMP dataset consists of some minimal pairs that have "sentence-good" (representing a correct grammatical sentence) and "sentence-bad" (representing an incorrect grammatical sentence). Based on the results shown in Chapter 5, there are some problematic instances in this dataset. One of the observed problems is the presence of multiple mistakes in one sentence, out of which the annotators annotated one of them. For example, "Most ladies confuses that people." is "sentence-bad" for determiner-noun-agreement error. Its "sentence-good" is "Most ladies confuses that person." in which the determiner error is fixed. However, this sentence also has a subject-verb-agreement error and its correct version should have been "Most ladies confuse that person.". Although O-RoBERTa could detect its "SVA" error, it is considered as a misclassification because the gold label was "G" because of being a "sentence-good". Moreover, there are some instances where it seems that the "sentence-good" and "sentence-bad" are confused in their annotations. For instance, "Some actresses distract Michael." is labeled as "sentence-bad", while it is grammatically correct. On the other hand, its pair ("Some actresses distracts Michael.") is labeled as "sentence-good", while it has a subject-verb-agreement error.

Warstadt et al. (2020) state that the BLiMP dataset is constructed through the automatic generation of sentences using grammar templates designed by linguists. To ensure the reliability of the automatically assigned labels, they validate the annotations using crowd-sourced human evaluations. Based on the observed problematic samples, it seems that crowd-sourced human judgments may not be reliable and maybe more strict annotation guidelines is needed to have high-quality annotations. Another problematic issue with the BLiMP dataset is the fact that its sentences are generated. These generated samples sometimes have unnatural structures that do not represent human sentence formations. For example, "Some waiter should notice this mirror that Meredith isn't revealing that will disgust a lot of pedestrians." does not seem humanly and its vague structure may have confused the models. This issue may have affected the general performance of the models in their GED classification task.

Another limitation is from part of the fine-tuning and evaluation datasets for the prepositions. Since the original BLiMP dataset does not contain preposition error types, I created a BLiMP-style dataset for the preposition error types from the BEA-2019 dataset (explained in Section 3.1.3). Although the created BLiMP-style preposition dataset was randomly checked, it needs to be validated completely. A few problematic instances were observed in the error analysis of the fine-tuning on the GED classification task in Chapter 5. For example, the fragmentary structure of some samples such as "The truth about Shopping" may confuse the models in their GED classification performance. Or, "from." is another problematic sample, whose corrected version by the BEA-2019 dataset is ".", which is not a sentence but a punctuation. According to Bryant et al. (2019), the BEA-2019 dataset is built upon CoNLL-2014 shared task (Ng et al., 2014) and introduces a newly annotated dataset from the Write&Improve+LOCNESS corpus. While Bryant et al. (2019) state that both the new corpus and the CoNLL-2014 dataset are filtered and validated by human annotators, some problematic samples persist within this dataset. Based on these findings, it seems that more precise approaches are required to ensure the validity of the datasets.

6.6 Future Research

Based on the findings and limitations identified in the thesis, there are several directions for future research, some of which are discussed in this section. Future works can focus on the development and evaluation of more BabyLM variants with broader architectural and training design explorations. While this thesis limited the variation to vocabulary size and number of hidden layers, future studies could experiment with diverse configurations. Moreover, an exhaustive hyperparameter search such as adjustments to learning rates and batch sizes could further help in identifying BabyLMs that better balance linguistic competence with efficiency, especially with constrained training dataset like that of the strict-small.

Moreover, future research would benefit from extending the evaluation beyond zero-shot BLiMP performance and three grammatical error types. Incorporating a wider range of linguistic benchmarks would provide a more complete understanding of BabyLM capabilities. This includes covering a broader range of grammatical error types and learner language phenomena, ideally sourced from more representative written ESL corpora rather than child-directed speech or transcribed dialogues. In addition, comparative fine-tuning experiments across multiple random seeds and epochs should be conducted not only for the BabyLM but also for BB-RoBERTa and O-RoBERTa to ensure fair and reliable comparisons. Finally, while BabyLMs offer a promising route for low-resource and explanatory experimentation, future work should also investigate the extent to which their learning dynamics align with those of LLMs.

Chapter 7

Conclusion

The objective of this thesis is to study the feasibility of using a RoBERTa-based BabyLM—a smaller and more resource-efficient language model—for grammaticality assessment and the GED task. In contrast to LLMs like RoBERTa, BabyLMs are trained on significantly smaller datasets, offering a promising path toward computational and environmental sustainability in NLP. However, their performance in some NLP tasks have not yet been fully explored, especially within the context of tasks such as grammaticality assessment and GED.

In the first phase of my experiment, six BabyLMs varying in vocabulary sizes and number of hidden layers are trained. The other parameters such as RoBERTa architecture and training objectives are fixed. All the BabyLMs are trained with the training dataset of the strict-small track (9.96M tokens) of the first BabyLM Challenge, introduced in 2023. Since there is no definition for a BabyLM's number of parameters, I reviewed the configurations of submitted papers that used RoBERTa architecture and competed in the strict-small track in the first BabyLM Challenge. The submitted models for the BabyLM Challenge span from 0.75M to 125M number of parameters. Since the winning model in the strict-small track of the BabyLM Challenge used 24M parameters, this value was adopted as the upper bound for the number of parameter in training the BabyLMs. To choose the promising model among these six BabyLMs, their BLiMP scores were compared against that of BB-RoBERTa's, which is a baseline BabyLM with RoBERTa architecture introduced by the first BabyLM Challenge. Among the six BabyLMs, the one with 20K vocabulary size and four hidden layers showed the highest and closest performance to BB-RoBERTa. This BabyLM was used for the following phases of the thesis experiments including the zero-shot evaluation and fine-tuning for the GED classification.

The thesis focused on three common grammatical error types among ESL learners including determiners, prepositions, and subject-verb-agreement to study the grammaticality assessment and the GED task. Two experimental settings are employed: (1) zero-shot evaluation using the BLiMP dataset along with the BLiMP-style dataset for preposition errors, which I created for this purpose, and (2) fine-tuned classification of the three targeted error types and the correct grammatical sentences. The performance of the selected BabyLM was compared to that of BB-RoBERTa and O-RoBERTa (the original RoBERTa model) in both settings.

To study the grammaticality assessment of the BabyLM over time, each training stage (epoch) of the BabyLM is saved and evaluated on the BLiMP dataset along with the BLiMP-style dataset for the preposition errors. The performance of the BabyLM's training stages are compared with those of BB-RoBERTa and O-RoBERTa in a zero-shot setting. This phase of experiment anwers the first sub-question of the thesis: How do these models compare in a zero-shot setting for the three targeted error types? Based on the zero-shot evaluation results, the BabyLM's performance in the zero-shot setting is lower than BB-RoBERTa, and much lower than O-RoBERTa. The results show an increasing performance trajectory from epochs one to four, after which the BabyLM's performance seems to plateau. The BabyLM's highest performance is at its eighth and ninth training stages. The zero-shot evaluation results indicate that the BabyLM learns determiner errors better than preposition and subject-verb-agreement error types. Among the three targeted errors, subject-verb-agreement seems the most challenging, as indicated by the lowest BLiMP scores.

The last phase of my thesis experiment involved fine-tuning the models for the GED classification task of labeling sentences either as grammatical or as containing one of the three targeted error types. First, the BabyLM is fine-tuned at its multiple training stages (per epoch) to study its performance over time. Then, BB-RoBERTa and O-RoBERTa are fine-tuned to compare their performance against those fine-tuned stages of the BabyLM. This phase of the experiment responds to the second sub-question of the thesis: How do these models compare in a fine-tuned setting for the three targeted error types? The results show that, once fine-tuned, the BabyLM achieves a level of performance comparable to those of BB-RoBERTa and O-RoBERTa on the GED task. While O-RoBERTa shows slightly better overall scores, the fine-tuned BabyLM narrows the performance gap significantly, demonstrating strong capability despite its smaller size.

Overall, all the fine-tuned models could classify the determiners easily, but misclassified some instances within the other error types. Most of the misclassifications occurred between grammatical sentences and those containing prepositional errors, and vice versa. Unlike the developmental findings of Dulay and Burt (1974) who indicate that the ESL learners generally learn prepositions earlier than the determiners or subject-verb-agreement grammar, the fine-tuned BabyLM shows difficulty in detecting the preposition error types and better classifies the determiner errors. This pattern is consistent across all three models, indicating that preposition errors are harder to detect in this classification setting, possibly due to their semantic and syntactic variability.

In conclusion, the results obtained from both the zero-shot and fine-tuned settings provide an answer to the main research question of the thesis, which investigates whether a RoBERTa-based BabyLM model can offer competitive grammaticality judgments in comparison to RoBERTa-base, specifically across the three targeted error types. Even though the BabyLM shows quite difference in the zero-shot evaluation in comparison with BB-RoBERTa and O-RoBERTa, its performance gets closer when it is fine-tuned. Overall, after fine-tuning, the BabyLM shows a competitive performance relative to the fine-tuned O-RoBERTa on the GED classification task. While there is a slight performance gap between the BabyLM fine-tuned at epoch ten and the fine-tuned O-RoBERTa model, it is important to note that the BabyLM is trained with approximately ten times fewer parameters (13M vs. 125M), roughly 3,000 times less training data (52MB vs. 160GB), and significantly reduced environmental impact and computational cost. These findings suggest that the BabyLM represents a competitive alternative to its larger language model counterpart for similar GED tasks. However, this conclusion is limited to the three targeted error types examined, and its effectiveness may not generalize to other categories of grammatical errors.

Appendix A

Additional Tables

Error Type	**Count**
Preposition	3001
anaphor_agreement	231
argument_structure	786
binding	798
control_raising	532
determiner_noun_agreement	5297
ellipsis	203
filler_gap_dependency	763
irregular_forms	228
island_effects	284
npi_licensing	754
quantifiers	458
s-selection	204
subject_verb_agreement	3931

Table A.1: The table shows the distribution of error types for the grammatical class ("G") in the finetuning dataset. As it is explained in Section 3.1.3, 30% of the G class is randomly selected from other error types in the filtered BLiMP dataset. This approach helps to study if the models can distinguish correct grammatical sentences with some grammatical structures other than the three targeted error types.

	Vocab-Size=20k				Vocab-Size=40k				Vocab-Size=50k				BB-RoBERTa
	Layer=4		Layer=6		Layer=4		Layer=6		Layer=4		Layer=6		
	Epoch=5	Epoch=10	Epoch=5	Epoch=10	Epoch=5	Epoch=10	Epoch=5	Epoch=10	Epoch=5	Epoch=10	Epoch=5	Epoch=10	Epoch=20
anaphor_agreement	64.62	64.98	69.68	66.31	69.43	69.38	66.72	62.12	72.39	69.99	70.91	63.55	81.54
argument_structure	61.47	62.43	59.88	59.69	62.48	63.54	61.93	62.38	62.75	62.57	63.07	62.91	67.12
binding	60.57	61.59	61.95	62.67	61.46	62.72	64.14	62.94	63.28	61.92	61.83	62.10	67.26
control_raising	57.78	59.77	59.37	59.63	60.65	60.30	61.20	61.27	62.64	62.44	61.18	61.60	67.85
determiner_noun_agreement	75.83	77.29	75.83	77.14	75.36	74.87	76.62	76.32	74.24	74.25	76.52	76.50	90.75
ellipsis	45.55	52.66	50.69	57.51	46.13	52.02	46.82	49.19	47.58	52.71	45.73	53.70	76.44
filler_gap	60.86	60.26	63.01	61.95	62.14	61.55	63.62	63.54	62.54	61.83	62.08	63.38	63.48
irregular_forms	86.72	84.68	84.58	82.60	88.70	84.07	83.72	79.24	83.21	82.44	85.04	80.76	87.43
island_effects	39.24	44.51	41.44	44.36	39.65	43.24	38.75	40.88	41.93	39.24	42.56	44.06	39.87
npi_licensing	56.71	59.52	54.57	51.69	46.05	48.79	48.33	47.98	50.00	43.99	51.18	55.04	55.92
quantifiers	65.71	67.62	65.89	67.90	75.04	71.92	73.80	73.96	71.05	70.66	63.68	62.83	70.53
subject_verb_agreement	54.42	55.01	54.60	55.34	54.07	55.18	54.72	57.27	53.80	55.23	56.62	56.60	65.42
Mean (All BLiMP)	60.79	**62.52**	61.80	62.23	**63.18**	62.30	61.70	61.42	62.12	61.44	61.70	61.91	**69.46**
Mean (DET and SVA)	65.12	66.15	65.21	66.24	64.71	65.02	65.67	66.79	64.02	64.74	66.57	66.55	78.08

Table A.2: The table presents the evaluation results of the six BabyLMs that are trained on the first BabyLM Challenge's BLiMP dataset in Comparison with BB-RoBERTa. This belongs to Section 4.1 where three tables separately display the results of the six BabyLMs. For a better comparison, all the results are put in one table.

	Vocab = 20k, Layer = 4, Epoch = 10						Vocab = 40k, Layer = 4, Epoch = 5						BB-RoBERTa
	Seed=42	Seed=56	Seed=162	Seed=354	Seed=670	Seed=1278	Seed=42	Seed=56	Seed=162	Seed=354	Seed=670	Seed=1278	Epoch=20
anaphor_agreement	64.98	66.16	62.78	61.55	68.35	65.29	69.43	69.84	69.02	62.63	77.75	62.68	81.54
argument_structure	62.43	60.67	59.53	61.14	61.30	60.92	62.48	62.60	62.80	61.83	62.84	61.92	67.12
binding	61.59	60.27	60.92	64.38	62.50	63.73	61.46	63.43	63.42	63.92	61.53	63.24	67.26
control_raising	59.77	56.94	59.66	59.30	60.08	60.83	60.65	60.05	61.31	59.94	61.31	61.07	67.85
determiner_noun_agreement	77.29	79.16	74.54	77.57	74.85	78.22	75.36	74.65	74.53	75.54	76.07	74.09	90.75
ellipsis	52.66	50.87	56.47	53.75	51.10	50.17	46.13	49.83	45.67	45.21	49.65	45.50	76.44
filler_gap	60.26	60.82	63.60	65.22	60.36	63.73	62.14	63.48	62.96	65.53	61.75	62.76	63.48
irregular_forms	84.68	78.02	81.93	83.41	80.10	82.95	88.70	85.55	82.54	72.82	84.68	69.97	87.43
island_effects	44.51	41.70	46.15	44.28	40.36	42.26	39.65	44.77	42.86	41.78	44.06	39.57	39.87
npi_licensing	59.52	46.46	50.15	54.11	58.94	62.25	46.05	55.30	57.01	50.94	56.35	53.81	55.92
quantifiers	67.62	64.68	71.35	79.16	60.54	74.57	75.04	70.17	72.020	63.60	72.75	62.73	70.53
subject_verb_agreement	55.01	55.68	53.86	55.00	56.57	58.43	54.07	54.54	54.54	53.19	52.18	54.47	65.42
Mean (All BLiMP)	62.52	60.12	61.75	63.24	61.25	63.61	63.18	62.85	62.39	59.74	63.41	59.32	**69.46**
Mean (DET and SVA)	66.15	67.42	64.20	66.29	65.71	68.33	64.71	64.60	64.54	64.37	64.13	64.28	**78.08**
Seed Average (Total)	-			**62.00**			-			**61.54**			-
Seed Average (DET and SVA)	-			**64.60**			-			**64.38**			-

Table A.3: The table presents the BLiMP-Score performance across multiple seeds for BabyLM-V20-L4 and BabyLM-V40-L4 in comparison with the performance of BB-RoBERTa and that of the default seed's (42). This table belongs to Section 4.1 where I run and experiment to ensure the BabyLMs' performance differences are not because of random initialization. For a more reliable comparison, I train and evaluate these two BabyLMs across five random seeds 56, 162, 354, 670, 1278) to compare their average performance with the default random seed that I do my zero-shot evaluations with. This comparison acts like a sanity check for my default random seed. The results are separately displayed for each BabyLM in Table 4.4 and Table 4.5 in Chapter 4.

	The BabyLM										BB-RoBERTa	O-RoBERTa
Category	**E=1**	**E=2**	**E=3**	**E=4**	**E=5**	**E=6**	**E=7**	**E=8**	**E=9**	**E=10**		
anaphor_agreement	44.63	58.74	60.48	61.81	63.85	65.75	67.74	67.59	66.56	64.98	81.54	90.70
argument_structure	60.81	59.93	60.18	60.90	61.60	61.91	62.26	62.00	62.08	62.43	67.12	83.04
binding	61.29	60.00	60.15	60.88	60.79	60.54	60.60	61.07	61.25	61.59	67.26	79.18
control_raising	58.44	58.02	58.24	57.95	58.59	58.84	59.37	59.32	59.99	59.77	67.85	81.95
determiner_noun_agreement	51.99	61.83	70.95	75.07	75.93	76.98	78.03	**78.37**	78.16	77.29	**90.75**	**97.28**
ellipsis	23.61	34.24	42.21	46.19	48.04	49.48	50.23	51.50	52.08	52.66	76.44	92.15
filler_gap	62.81	60.04	60.83	60.99	60.96	60.78	60.58	60.04	59.77	60.26	63.48	89.39
irregular_forms	61.27	82.19	85.80	87.38	88.09	87.38	86.67	86.11	85.29	84.68	87.43	95.67
island_effects	49.63	45.70	43.31	41.37	41.59	42.41	43.27	43.98	43.80	44.51	39.87	79.71
npi_licensing	42.03	41.09	47.07	54.33	56.98	57.96	57.82	58.43	58.96	59.52	55.92	82.61
quantifiers	44.28	53.09	61.39	64.61	67.34	68.13	68.88	69.78	68.26	67.62	70.53	70.79
subject_verb_agreement	50.46	51.83	52.72	53.01	53.93	54.58	55.00	**55.52**	**55.52**	55.01	**65.42**	**91.47**
preposition	49.75	55.85	57.85	59.24	60.63	61.88	62.04	61.65	**62.24**	61.99	**73.08**	**91.28**
Mean (All BLiMP)	50.85	55.58	58.55	60.29	61.41	62.04	62.50	**62.72**	62.61	62.49	**73.59**	**86.55**
Mean (DET, SVA, PREP)	50.73	56.50	60.51	62.44	63.50	64.48	65.02	65.18	**65.30**	64.76	**76.42**	**93.34**

Table A.4: The table shows the zero-shot evaluation of the BabyLM in comparison with those of BB-RoBERTa and O-RoBERTa. The zero-shot evaluation is done across the BabyLM's epochs, which represent its training stages. The BabyLM refers to "BabyLM-V20-L4", which is the best BabyLM among the six trained BabyLMs (as reported in Section4.1. The table belongs to Section 4.2 where a selection of the model's training stages and the results of BB-RoBERTa's and O-RoBERTa's are displayed in Table 4.6 for an analysis. "E" represents the BabyLM's epoch number.

	The BabyLM at Epoch 1		**The BabyLM at Epoch 10**	
Train Data	**50%**	**100%**	**50%**	**100%**
1 Epoch	0.59031705678386	0.5869398245064774	0.7119173577904299	0.7370329045762579
2 Epochs	0.6620179109912859	0.8035347774653994	0.8260227762585975	0.8906185760119897
3 Epochs	0.6903315633724443	0.8690032260433115	0.8143435574424752	0.9090168753864079
4 Epochs	0.763186098734778	0.8792684755627849	0.8380354860635373	0.8958703090052436
5 Epochs	0.7538632791528652	0.8645069805680621	0.8422210119242948	**0.9056979753954882**
6 Epochs	0.791461875827684	**0.8821109348523397**	0.7832795664805713	0.8896873347700005
7 Epochs	**0.8043349759627726**	0.8689473962820986	**0.8499169900421547**	0.8799197755646971
8 Epochs	0.7604117339714573	0.8309469950106507	0.8172135137560359	0.8439704296950851
9 Epochs	0.7910931308693561	0.8309543708959224	0.8228752684361242	0.8528419464543615
10 Epochs	0.7836947995405191	0.841388851916296	0.8209056798376733	0.845804050758941

Table A.5: The table illustrates the experiment to decide about the epoch numbers for fine-tuning the models (explained in Section 4.3). The BabyLM's epochs one and ten are fine-tuned up to ten epochs with the training subset of the fine-tuning dataset. "BabyLM at Epoch 1" refers to the fine-tuned BabyLM at its first training stage (epoch). On the other hand, "BabyLM at Epoch 10" refers to the fine-tuned BabyLM at its last training stage (epoch). The experiment is done twice across 50% and 100% of the training data. The results are macro-average (F0.5-score) on the DEV dataset.

Appendix B

Additional Results

B.1 Multiple Seeds Experiment in the Fine-tuning Phase

This experiment belongs to the fine-tuning phase of my thesis experiments, explained in Section 4.3. The average score of multiple random seeds (20, 58, 150, 342, 613) are compared with the main random seed (464), with which all the fine-tuning phase is done. This comparison can help check if the main seed is not a bad seed. The mean of the macro avg f0.5-score of these random seeds is shown in Table 4.9 in Chapter 4. This additional experiment is done at epochs one, five, and ten of the BabyLM. This choice gives a view of fine-tuning the BabyLM for the GED classification task at its initial, middle, and final epochs, which represent the BabyLM's training stages.

GED Evaluation of the BabyLM (with Random Seed of 20) at Epoch One:

	precision	**recall**	**f0.5-score**	**support**
DET	0.985	0.998	0.988	1508
G	0.909	0.782	0.880	3494
PREP	0.563	0.873	0.606	879
SVA	0.863	0.850	0.861	1107
accuracy			0.851	6988
macro avg	0.830	0.876	0.833	6988
weighted avg	0.874	0.851	0.866	6988

Table B.1: The Classification Report of the Results for the BabyLM at Epoch One Fine-Tuned with Seed 20 on the GED Classification Task.

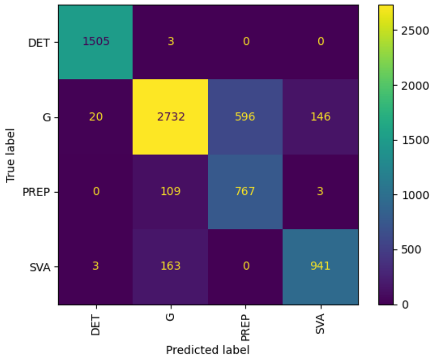

Figure B.1: The Confusion Matrix of the Results for the BabyLM at Epoch One Fine-Tuned with Seed 20 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 20) at Epoch Five:

	precision	recall	f0.5-score	support
DET	0.997	0.999	0.997	1508
G	0.938	0.813	0.910	3494
PREP	0.558	0.843	0.599	879
SVA	0.945	0.955	0.947	1107
accuracy			0.880	6988
macro avg	0.859	0.903	0.863	6988
weighted avg	0.904	0.880	0.896	6988

Table B.2: The Classification Report of the Results for the BabyLM at Epoch Five Fine-Tuned with Seed 20 on the GED Classification Task.

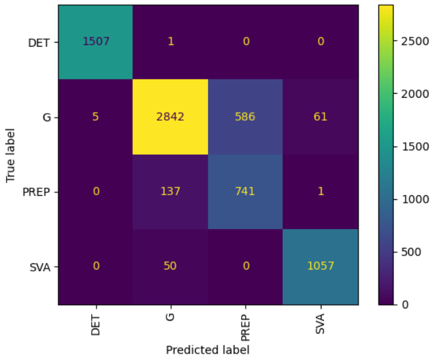

Figure B.2: The Confusion Matrix of the Results for the BabyLM at Epoch Five Fine-Tuned with Seed 20 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 20) at Epoch Ten:

	precision	recall	f0.5-score	support
DET	0.995	0.999	0.996	1508
G	0.918	0.810	0.894	3494
PREP	0.551	0.812	0.588	879
SVA	0.929	0.920	0.927	1107
accuracy			0.868	6988
macro avg	0.848	0.885	0.851	6988
weighted avg	0.890	0.868	0.883	6988

Table B.3: The Classification Report of the Results for the BabyLM at Epoch Ten Fine-Tuned with Seed 20 on the GED Classification Task.

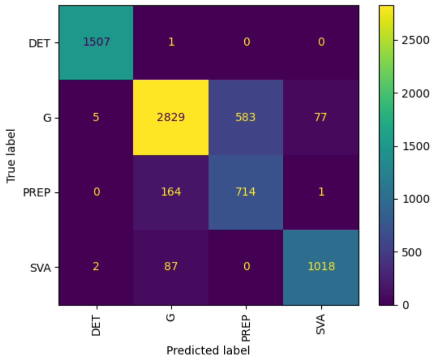

Figure B.3: The Confusion Matrix of the Results for the BabyLM at Epoch Ten Fine-Tuned with Seed 20 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 58) at Epoch One:

	precision	recall	f0.5-score	support
DET	0.989	0.999	0.991	1508
G	0.910	0.775	0.879	3494
PREP	0.564	0.892	0.609	879
SVA	0.849	0.842	0.847	1107
accuracy			0.849	6988
macro avg	0.828	0.877	0.832	6988
weighted avg	0.874	0.849	0.864	6988

Table B.4: The Classification Report of the Results for the BabyLM at Epoch One Fine-Tuned with Seed 58 on the GED Classification Task.

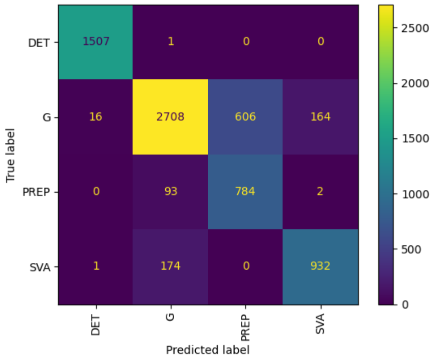

Figure B.4: The Confusion Matrix of the Results for the BabyLM at Epoch One Fine-Tuned with Seed 58 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 58) at Epoch Five:

	precision	recall	f0.5-score	support
DET	0.995	1.000	0.996	1508
G	0.943	0.812	0.913	3494
PREP	0.563	0.876	0.607	879
SVA	0.948	0.939	0.946	1107
accuracy			0.881	6988
macro avg	0.862	0.907	0.866	6988
weighted avg	0.907	0.881	0.898	6988

Table B.5: The Classification Report of the Results for the BabyLM at Epoch Five Fine-Tuned with Seed 58 on the GED Classification Task.

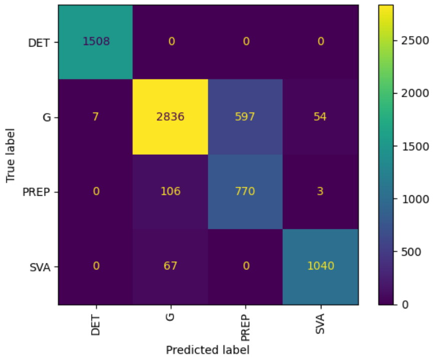

Figure B.5: The Confusion Matrix of the Results for the BabyLM at Epoch Five Fine-Tuned with Seed 58 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 58) at Epoch Ten:

	precision	recall	f0.5-score	support
DET	0.996	1.000	0.997	1508
G	0.925	0.810	0.899	3494
PREP	0.553	0.832	0.592	879
SVA	0.933	0.920	0.930	1107
accuracy			0.871	6988
macro avg	0.852	0.890	0.855	6988
weighted avg	0.895	0.871	0.887	6988

Table B.6: The Classification Report of the Results for the BabyLM at Epoch Ten Fine-Tuned with Seed 58 on the GED Classification Task.

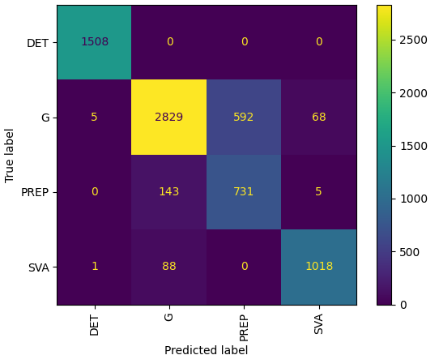

Figure B.6: The Confusion Matrix of the Results for the BabyLM at Epoch Ten Fine-Tuned with Seed 58 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 150) at Epoch One:

	precision	recall	f0.5-score	support
DET	0.990	1.000	0.992	1508
G	0.905	0.793	0.880	3494
PREP	0.553	0.805	0.590	879
SVA	0.874	0.887	0.877	1107
accuracy			0.854	6988
macro avg	0.830	0.871	0.835	6988
weighted avg	0.874	0.854	0.867	6988

Table B.7: The Classification Report of the Results for the BabyLM at Epoch One Fine-Tuned with Seed 150 on the GED Classification Task.

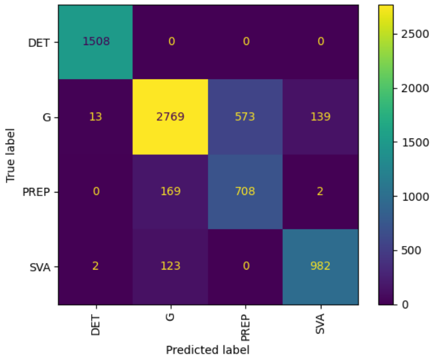

Figure B.7: The Confusion Matrix of the Results for the BabyLM at Epoch One Fine-Tuned with Seed 150 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 150) at Epoch Five:

	precision	recall	f0.5-score	support
DET	0.995	1.000	0.996	1508
G	0.930	0.812	0.904	3494
PREP	0.554	0.818	0.592	879
SVA	0.937	0.951	0.940	1107
accuracy			0.876	6988
macro avg	0.854	0.895	0.858	6988
weighted avg	0.898	0.876	0.890	6988

Table B.8: The Classification Report of the Results for the BabyLM at Epoch Five Fine-Tuned with Seed 150 on the GED Classification Task.

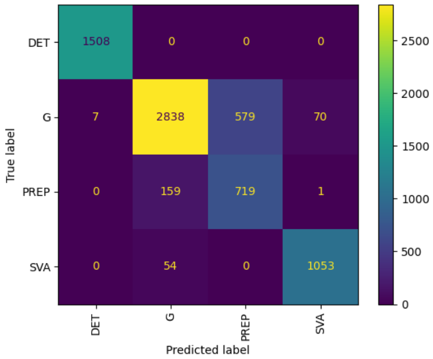

Figure B.8: The Confusion Matrix of the Results for the BabyLM at Epoch Five Fine-Tuned with Seed 150 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 150) at Epoch Ten:

	precision	recall	f0.5-score	support
DET	0.993	1.000	0.995	1508
G	0.911	0.819	0.891	3494
PREP	0.548	0.785	0.584	879
SVA	0.944	0.915	0.938	1107
accuracy			0.869	6988
macro avg	0.849	0.880	0.852	6988
weighted avg	0.888	0.869	0.882	6988

Table B.9: The Classification Report of the Results for the BabyLM at Epoch Ten Fine-Tuned with Seed 150 on the GED Classification Task.

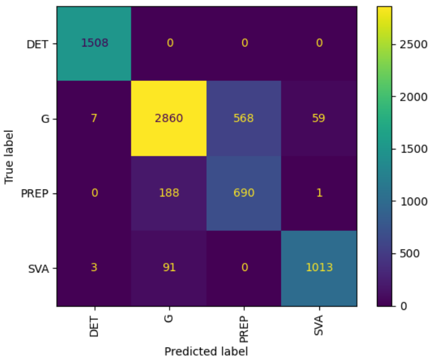

Figure B.9: The Confusion Matrix of the Results for the BabyLM at Epoch Ten Fine-Tuned with Seed 150 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 342) at Epoch One:

	precision	recall	f0.5-score	support
DET	0.988	1.000	0.990	1508
G	0.924	0.787	0.893	3494
PREP	0.569	0.901	0.614	879
SVA	0.881	0.870	0.879	1107
accuracy			0.860	6988
macro avg	0.840	0.890	0.844	6988
weighted avg	0.886	0.860	0.877	6988

Table B.10: The Classification Report of the Results for the BabyLM at Epoch One Fine-Tuned with Seed 342 on the GED Classification Task.

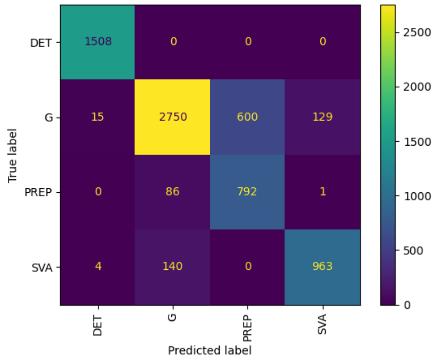

Figure B.10: The Confusion Matrix of the Results for the BabyLM at Epoch One Fine-Tuned with Seed 342 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 342) at Epoch Five:

	precision	recall	f0.5-score	support
DET	0.993	1.000	0.994	1508
G	0.924	0.821	0.902	3494
PREP	0.556	0.811	0.593	879
SVA	0.959	0.937	0.955	1107
accuracy			0.877	6988
macro avg	0.858	0.892	0.861	6988
weighted avg	0.898	0.877	0.891	6988

Table B.11: The Classification Report of the Results for the BabyLM at Epoch Five Fine-Tuned with Seed 342 on the GED Classification Task.

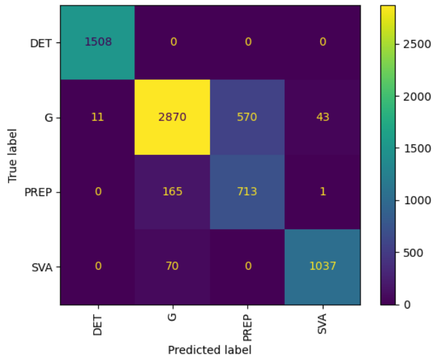

Figure B.11: The Confusion Matrix of the Results for the BabyLM at Epoch Five Fine-Tuned with Seed 342 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 342) at Epoch Ten:

	precision	recall	f0.5-score	support
DET	0.995	0.999	0.996	1508
G	0.925	0.811	0.900	3494
PREP	0.556	0.833	0.595	879
SVA	0.935	0.923	0.933	1107
accuracy			0.872	6988
macro avg	0.853	0.892	0.856	6988
weighted avg	0.895	0.872	0.887	6988

Table B.12: The Classification Report of the Results for the BabyLM at Epoch Ten Fine-Tuned with Seed 342 on the GED Classification Task.

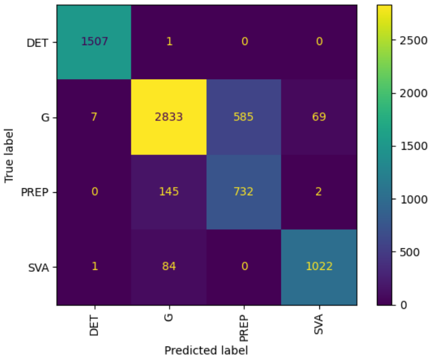

Figure B.12: The Confusion Matrix of the Results for the BabyLM at Epoch Ten Fine-Tuned with Seed 342 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 613) at Epoch One:

	precision	recall	f0.5-score	support
DET	0.988	1.000	0.990	1508
G	0.897	0.770	0.869	3494
PREP	0.558	0.843	0.598	879
SVA	0.821	0.839	0.824	1107
accuracy			0.840	6988
macro avg	0.816	0.863	0.820	6988
weighted avg	0.862	0.840	0.854	6988

Table B.13: The Classification Report of the Results for the BabyLM at Epoch One Fine-Tuned with Seed 613 on the GED Classification Task.

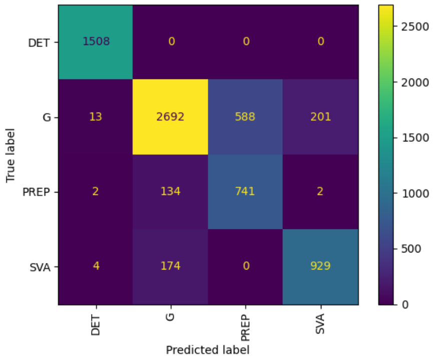

Figure B.13: The Confusion Matrix of the Results for the BabyLM at Epoch One Fine-Tuned with Seed 613 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 613) at Epoch Five:

	precision	recall	f0.5-score	support
DET	0.995	1.000	0.996	1508
G	0.940	0.814	0.912	3494
PREP	0.559	0.862	0.601	879
SVA	0.958	0.944	0.955	1107
accuracy			0.881	6988
macro avg	0.863	0.905	0.866	6988
weighted avg	0.907	0.881	0.898	6988

Table B.14: The Classification Report of the Results for the BabyLM at Epoch Five Fine-Tuned with Seed 613 on the GED Classification Task.

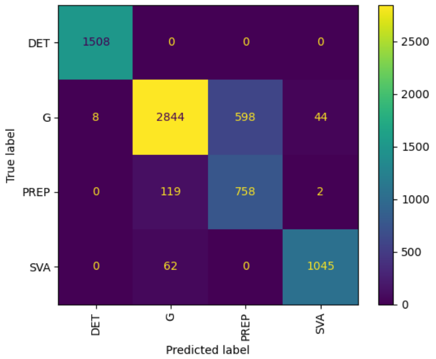

Figure B.14: The Confusion Matrix of the Results for the BabyLM at Epoch Five Fine-Tuned with Seed 613 on the GED Classification Task.

GED Evaluation of the BabyLM (with Random Seed of 613) at Epoch Ten:

	precision	recall	f0.5-score	support
DET	0.994	1.000	0.995	1508
G	0.930	0.809	0.903	3494
PREP	0.558	0.843	0.599	879
SVA	0.929	0.927	0.928	1107
accuracy			0.873	6988
macro avg	0.853	0.895	0.856	6988
weighted avg	0.897	0.873	0.888	6988

Table B.15: The Classification Report of the Results for the BabyLM at Epoch Ten Fine-Tuned with Seed 613 on the GED Classification Task.

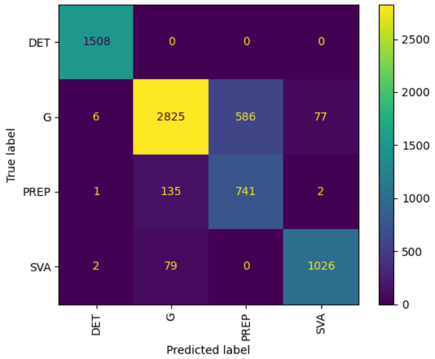

Figure B.15: The Confusion Matrix of the Results for the BabyLM at Epoch Ten Fine-Tuned with Seed 613 on the GED Classification Task.

B.2 The Results of Fine-tuning the BabyLM's Remaining Pretraining Epochs on the GED Task

As stated in Section 4.3 of Chapter 4, all ten of the BabyLM's pretraining epochs were fine-tuned on the GED classification task using the main seed 464, which was randomly selected. However, only a subset of these epochs (epochs one, three, five, nine, and ten) were included in the main analysis. This section presents the classification reports and confusion matrices for the remaining fine-tuned epochs of the BabyLM (epochs two, four, six, seven, and eight).

The GED Evaluation of Fine-tuning the BabyLM Pretrained for Two Epochs:

	precision	recall	f0.5-score	support
DET	0.988	1.000	0.990	1508
G	0.908	0.821	0.889	3494
PREP	0.556	0.778	0.590	879
SVA	0.940	0.912	0.935	1107
accuracy			0.869	6988
macro avg	0.848	0.878	0.851	6988
weighted avg	0.886	0.869	0.881	6988

Table B.16: The Classification Report of the Results for Fine-tuning the BabyLM Pretrained for Two Epochs on the GED Classification Task.

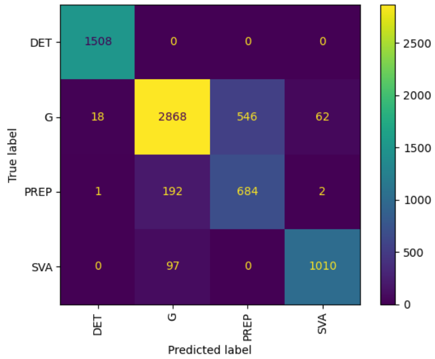

Figure B.16: The Confusion Matrix of Fine-tuning the BabyLM Pretrained for Two Epochs on the GED Classification Task.

The GED Evaluation of Fine-tuning the BabyLM Pretrained for Four Epochs:

	precision	**recall**	**f0.5-score**	**support**
DET	0.991	1.000	0.993	1508
G	0.922	0.824	0.900	3494
PREP	0.553	0.782	0.587	879
SVA	0.955	0.948	0.954	1107
accuracy			0.876	6988
macro avg	0.855	0.888	0.858	6988
weighted avg	0.895	0.876	0.889	6988

Table B.17: The Classification Report of the Results for Fine-tuning the BabyLM Pretrained for Four Epochs on the GED Classification Task.

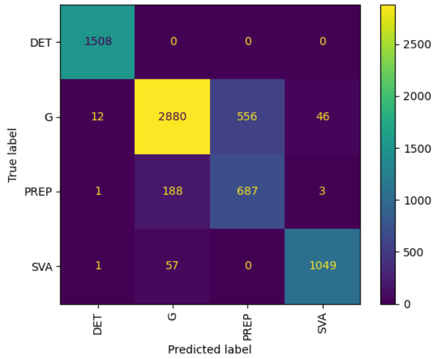

Figure B.17: The Confusion Matrix of Fine-tuning the BabyLM Pretrained for Four Epochs on the GED Classification Task.

The GED Evaluation of Fine-tuning the BabyLM Pretrained for Six Epochs:

	precision	recall	f0.5-score	support
DET	0.995	1.000	0.996	1508
G	0.920	0.828	0.900	3494
PREP	0.553	0.776	0.587	879
SVA	0.960	0.948	0.957	1107
accuracy			0.878	6988
macro avg	0.857	0.888	0.860	6988
weighted avg	0.896	0.878	0.890	6988

Table B.18: The Classification Report of the Results for Fine-tuning the BabyLM Pretrained for Six Epochs on the GED Classification Task.

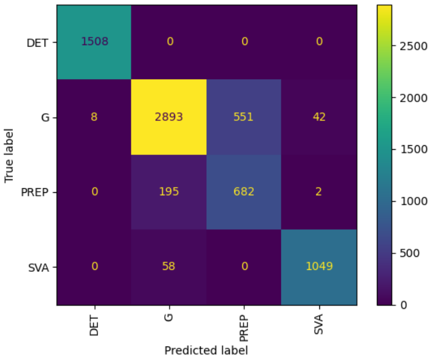

Figure B.18: The Confusion Matrix of Fine-tuning the BabyLM Pretrained for Six Epochs on the GED Classification Task.

The GED Evaluation of Fine-tuning the BabyLM Pretrained for Seven Epochs:

	precision	recall	f0.5-score	support
DET	0.994	1.000	0.995	1508
G	0.921	0.826	0.900	3494
PREP	0.549	0.776	0.583	879
SVA	0.962	0.950	0.960	1107
accuracy			0.877	6988
macro avg	0.856	0.888	0.860	6988
weighted avg	0.896	0.877	0.890	6988

Table B.19: The Classification Report of the Results for Fine-tuning the BabyLM Pretrained for Seven Epochs on the GED Classification Task.

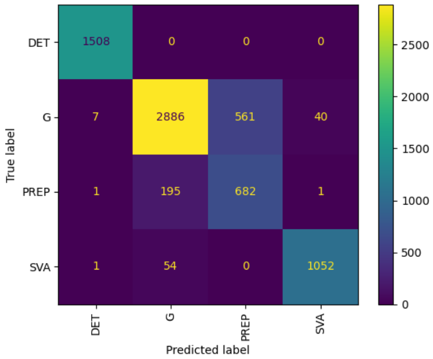

Figure B.19: The Confusion Matrix of Fine-tuning the BabyLM Pretrained for Seven Epochs on the GED Classification Task.

The GED Evaluation of Fine-tuning the BabyLM Pretrained for Eight Epochs:

	precision	recall	f0.5-score	support
DET	0.995	1.000	0.996	1508
G	0.920	0.823	0.899	3494
PREP	0.551	0.793	0.586	879
SVA	0.960	0.937	0.955	1107
accuracy			0.876	6988
macro avg	0.857	0.888	0.859	6988
weighted avg	0.896	0.876	0.890	6988

Table B.20: The Classification Report of the Results for Fine-tuning the BabyLM Pretrained for Eight Epochs on the GED Classification Task.

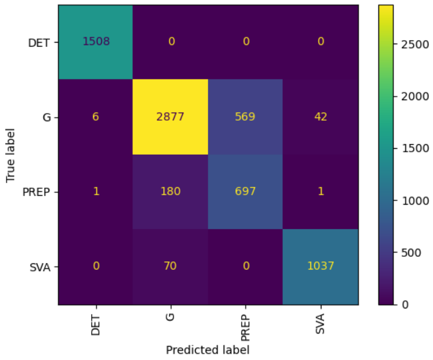

Figure B.20: The Confusion Matrix of Fine-tuning the BabyLM Pretrained for Eight Epochs on the GED Classification Task.

Bibliography

A. Abdelali, F. Guzman, H. Sajjad, and S. Vogel. 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)*, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).

A. Abu-Akel, A. L. Bailey, and Y.-M. Thum. Describing the acquisition of determiners in english: A growth modeling approach. *Journal of Psycholinguistic Research*, 33: 407–424, 2004.

E. M. Bender, T. Gebru, A. McMillan-Major, and 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*, pages 610–623. ACM, 2021. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

G. Berend. Better together: Jointly using masked latent semantic modeling and masked language modeling for sample efficient pre-training. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 298–307, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.26. URL https://aclanthology.org/2023.conll-babylm.26/.

D. Biber. *Variation Across Speech and Writing.* Cambridge University Press, 1991.

J. Bitchener, S. Young, and D. Cameron. The effect of different types of feedback on esl student writing. *Journal of Second Language Writing*, 14(3):191–205, 2005. URL https://www.researchgate.net/publication/222682464_The_effect_of_different_types_of_feedback_on_ESL_student_writing.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

C. Bryant, M. Felice, Ø. E. Andersen, and T. Briscoe. The BEA-2019 shared task on grammatical error correction. In H. Yannakoudakis, E. Kochmar, C. Leacock,

N. Madnani, I. Pilán, and T. Zesch, editors, *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4406. URL `https://aclanthology.org/W19-4406/`.

C. Bryant, Z. Yuan, M. R. Qorib, H. Cao, H. T. Ng, and T. Briscoe. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3): 643–701, 2023. doi: 10.1162/coli_a_00478. URL `https://doi.org/10.1162/coli_a_00478`.

B. Bunzeck and S. Zarrieß. GPT-wee: How small can a small language model really get? In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 35–46, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.2. URL `https://aclanthology.org/2023.conll-babylm.2/`.

M. Celce-Murcia and D. Larsen-Freeman. *The Grammar Book: An ESL/EFL Teacher's Course.* Heinle and Heinle, Boston, MA, 2nd edition, 1999. ISBN 978-0838447253. URL `https://flaviamcunha.files.wordpress.com/2013/03/the-grammar-book-an-eslefl-teachers-course-second-editiona4.pdf`.

G. G. L. Charpentier and D. Samuel. Not all layers are equally as important: Every layer counts BERT. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 238–252, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.20. URL `https://aclanthology.org/2023.conll-babylm.20/`.

L. Charpentier, L. Choshen, R. Cotterell, M. O. Gul, M. Y. Hu, J. Jumelet, T. Linzen, J. Liu, A. Mueller, C. Ross, R. S. Shah, A. Warstadt, E. Wilcox, and A. Williams. Babylm turns 3: Call for papers for the 2025 babylm workshop. *arXiv preprint arXiv:2502.10645*, 2025. URL `https://arxiv.org/abs/2502.10645`.

X. Chen and E. Portelance. Grammar induction pretraining for language modeling in low resource contexts. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 69–73, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.5. URL `https://aclanthology.org/2023.conll-babylm.5/`.

Z. Cheng, R. Aralikatte, I. Porada, C. Spinoso-Di Piano, and J. C. Cheung. McGill BabyLM shared task submission: The effects of data formatting and structural biases. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 207–220, Singapore, Dec. 2023. Association for

Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.18. URL https://aclanthology.org/2023.conll-babylm.18/.

N. Chomsky. *Syntactic Structures*. Walter de Gruyter, Berlin, 2nd edition, 2002.

R. Dale and A. Kilgarriff. Helping our own: The HOO 2011 pilot shared task. In C. Gardent and K. Striegnitz, editors, *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France, Sept. 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-2838/.

R. Dale, I. Anisimoff, and G. Narroway. HOO 2012: A report on the preposition and determiner error correction shared task. In J. Tetreault, J. Burstein, and C. Leacock, editors, *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-2006/.

G. Dalgish. Computer-assisted error analysis and courseware design: Applications for esl in the swedish context. *CALICO Journal*, 9, 1991.

W. de Vries, M. Wieling, and M. Nissim. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.529. URL https://aclanthology.org/2022.acl-long.529/.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020. URL https://arxiv.org/abs/2002.06305.

H. C. Dulay and M. K. Burt. Natural sequences in child second language acquisition. *Language Learning*, 24(1):37–53, 1974. doi: 10.1111/j.1467-1770.1974.tb00234.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-1770.1974.tb00234.x.

L. Edman and L. Bylinina. Too much information: Keeping training simple for BabyLMs. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 89–97, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.8. URL https://aclanthology.org/2023.conll-babylm.8/.

L. Edman, L. Bylinina, F. Ghorbanpour, and A. Fraser. Are BabyLMs second language learners?, 2024. URL http://arxiv.org/abs/2410.21254.

R. Eldan and Y. Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

N. Ellis and L. Collins. Input and second language acquisition: The roles of frequency, form, and function introduction to the special issue. *The Modern Language Journal*, 93:329 – 335, 09 2009. doi: 10.1111/j.1540-4781.2009.00893.x.

M. Gerlach and F. Font-Clos. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Computing Research Repository*, 2018.

D. Hiemstra. Language models. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, pages 1591–1594. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_923. URL https://doi.org/10.1007/978-0-387-39940-9_923.

F. Hill, A. Bordes, S. Chopra, and J. Weston. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016a.

F. Hill, A. Bordes, S. Chopra, and J. Weston. The goldilocks principle: Reading children's books with explicit memory representations, 2016b. URL https://arxiv.org/abs/1511.02301.

J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031/.

M. Y. Hu, A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, R. Cotterell, L. Choshen, A. Warstadt, and E. G. Wilcox. Findings of the second babylm challenge: Sample-efficient pretraining on developmentally plausible corpora, 2024. URL https://arxiv.org/abs/2412.05149.

P. A. Huebner and J. A. Willits. Using lexical context to discover the noun category: Younger children have it easier. In K. D. Federmeier and L. Sahakyan, editors, *The Context of Cognition: Emerging Perspectives*, volume 75 of *Psychology of Learning and Motivation*, pages 279–331. Academic Press, 2021.

P. A. Huebner, E. Sulem, F. Cynthia, and D. Roth. BabyBERTa: Learning more grammar with small-scale child-directed language. In A. Bisazza and O. Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.49. URL https://aclanthology.org/2021.conll-1.49/.

X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu. Tinybert: Distilling bert for natural language understanding, 2020. URL https://arxiv.org/abs/1909.10351.

V. A. Johnson, J. G. de Villiers, and H. N. Seymour. Agreement without understanding? the case of third person singular /s/. *First Language*, 25:317–330, 2005.

P. M. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In *Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://api.semanticscholar.org/CorpusID:215828350.

D. Jurafsky and J. H. Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models, 2025. URL https://web.stanford.edu/~jurafsky/slp3/. Online manuscript, released January 12, 2025, 3rd edition.

K. Killie. The acquisition of subject-verb agreement among norwegian (teenage) learners of english: Focus on the subject. *Linguists Journal of Linguistics and Language Teaching*, 10:183–206, 01 2020.

P. Lison and J. Tiedemann. 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, 2016. European Language Resources Association (ELRA).

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

B. MacWhinney. *The CHILDES Project: The Database, Volume 2.* Psychology Press, 2000.

A. Masciolini, A. Caines, O. De Clercq, J. Kruijsbergen, M. Kurfali, R. Muñoz Sánchez, E. Volodina, R. Östling, K. Allkivi-Metsoja, Š. Arhar Holdt, I. Auzina, R. Darğis, E. Drakonaki, J.-C. Frey, I. Glišić, P. Kikilintza, L. Nicolas, M. Romanyshyn, A. Rosen, A. Rozovskaya, K. Suluste, O. Syvokon, A. Tantos, D.-O. Touriki, K. Tsiotskas, E. Tsourilla, V. Varsamopoulos, K. Wisniewski, A. Žagar, and T. Zesch. Multigec, 2025. URL https://spraakbanken.gu.se/resurser/multigec.

K. Misra and K. Mahowald. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.53. URL https://aclanthology.org/2024.emnlp-main.53/.

A. Morgenstern and M. Sekali. What can child language tell us about prepositions? *Studies in language and cognition*, pages 261–275, 2009.

C. Napoles, K. Sakaguchi, and J. Tetreault. JFLEG: A fluency corpus and benchmark for grammatical error correction. In M. Lapata, P. Blunsom, and A. Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2037/.

National Center for Education Statistics. English learners in public schools, 2024. URL https://nces.ed.gov/programs/coe/indicator/cgf/english-learners. Accessed: 2025-01-25.

H. T. Ng, A. Kunchukuttan, H. L. Chieu, H. Mital, R. Das, C. Bryant, P. Cook, M. Hermet, B. Hladka, E. Kochmar, et al. The conll-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria, August 8–9 2013. Association for Computational Linguistics. URL https://www.comp.nus.edu.sg/~nlp/conll13st/CoNLLST01.pdf.

H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant. The CoNLL-2014 shared task on grammatical error correction. In H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant, editors, *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-1701. URL https://aclanthology.org/W14-1701/.

M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. *Advances in neural information processing systems*, 22, 2009.

M. Pienemann. An outline of processability theory and its relationship to other approaches to sla. *Language Learning*, 65(S1):123–151, 2015. doi: 10.1111/lang.12095. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12095.

I. Proskurina, G. Metzler, and J. Velcin. Mini minds: Exploring bebeshka and zlata baby models. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 58–68, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.4. URL https://aclanthology.org/2023.conll-babylm.4/.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

C. Rozema. Language models as second language learners: A second language acquisition-inspired curriculum learning approach to training a babylm. Research master thesis, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, Aug. 2024. Submitted in partial fulfilment of the requirements for the degree of ReMA Humanities (Human Language Technology). Supervised by Dr. Lucia Donatelli. 2nd reader: Dr. Luís de Passos Morgado da Costa. Submitted: August 15, 2024.

D. Samuel. Mean BERTs make erratic language teachers: the effectiveness of latent bootstrapping in low-resource settings. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 221–237, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.19. URL https://aclanthology.org/2023.conll-babylm.19/.

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units, 2016. URL https://arxiv.org/abs/1508.07909.

M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45:427–437, 07 2009. doi: 10.1016/j.ipm.2009.03.002.

S. H. Stapa and M. M. Izahar. Analysis of errors in subject-verb agreement among malaysian esl learners. *3L: The Southeast Asian Journal of English Language Studies*, 16(1):56–73, 2010. URL https://www.researchgate.net/publication/228514181_Analysis_of_errors_in_subject-verb_agreement_among_Malaysian_ESL_learners.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, M. Meteer, and C. V. Ess-Dykema. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–371, 2000.

I. Timiryasov and J.-L. Tastet. Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 279–289, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.24. URL https://aclanthology.org/2023.conll-babylm.24/.

Ö. Veysel Çağatan. ToddlerBERTa: Exploiting BabyBERTa for grammar learning and language understanding. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 171–179, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.14. URL https://aclanthology.org/2023.conll-babylm.14/.

E. Volodina, C. Bryant, A. Caines, O. De Clercq, J.-C. Frey, E. Ershova, A. Rosen, and O. Vinogradova. MultiGED-2023 shared task at NLP4CALL: Multilingual grammatical error detection. In D. Alfter, E. Volodina, T. François, A. Jönsson, and E. Rennes, editors, *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 1–16, Tórshavn, Faroe Islands, May 2023. LiU Electronic Press. URL https://aclanthology.org/2023.nlp4call-1.1/.

A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, and S. R. Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL https://doi.org/10.1162/tacl_a_00321.

A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore, Dec. 2023a. Association for

Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.1. URL https://aclanthology.org/2023.conll-babylm.1/.

A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors. *Proceedings of the CoNLL 2023 Shared Task: BabyLM*, Singapore, 2023b. Association for Computational Linguistics. URL https://aclanthology.org/volumes/2023.conll-babylm/.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

M. Wynne. British national corpus. http://www.natcorp.ox.ac.uk/, 2022. Accessed: October 2022.

Y. Yang, E. Sulem, I. Lee, and D. Roth. Penn & BGU BabyBERTa+ for strict-small BabyLM challenge. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 86–88, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.7. URL https://aclanthology.org/2023.conll-babylm.7/.

S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

Y. Zhong and X. Yue. On the correction of errors in english grammar by deep learning. *Journal of Intelligent Systems*, 31(1):260–270, 2022. ISSN 2191-026X. doi: 10.1515/jisys-2022-0013. URL https://www.degruyter.com/document/doi/10.1515/jisys-2022-0013/html. Publisher: De Gruyter.