Code-Switching Detection in Multilingual Child Speech with SwissBERT

Agnese D'Angelo[1,*] **Sina Ahmadi**[2] **Moritz M. Daum**[1] **Stephanie Wermelinger**[1]
[1]Department of Psychology, University of Zurich
[2]Department of Computational Linguistics, University of Zurich
[*]a.dangelo@psychologie.uzh.ch

Abstract

Code-switching is widespread in multilingual speech, yet its automatic detection remains challenging, especially for low-resource languages. In Switzerland, a context with multiple languages and Swiss German varieties, these challenges are amplified by variable orthography and limited annotated data. We present a supervised word-level language-identification system for code-switching detection in multilingual everyday child and adult speech, obtained by fine-tuning SwissBERT. We constructed a dataset of four languages and an *other* category, implemented controlled subword–label alignment, and evaluated performance using token-level F1. To contextualize SwissBERT's performance, we additionally fine-tuned mBERT as a multilingual baseline. SwissBERT achieves robust word-level predictions and outperforms mBERT. We release the full training pipeline and evaluation scripts to facilitate reproducibility.

🤗 | ZurichNLP/SwissBERT-CS

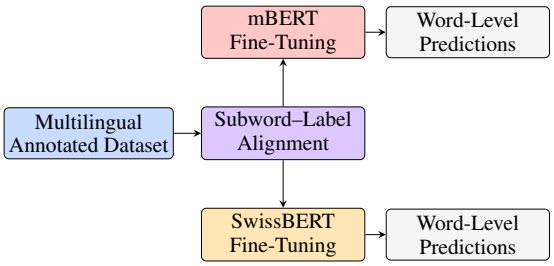

Figure 1: Overview of the methodology. A multilingual annotated dataset is aligned at the subword level and used to fine-tune SwissBERT and mBERT.

1 Introduction

Multilingualism shapes Switzerland's identity: one-third of the population regularly uses more than one language (Bundesamt für Statistik, 2021). In multilingual settings, speakers often engage in code-switching (CS), the alternation between two or more languages within a conversation. CS emerges early in development (Smolak et al., 2020) and is a natural, systematic feature of multilingual communication rather than a sign of confusion or delay. CS occurs across sentences (i.e., intersentential; "I went to the park yesterday! *Ich ha Spass gha.*") or within the same sentence (i.e., intrasentential; "Could you give me *s Buech bitte*?"). Swiss German presents unique challenges for computational models: it lacks a standardized orthography, exhibits substantial regional variation, and is underrepresented in existing multilingual pretrained models.

These difficulties are amplified in child speech. Young children frequently produce phonological reductions, truncations, and non-standard word forms that deviate from adult speech (Dodd et al., 2003), making word boundaries and lexical identity less predictable. In multilingual contexts, children may additionally blend features from multiple languages within a single word or use language-specific morphology in non-target-language utterances (Paradis and Genesee, 1996), further complicating word-level language identification. Moreover, child-directed speech from caregivers often contains simplified or exaggerated forms that differ from standard adult registers (Foulkes et al., 2005), adding another layer of variation. Unlike social-media text, where non-standard spelling follows some community conventions that models can learn, the variability in child speech is less systematic and harder to anticipate from pre-training data alone.

In this work, we address these gaps by constructing a multilingual, manually annotated dataset across four languages and by fine-tuning SwissBERT (Vamvas et al., 2023)—a transformer model pre-trained on Swiss national languages and dialects—for word-level CS detection. Because SwissBERT is pre-trained on Swiss German, we hy-

pothesize that it will outperform other multilingual models. To test this hypothesis, we additionally fine-tune mBERT (Wu and Dredze 2020) as a contrastive baseline using the same training pipeline.

To our knowledge, this is the first work to apply transformer-based word-level CS detection to multilingual child speech involving Swiss German.

2 Related Work

Previous work on CS relied heavily on manual annotation to identify language boundaries (e.g., Lachemat et al., 2025). Most research on CS detection has focused on social-media text. The Second Shared Task on Language Identification in Code-Switched Data (Molina et al., 2016) provided one of the first large-scale word-level benchmarks for bilingual social-media corpora, highlighting persistent challenges, including noisy orthography, short words, and limited annotated data. Rijhwani et al. (2017) proposed a generalized word-level language identification model based on a Hidden Markov Model, enabling multilingual CS detection without manually annotated training data and demonstrating its effectiveness on Twitter data.

Recent work has increasingly turned to transformer-based architectures. Multilingual models like mBERT (Wu and Dredze, 2020) and XLM-R (Conneau et al., 2020) outperform traditional sequence models on CS language identification, especially when fine-tuned on mixed-language data (e.g., Aguilar et al., 2020; Khanuja et al., 2020; Das et al., 2023). Other studies have explored syntactic and discourse information: Sterner and Teufel (2025) showed that syntactic structure alone can support human-level CS acceptability judgments in a graph-neural-network model (CSntax-GNN), with patterns generalizing across unseen language pairs.

Data-augmentation strategies have also been proposed to improve CS language modeling. Hu et al. (2020) combined monolingual sentence selection, syntactic-constraint substitution, and a pointer-generator network, achieving substantial perplexity reductions on Mandarin–English CS corpora.

A comprehensive survey by Winata et al. (2023) reviewed more than 400 CS studies and documented a rapid increase in the number of publications. However, two gaps remain: (1) most work targets high-resource language pairs (e.g., Spanish–English, Mandarin–English), and (2) fine-grained word-level CS detection remains difficult for informal registers with high variability. Furthermore, automatic CS detection in child speech, particularly for low-resource languages, remains unexplored. This gap is especially relevant for developmental psychology, where CS annotation is typically performed manually. Automating this process provides a methodological bridge between computational linguistics and developmental research, enabling scalable, reproducible analyses of multilingual child language.

Language	Words
Swiss German	50,733
English	49,658
Italian	48,844
French	46,281
Other	9,956

Table 1: Distribution of the 205,472 manually labeled words across Swiss German, English, French, Italian and *other*.

3 Data

Existing CS datasets and benchmarks focus primarily on adult speakers, social-media text, or high-resource language pairs. No existing resources cover multilingual child speech in Swiss German, so we created a manually annotated dataset across four languages (see Table 1), including an *other* category.

3.1 Sources

In total, we assembled a corpus of 205,472 manually labeled words across five labels (Swiss German, English, French, Italian, Other). The dataset contains 46,501 utterances with a mean utterance length of 4.42 words (SD = 5.16), ranging from 1 to 127 words per utterance. This level of variability is characteristic of spontaneous child speech, where children often produce one-word utterances while caregivers may produce much longer utterances.

English, French, and Italian words were extracted from publicly available transcribed child-speech corpora in the CHILDES database (MacWhinney, 2000), including Antelmi and Morlacchi (2005); Stine and Bohannon (1983); Burgato (2025); Hamann et al. (2003); Genesee et al. (2004); Pizzuto (2004); Tonelli (2004); Watkins (2004), both monolingual corpora and multilingual corpora with CS. More than one third (17,624 words) of the Swiss German words were obtained

from unpublished child speech data collected as part of a pre-registered[1] project in our research unit. The dataset consists of spontaneous everyday speech produced by three-year-old children and their interlocutors (i.e., caregivers, siblings, friends). The recordings were collected using microphones that the children wore for approximately 12 hours across one week in their everyday environments. Because no open-source Swiss German child speech corpora currently exist, we expanded the Swiss German portion of the dataset with an additional 33,109 tokens from the SwissDial corpus (Dogan-Schönberger et al., 2021). Although SwissDial is not child speech, its inclusion increases lexical diversity and improves coverage of Swiss German orthographic variation. Because the dataset was constructed over an extended period, the earliest stage involved artificially enriching code-switching patterns. Specifically, within a segment of 1,619 monolingual Swiss German words, we added 139 English words by translating parts of existing utterance to simulate code-switching. This approach was only used in the initial phase; as more recordings were transcribed and real-life instances of child code-switching became available, the dataset was expanded using real instances instead of artificial ones.

3.2 Preprocessing

Before word annotation, the data were preprocessed to ensure consistent tokenization across heterogeneous sources. This included: (i) splitting contractions in English, French, and Italian (e.g., *c'/est*, *don'/t*) to ensure that each meaningful unit receives an independent label; (ii) separating Swiss German clitics (e.g., *s/Auto*); (iii) removing punctuation; (iv) normalizing white-space and removing transcription artifacts. These steps ensured that word boundaries aligned with meaningful linguistic units, which is essential for reliable word-level CS detection.

3.3 Annotation

After preprocessing, the dataset was manually labeled at the word level. The annotator assigned one of four languages (Swiss German, French, English, or Italian) to each token based on its lexical form. Proper names, place names, interjections, fillers, and words whose language could not be reliably assigned to any of the four languages (e.g., *super*, *okay*) were assigned to an additional *other* category. This category prevents false code-switch detections for language-neutral or ambiguous lexical items that do not clearly belong to a single language. All annotations were first performed by an annotator fluent in all languages in the dataset.

[1] https://osf.io/57wt3/overview?view_only=3fb372b514e4413ca8dbbde2056f6011

Comparison	Percent Agreement	Cohen's κ
A vs. B	0.99	0.98
A vs. GOLD	0.98	0.97
B vs. GOLD	0.98	0.97

Table 2: Inter-annotator agreement for language-label annotations.

To calculate inter-annotator agreement (IAA), we extracted a total of 200 utterances (1,831 words), equally divided into 100 monolingual utterances (balanced across Swiss German, English, French, and Italian) and 100 multilingual utterances. Only utterances containing at least three words were included, as shorter utterances do not provide enough lexical or contextual information to be reliably categorized. Two annotators (A and B), both fluent in all languages of the dataset, completed the task. Before the main annotation, they annotated 30 utterances as a training exercise, after which they proceeded independently following the annotation guidelines.

IAA was computed between the two annotators and relative to the gold standard (i.e., the labeled dataset). Percent agreement and Cohen's κ are reported in Table 2. The agreement values are extremely high, indicating that the task is straightforward and that the annotation scheme is well-defined. Disagreements were mostly found in short utterances, where the lack of contextual and lexical cues made language identification more difficult.

3.4 Sentence IDs

Each label word in the dataset is associated with a sentence identifier (i.e., Sentence ID), allowing for the reconstruction of utterance boundaries after preprocessing and tokenization. Maintaining utterance structure is crucial for analyzing code-switching patterns, as CS can occur both within and across utterances.

4 Methodology

Our approach combines manual dataset construction, controlled preprocessing, and transformer-based token classification. Figure 1 provides an overview of the methodology.

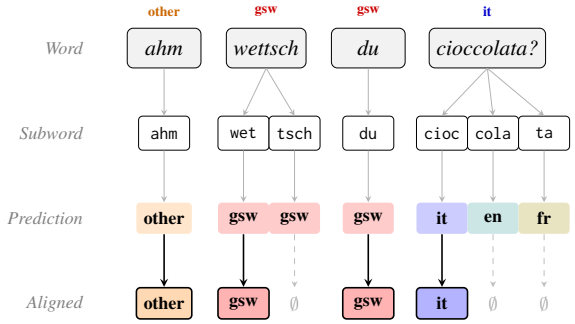

Figure 2: Example of WordPiece tokenization with corresponding word-level labels, model predictions, and loss masking. Only the first subword of each word contributes to the supervised loss.

4.1 Model Architecture

We fine-tuned SwissBERT (Vamvas et al., 2023), a multilingual transformer pre-trained on the Swiss national languages (Standard German, French, Italian, and Romansh), and more importantly, including recently-added Swiss German. A linear classification head predicts one of the five labels (Swiss German, English, French, Italian or *other*) for each word.

4.2 Baseline Model

To contextualize the performance of SwissBERT, we additionally fine-tuned mBERT (Devlin et al., 2019) using the same preprocessing, subword–label alignment, and training pipeline. mBERT serves as a contrastive baseline: unlike SwissBERT, it is not pre-trained on Swiss German. This comparison allows us to isolate the contribution of SwissBERT's dialect-specific pretraining and assess whether exposure to Swiss German during pretraining yields measurable improvements in word-level CS detection.

4.3 Subword Tokenization and Alignment

SwissBERT uses WordPiece tokenization, which frequently splits words into multiple subword units. However, code-switching happens at the *word* level, and accurate CS detection requires a single language label per word. To reconcile word-level labels with subword tokenization, we adopt a first-subword labeling strategy: the first subword inherits the word's language label, while all subsequent subwords are ignored during loss computation. This alignment ensures that the model learns word-level language boundaries while remaining compatible with subword-based transformer architectures.

4.4 Inference

At inference time, the model applies the same preprocessing and WordPiece tokenization used during training. As illustrated in Figure 2, SwissBERT outputs a label for every subword token. To obtain word-level predictions, we use the same strategy as in training: only the first subword is kept, and all later subwords are ignored. This mirrors the loss-masking scheme showed in Figure 2, where non-initial subwords receive a null label (ø), ensuring that each word receives exactly one linguistically meaningful prediction.

4.5 Training Procedure

We fine-tuned SwissBERT with a classification head using the Hugging Face Trainer API (Wolf et al., 2020) for five epochs, a batch size of 8, and a learning rate of 5×10^{-5}. We used the AdamW optimizer with weight decay 0.01. Validation was performed at the end of each epoch using the F1 metric computed over non-masked tokens, and the best-performing checkpoint was selected.

To train the model, we randomly split the annotated dataset into training and validation sets using a 90/10 ratio. The split was performed at the utterance (sentence ID) level, ensuring that all tokens within the same utterance remained in the same partition. This prevents contextual leakage across splits and avoids artificially inflated performance. Each utterance was truncated to a maximum length of 128 tokens, which standardizes input size and ensures consistent batching during training.

We evaluate model performance using the standard F1 score, computed with the Seqeval library (Nakayama, 2018). Seqeval is widely used for sequence-labeling tasks such as named entity recognition and token-level classification, and provides a reliable implementation of precision, recall, and F1 over label sequences. We report token-level F1 computed only on non-masked tokens (i.e., the first subwords), thereby aligning the evaluation with our word-level labeling scheme. Following common practice in token-level evaluation, we use Seqeval's micro-averaged F1 across all tokens to select the best model checkpoint during training.

5 Results

5.1 Test Setup

The models were evaluated on an independent dataset of 15,819 manually labeled words grouped into 3,420 utterances. This dataset was collected

Child	Languages	Words
A	French, Swiss German	3,442
B	Italian, Swiss German	1,960
C	English, Swiss German	5,337
D	Italian, French, English	5,080
Total		**15,819**

Table 3: Composition of the evaluation dataset, based on transcriptions of naturalistic everyday interactions between multilingual children and their caregivers.

	SwissBERT			mBERT		
Language	P	R	F1	P	R	F1
English	0.87	0.99	0.93	0.82	0.92	0.87
French	0.98	1.00	0.99	0.88	0.90	0.89
Italian	0.91	0.90	0.90	0.83	0.88	0.85
Swiss German	0.98	0.95	0.96	0.92	0.85	0.88
Other	0.99	0.95	0.97	0.90	0.87	0.89
Overall	**0.96**	**0.96**	**0.96**	**0.88**	**0.88**	**0.88**

Table 4: Token-level precision (P), recall (R), and F1 for SwissBERT and mBERT on the external evaluation set. SwissBERT consistently outperforms mBERT across all languages.

within the same pre-registered[2] project described in Section 3.1. Because it originates from a different population and recording context than the training corpus, it provides an ecologically valid estimate of model generalization. The composition of the evaluation set is provided in Table 3.

5.2 Token-level Prediction

Table 4 reports precision, recall, and F1 for each language category for both models. Across the test dataset, the fine-tuned SwissBERT model achieved an overall F1 score of 0.96. Performance was high across all languages, with F1 scores of 0.93 (English), 0.99 (French), 0.90 (Italian), 0.97 (Other), and 0.96 (Swiss German). These results indicate that SwissBERT generalizes well to naturalistic multilingual child speech despite dialectal variation and informal registers.

To assess the contribution of dialect-specific pre-training, we compared SwissBERT with mBERT, both fine-tuned on the same training data and alignment strategy. mBERT achieved an overall F1 score of 0.88. While mBERT performed reasonably well, SwissBERT outperformed it across all languages.

Overall, the results show that SwissBERT's pre-training on Swiss German leads to better subword representations of dialectal spellings.

[2] https://osf.io

Metric	Value
Gold switch points	736
Predicted switch points	862
Precision	0.78
Recall	0.91
F1 score	0.84

Table 5: Switch-point detection performance. A switch is defined as a change in language label between consecutive tokens, excluding the label *other*.

5.3 Switch-Point Detection

To further evaluate the performance of our model, we compute its ability to correctly identify *switch-points*. These correspond to transitions between two consecutive words where the language label changes, excluding any words labeled *other*. The *other* label is treated as non-linguistic and is removed from the evaluation stream prior to computing transitions. This ensures that sequences such as English → other → Swiss German are treated as a direct English → Swiss German transition.

Switch-points are therefore computed on a cleaned word sequence in which all words with the label *other* have been removed. A predicted switch is counted as correct only if it occurs at the same position in the global sequence and exhibits the same direction of transition (e.g., Swiss German → English) as in the gold annotation. Experiments on the switch-point detection, as summarized in Table 5, show that the model achieves a switch-point precision of 0.78, recall of 0.91, and F1 score of 0.84. These results suggest that the model captures the majority of true code-switching transitions, although it tends to over-predict switch points. This additionally indicates that even though the model achieves high token-level F1, as discussed in Section 5.2, it is less effective at identifying the exact locations of language transitions.

6 Discussion & Conclusion

Our aim was to develop an automatic approach to detect code-switching in multilingual child speech. We hypothesized that SwissBERT's pre-training on Swiss German and the Swiss national languages would provide an advantage for multilingual word-level classification. The evaluation confirmed this: SwissBERT outperformed mBERT across all languages.

A likely explanation for this performance difference lies in the training data and model specialization. SwissBERT is trained with a focus on language varieties relevant to Switzerland, allowing it to better capture linguistic nuances, vocabulary, and orthographic variation present in the dataset. In contrast, mBERT is designed as a broadly multilingual model trained on a wide range of languages, which can limit its ability to model specific regional language varieties.

The strong performance of both models may also reflect that the languages in our dataset are linguistically distant, which simplifies the classification problem. Future work with more closely related language pairs (e.g., Standard German vs. Swiss German) may pose greater challenges, because lexical overlap and shared morphology reduce the distinctiveness of subword patterns. Moreover, code-switching frequency varies widely across speakers and contexts, introducing additional variability that models must learn to handle.

A qualitative analysis also highlighted a structural limitation shared by both models: words that are orthographically identical across languages are difficult to classify reliably when they appear in isolation or in contexts with limited syntactic information. For example, the word *da* (Italian preposition and Swiss German adverb) is sometimes misclassified when isolated. In such cases, the model has no access to phonetic cues (which would distinguish the two pronunciations) or semantic cues (which would clarify the intended meaning), making misclassifications almost unavoidable. This limitation is not specific to our models but reflects an inherent ambiguity in written speech transcripts.

Finally, although the training pipeline is fully generalizable, the current model's advantage stems from SwissBERT's region-specific pre-training and is therefore limited to the Swiss context. As our research unit continues to collect more diverse multilingual child-speech data, the same corpus can be extended to train classifiers that recognize an increasingly broad set of languages, enabling progressively richer and more inclusive analyses of code-switching.

Limitations

The dataset is imbalanced. The *other* category is heterogeneous, grouping together diverse linguistic phenomena that are not easily comparable. The Swiss German portion of the dataset is not exclusively composed of child-speech data, as it also includes corpus data from SwissDial that differ substantially from naturalistic child language. This mismatch introduces distributional differences that may affect model performance and limit the validity of the results. Our long-term goal is to retrain the model on a corpus composed exclusively of child-speech data as soon as additional recordings become available. Although the current model shows high performance and strong inter-annotator agreement, these metrics may shift as more closely related languages are added. This is especially true for distinctions between language varieties (e.g., Swiss German vs. Standard German).

Manual annotation is time-consuming and introduces the possibility of bias, particularly for ambiguous or context-dependent words. These challenges are amplified by the absence of phonetic information, which limits the ability to disambiguate homographs or reduced child-speech forms. Future work should integrate both phonetic and orthographic information to improve the robustness of code-switching detection.

Ethical Considerations

All data involving families were collected in accordance with the ethical standards of the 1964 Helsinki Declaration and its later amendments. The recordings contain sensitive everyday conversations between children and their caregivers; therefore, raw audio and transcripts cannot be publicly released. The dataset reflects multilingual families in the Zurich area, which may limit generalizability and introduce demographic biases. Our work is intended solely for research on early multilingual development. The authors declare no conflicts of interest.

Acknowledgments

The project is financed by the Swiss National Science Foundation (reference number 10001585). Sina Ahmadi gratefully thanks the support of the UZH Grant (reference number 269093). We also thank the Kleine Weltentdecker*innen Lab and the Jacobs Center for Productive Youth Development at Universität Zürich for the support and all children and their caregivers for their participation in our studies.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Donatella Antelmi and Alessandra Morlacchi. 2005. L'interpretazione del linguaggio figurato nel ritardo mentale. *Rassegna Italiana di Linguistica Applicata*, 37(2):355–380.

Bundesamt für Statistik. 2021. Zunahme der mehrsprachigkeit in der schweiz: 68% verwenden regelmässig mehr als eine sprache. Swiss Federal Statistical Office report.

Laura Burgato. 2025. CHILDES Italian Burgato Corpus. CHILDES Database. 30 samples. Corpus date: 2025.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Richeek Das, Sahasra Ranjan, Shreya Pathak, and Preethi Jyothi. 2023. Improving pretraining techniques for code-switched NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1176–1191, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Barbara Dodd, Alison Holm, Zhu Hua, and Sharon Crosbie. 2003. Phonological development: a normative study of british english-speaking children. *Clinical Linguistics & Phonetics*, 17(8):617–643.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. SwissDial: Parallel multidialectal corpus of spoken Swiss German. *CoRR*, abs/2103.11401.

Paul Foulkes, Gerard Docherty, and Dominic Watt. 2005. Phonological variation in child-directed speech. *Language*, 81(1):177–206.

Fred Genesee, Elena Nicoladis, and Johanne Paradis. 2004. Childes French–English gnp corpus. CHILDES Database. French–English bilingual children in Montreal. Corpus date: 2004-03-30.

Cornelia Hamann, Sharon Ohayon, Sophie Dubé, Ulrich H. Frauenfelder, Luigi Rizzi, Michal Starke, and Pascal Zesiger. 2003. Aspects of grammatical development in young French children with sli. *Developmental Science*, 6(2):151–158.

Xuenan Hu, Qi Zhang, Liner Yang, Bin Gu, and Xun Xu. 2020. Data augmentation for code-switch language modeling by fusing multiple text generation methods. In *Proceedings of Interspeech 2020*, pages 1062–1066.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Houssam Eddine-Othman Lachemat, Abbas Akli, Nourredine Oukas, Yassine El Kheir, Samia Haboussi, and Shammur Absar Chowdhury. 2025. CAFE: Spontaneous code-switching speech dataset in Algerian dialect, French and English. *Data in Brief*, 63:112150.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk: The Database*, 3rd edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas. Association for Computational Linguistics.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Johanne Paradis and Fred Genesee. 1996. Syntactic acquisition in bilingual children: Autonomous or interdependent? *Studies in Second Language Acquisition*, 18(1):1–25.

Elena Pizzuto. 2004. CHILDES Italian Roma Corpus. CHILDES Database. Longitudinal study of a single Italian child. Corpus date: 2004-04-02.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.

Emily Smolak, Stephanie de Anda, Bianca Enriquez, Diane Poulin-Dubois, and Margaret Friend. 2020. Code-switching in young bilingual toddlers: A longitudinal, cross-language investigation. *Bilingualism: Language and Cognition*, 23(3):500–518.

Igor Sterner and Simone Teufel. 2025. Code-switching and syntax: A large-scale experiment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11526–11533, Vienna, Austria. Association for Computational Linguistics.

Elaine L. Stine and John N. III Bohannon. 1983. Imitations, interactions, and language acquisition. *Journal of Child Language*, 10(3):589–603.

Livia Tonelli. 2004. Childes Italian Tonelli corpus. CHILDES Database. Transcripts from three Italian children. Corpus date: 2004-04-01.

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. SwissBERT: The multilingual language model for Switzerland. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 54–69, Neuchatel, Switzerland. Association for Computational Linguistics.

Charles Watkins. 2004. Childes French–English Watkins corpus. CHILDES Database. French–English bilingual children studied for their use of deixis. Corpus created at Université de Paris XIII.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.