Learning from Child-Directed Speech in Two-Language Scenarios: A Controlled English–French Study

Liel Binyamin, Elior Sulem
The Faculty of Computer and Information Science, Ben-Gurion University of the Negev, Israel

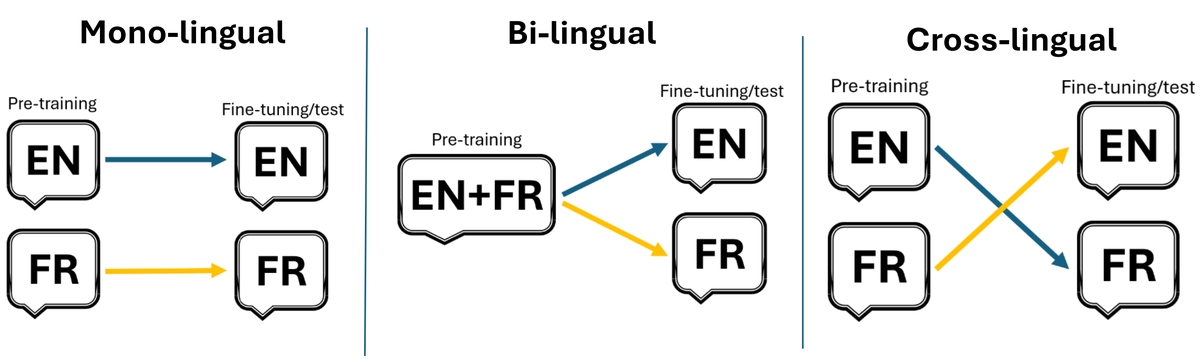
Ben-Gurion University of the Negev
Ministry of Innovation, Science & Technology
EACL 2026 RABAT • MOROCCO
Mars • March 24-29, 2026 • مارس

Motivation

- CDS-based language models have been studied primarily in monolingual English and mainly evaluated on grammatical competence.
- It remains unclear whether efficiency gains extend to multilingual, size-matched scenarios.
- A controlled comparison of monolingual, bilingual, and cross-lingual training across corpus types is still missing.

Controlled Experimental Design

Training Setups



Pre-training Corpora

Corpus Type	Size (tokens)	Comparison Purpose
CHILDES	2.5M	Human-scale CDS baseline
Wikipedia	2.5M	Matched semantic baseline
Multidomain	10M	Larger-scale natural data
Wikipedia	10M	Matched 10M semantic baseline

All pre-training corpora are strictly size-matched across languages (EN/FR) and training setups. Downstream fine-tuning is performed in the evaluation language for Question Answering and Textual Entailment tasks.

Evaluation Tasks

Task	Evaluation Dataset(s)	Metric
Question Answering (QA)	SQuAD, QAMR, QASRL	F1
Textual Entailment	XNLI	F1
Grammatical Competence	CLAMS	Accuracy

All tasks were evaluated in both English and French under identical fine-tuning protocols.

Models

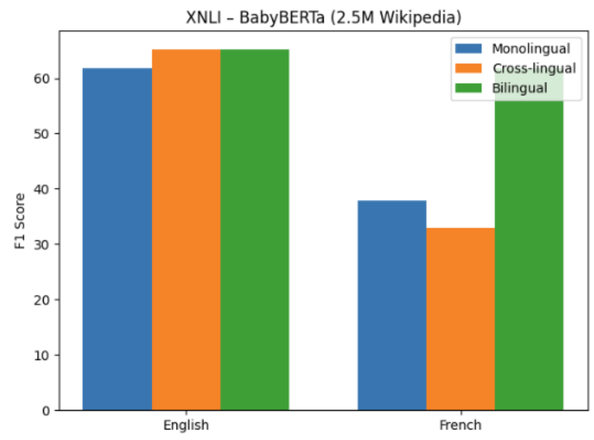
- BabyBERTa (main compact model)
- RoBERTa (retrained on matched data)
- LTG-BERT
- T5-tiny (analysis)

Research Questions

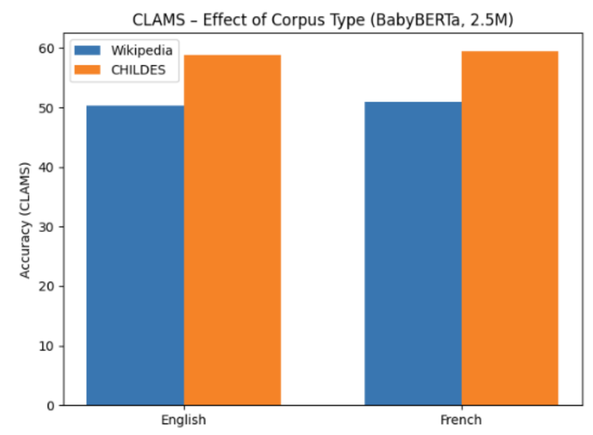
1. What are the effects of different language pairing strategies (monolingual, bilingual, cross-lingual)?
2. How does corpus type (CHILDES vs. Wikipedia vs. multi-domain) influence model competence?
3. Do multilingual effects persist across scale (2.5M → 10M tokens)?
4. Are the observed patterns consistent across architectures?

Results

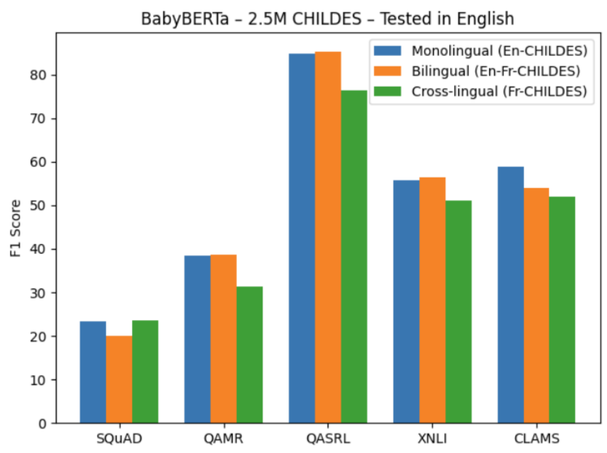
Bilingual Advantage on XNLI



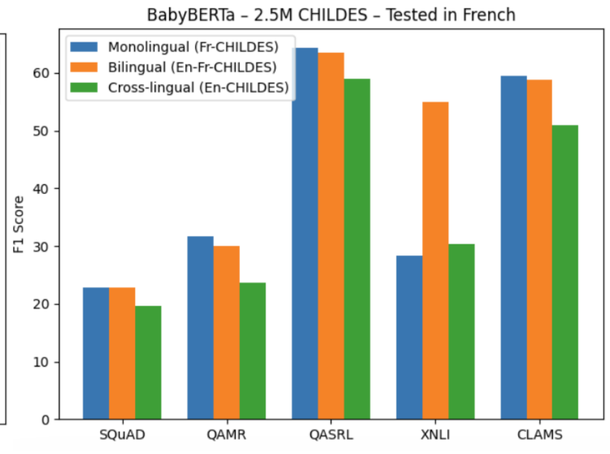
CDS Improves Grammar (CLAMS)



BabyBERTa – 2.5M CHILDES (English Evaluation)



BabyBERTa – 2.5M CHILDES (French Evaluation)



Key Findings

1. **Bilingual pretraining strongly improves textual inference (XNLI),** with dramatic gains for French at 2.5M tokens.
2. **Corpus type determines competence focus**: Wikipedia favors semantic tasks, while CHILDES improves grammatical performance (CLAMS).
3. **Bilingual advantages weaken with scale but persist at 10M tokens**, indicating diminishing yet stable multilingual benefits.
4. **Patterns replicate across architectures**, suggesting data composition and language pairing matter more than model size.

Future Work

- Experimenting on additional language pairs beyond English-French.
- Extending the analysis to decoder-only and larger-scale language models.
- Broadening evaluation to additional semantic and syntactic benchmarks.

References

- Huebner et al. (2021). BabyBERTa: Learning More Grammar with Small-Scale Child-Directed Language. CoNLL.
- Warstadt et al. (2023). Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. CoNLL.
- Conneau et al. (2018). XNLI: Evaluating Cross-Lingual Sentence Representations. EMNLP.
- Mueller et al. (2020). Cross-Linguistic Syntactic Evaluation of Word Prediction Models. ACL.
- Rajpurkar et al. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. EMNLP.
- Michael et al. (2018). Crowdsourcing Question-Answer Meaning Representations (QAMR). NAACL-HLT.
- He et al. (2015). Question-Answer Driven Semantic Role Labeling (QASRL). EMNLP.
- MacWhinney (2000). The CHILDES Project: Tools for Analyzing Talk.