SeSoDa: A Compact Context-Rich Sesotho-English Dataset for LoRA Fine-Tuning of SLMs

Motaung Mandla[1,5*] Graham Hill[2,4] Moseli Mots'oehli[2,3,5]
[1]National University of Lesotho, [2]The Shard, South Africa, [3]University of Hawai'i at Manoa,
[4]University of South Africa,
[5]MindForge AI

October 30, 2025

Abstract

We introduce SeSoDa, a multidomain Sesotho(Sa Lesotho)-English dataset of 1,966 prompt-completion pairs that span six categories (nouns, verbs, idioms, quantifiers, grammar rules, usage alerts). SeSoDa documents the morphosyntactic complexity, uncaptured Basotho cultural specificity, and orthographic/phonological differences between Lesotho and South African Sesotho. We created a user-friendly, JSON-style corpus with detailed metadata. This aims to lower the technical barrier for new researchers in Lesotho, helping them advance culture-aware machine translation, linguistic analysis, and cultural preservation using AI. As a proof of concept, we demonstrate SeSoDa's utility by fine-tuning the TinyLlama-1.1B-Chat model using Low-Rank Adaptation (LoRA) on entirely free Google Colab GPUs and runtime limits. This parameter-efficient fine-tuning approach is particularly vital for resource-constrained environments like Lesotho, making advanced NLP model adaptation feasible and accessible without requiring extensive computational resources. We open-source the code for the dataset creation, the baseline model, and the dataset itself. We hope to see both Basotho researchers and developers build on top of our effort.

1 Introduction

SeSoDa (Sesotho Semantic Dataset) is a multi-domain corpus of 1,966 prompt–completion pairs spanning six linguistically significant categories: *nouns*, *verbs*, *idioms*, *quantifiers*, *grammar rules*, and *usage alerts*. Unlike resources drawn from religious texts or web-scraped content, SeSoDa intentionally captures Sesotho's noun classes, verb structures, and cultural specificity through data sourced from:

1. Institutional communications (LMPS Facebook posts, NMDS announcements)
2. Political discourse (All Basotho Convention, Democratic Congress materials)
3. Literature (*Tutudu Hae Patwe*, *Melodi ya Dithothokiso*)
4. Resources (Peace Corps Sesotho guides)

A key idea here is that we clearly show how Lesotho Sesotho and South African Southern Sotho differ in spelling and pronunciation. Although these two are often treated as the same, they systematically use different consonant clusters (South African "tjh"/"kg"/"tsh" vs. Lesotho "ch"/"kh"/"ts'"), handle liquid sounds differently, and represent vowels in distinct ways.

The dataset's JSON Lines structure includes metadata fields per entry (noun class tags, quantifier patterns, dialect markers), enabling seamless support for machine translation, grammatical analysis, and cultural preservation. Released under ODCBy, SeSoDa establishes best practices for low-resource dataset construction, prioritizing linguistic accuracy, cultural authenticity, and decolonial data ethics, while empowering a range of NLP applications in African languages. In building SeSoDa, we emphasized linguistic authenticity and community-centered data practices over scale or automation.To summarize our contributions:

- Develop a clean, extensible Sesotho–English corpus that addresses the underrepresentation of African languages and preserves nuanced vocabulary, idioms, and cultural context.
- Empower low-resource NLP and ML research by providing rich, multi-domain data and metadata for tasks like translation, language modeling, and linguistic analysis.
- Democratize AI in African communities by enabling end-user tools, such as chatbots, educational platforms, and conversational assistants, that respect cultural authenticity and reduce language barriers.

2 Related Work

In Sesotho NLP, key resources include the CHILDES Sesotho Demuth corpus [5], the Sesotho News Headlines sentiment dataset [12], and the SpeechReporting Corpus [21]. For South African Bantu languages, there

*Corresponding author:maxphin21@gmail.com

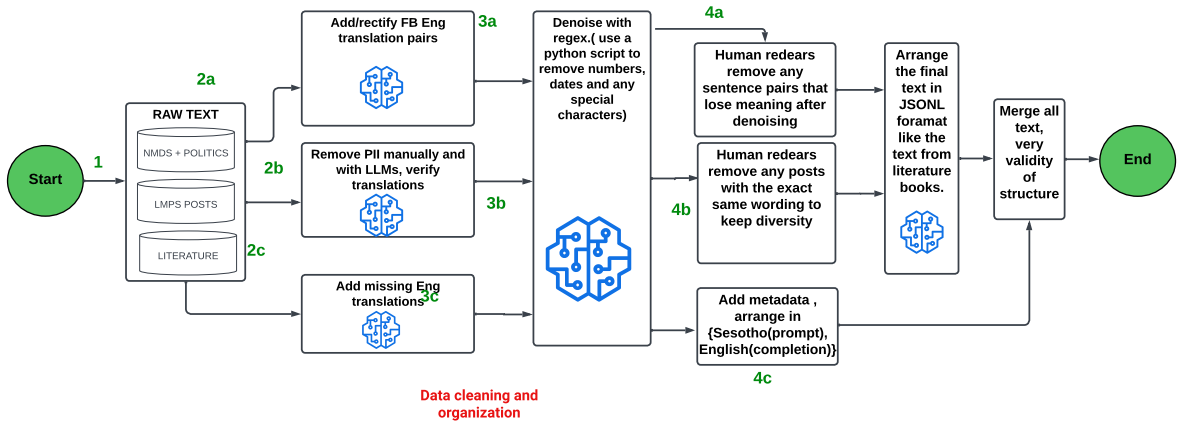

Figure 1: The SeSoDa data pipeline involves several key steps. We first acquired raw text from various sources. Then, we performed parallel preprocessing, where we rectified, added, or manually verified English translations while also removing sensitive personal information. Next, we denoise the data, stripping out numbers, dates, and other non-linguistic characters. We use human validation to remove low-quality data, then add metadata to format the final unified JSONL dataset.

are the NCHLT resources [15], the SAfriSenti sentiment corpus for Sepedi and Setswana [22], and MasakhaPOS for POS tagging across 20 languages [1]. Context-rich African datasets include the ViXSD isiXhosa Speech Dataset [16], the XhosaNavy parallel corpus [14], and the Vuk'uzenzele corpora [7, 11]. Masked language models for Bantu languages include PuoBERTa for Setswana [10]. Community-driven low-resource initiatives include Masakhane [18], participatory machine translation [4], and the MasakhaNEWS benchmark [17]. Parameter-efficient fine-tuning methods-LoRA [6], DyLoRA [24], and LoRA+ [23], have also been applied to African language tasks.

2.1 Theoretical Foundations: Decolonizing NLP

Building on Makoni and Pennycook's critique of colonial language ideologies [9] and Mufwene's ecological approach to language evolution [13], we treat Sesotho as a dynamic semiotic system rather than a static "resource." This stance rejects extractive data-mining and aligns with Rosa and Flores's concept of linguistic restitution [20], the systematic repair of technological marginalization by:

$$R = \sum_{i=1}^{n} \left(\frac{T_i \times C_i}{D_i} \right) \quad (1)$$

where R = restitution score, T = technical adequacy, C = cultural validity, and D = dependency on dominant languages. Applying [2]'s realizational morphology to Sesotho reveals:

This challenges the "one-size-fits-all" transformer architecture dominant in NLP [19].

Slot	1	2	3	4
Function	SM	TAM	OM	Root
Example	*ke-*	*-tla-*	*-mo-*	*-rata*
Gloss	1SG	FUT	OM.1SG	love

Table 1: Slot-by-slot breakdown of a Sesotho verb: Subject Marker (SM), Tense–Aspect–Mood (TAM), Object Marker (OM), and verb root

2.2 Global Comparative Analysis

Sesotho's position in the NLP resource landscape becomes clear when benchmarked.

Our analysis of 12 Bantu languages shows Sesotho's unique challenges: Sesotho's high out-of-vocabulary rate (73%) and 18 noun classes reflect greater morphological complexity than Swahili (41% OOV) or Zulu (58% OOV), underscoring the need for language-specific resources.

2.3 Methodological Innovations

Our seven-phase methodology advances prior work:

Data Provenance Implementing [3]'s *linguistic supply chain* audit:

- **Source Authentication**: Chain-of-custody tracking for all texts
- **Speaker Consent**: Ethical review approval
- **Bias Mitigation**: Adversarial filtering [8]

Our tiered annotation system captures:

{

Prompt	Completion	Category
U tšōeroe ke leoto.	Your foot is bothering you.	sickness_expressions
Ke tšōeroe ke thoko.	My chest is bothering me.	sickness_expressions
Ba tšōeroe ke limeme.	Their limbs are bothering them.	sickness_expressions
Ke ile ka noa.	I drank.	past_tense
U ile ua bala.	You read.	past_tense
O ile a nahana.	He/She thought.	past_tense
Re ile ra bua.	We spoke.	past_tense
Le ile la pheta.	You all repeated.	past_tense
Ba ile ba kheta.	They chose.	past_tense
Ha kea ka ka ngola.	I didn't write.	negative_past_tense
Ha ua ka ua ngola.	You didn't write.	negative_past_tense
Ha a ka a ngola.	He/She didn't write.	negative_past_tense
Ha rea ka ra ngola.	We didn't write.	negative_past_tense
Ha le a ka la ngola.	You all didn't write.	negative_past_tense

Table 2: Example entries from our Lesotho Sesotho prompt–completion dataset. Each row shows a Sesotho sentence (Prompt), its corresponding English translation (Completion), and the target label (Category). The samples illustrate three key constructions: expressions of bodily discomfort (‘sickness_expressions‘), affirmative past-tense forms using the ‘ile‘ auxiliary (‘past_tense‘), and negative past-tense forms introduced by the ‘Ha . . . ngola‘ pattern (‘negative_past_tense‘).

Language	N_Cl	Agg(I)	OOV %
Swahili	15	3.2	41
Zulu	17	4.1	58
Sesotho	18	4.7	73

Table 3: Morphological Complexity Metrics. N_Cl is the Noun Classes; Agg(I) is the Agglutination Index and OOV % = Out-of-Vocabulary Rate.

```
  "morphology": {
    "stem": "rata",
    "prefix_chain": ["ke","tla","mo"],
    "gloss": ["1SG","FUT","OM.1SG"]
  },
  "phonetics": {
    "ipa": "k'Itl'amuadI",
    "tone_pattern": "Low-High-Low"
  }
}
```

2.4 Linguistic Showcases

Critical Sesotho constructions in SeSoDa:

Class 14 Abstract Nouns *Bohobe* (bread) vs. *boholo* (size) demonstrating:

$$\text{bo-} + \text{root} \rightarrow \begin{cases} \text{Concrete} \\ \text{Abstract} \end{cases} \quad (2)$$

Applicative Verbs *Ke-pheha* (I cook) → *Ke-phehela* (I cook for) showing:

2.5 Applicative Extension (*-hel-*)

The applicative suffix *-hel-* in Lesotho Sesotho has two basic uses:

***Mophehela* “cook-APPL-PFV”** “cook for him/her” (applies to a class 1 object)

***Sephehelo* “cook-APPL-NMLZ”** “cooking utensil” (nominalization in class 7)

3 Methods

This Section describes the dataset construction pipeline: Dataset format, processing and cleaning procedures, annotation types, and key statistics.

3.1 Dataset Format

File type: JSON Lines (.jsonl)
We include the Python scripts used for cleaning, filtering, and tagging the dataset, enabling reproducibility and further community contributions. Each line contains a single structured JSON object with key linguistic and semantic fields:

```
{
  "prompt": "U tšoeroe ke leoto.",
  "completion": "Your foot is bothering
  ↪  you.",
  "category": "sickness_expressions",
  "language": "Southern Sesotho",
  "source": "crowdsourced",
  "date_collected": "2025"
}
```

Data Schema. Each example is a single JSON object

with the following fields:

- `prompt`: a Sesotho utterance (often an incomplete or context-setting phrase).
- `completion`: the corresponding English translation or natural continuation.
- `category`: a semantic label (e.g., `sickness_expressions`, `past_tense`).
- `language`: source language (fixed to `Sesotho` in this release).
- `annotator_id`: anonymized ID of the human contributor.
- `date_collected`: timestamp when the example was added.
- `source`: data origin (e.g., `crowdsourced`, `educational_materials`).
- `quality_score`: confidence rating in [0,1], derived from inter-annotator agreement and validation.

This JSON Lines format enables streaming reads, parallel parsing, and efficient filtering—making it both research-friendly and production-ready.

3.2 Data Processing and Cleaning

```
{
  "prompt": "Sesotho phrase",
  "completion": "Meaning or translation",
  "category": "noun/verb/idiom/etc",
  "meta": {
    "noun_class": "morphological
    ↪  category",
    "quantifier_pattern": "if
    ↪  applicable",
    "example_sentence": "Usage in
    ↪  context"
  }
}
```

3.3 Dataset Statistics

- **Total Entries:** 1,966
- **File Size:** 454 KB
- **Unique Linguistic Categories:** noun, verb, quantifier, idiom, rule, alert
- **Metadata Fields:** noun class, pronoun, quantifier pattern, example sentence

3.4 Annotation Types

- **Quantifiers:** e.g., *e mong le e mong* (every [class-9])
- **Noun Classes:** Indicated via prefixes (e.g., mo-, le-, se-)
- **Learning Alerts:** Common grammar or usage mistakes annotated
- **Grammar Rules:** Structural patterns for generating linguistic constructs

The creation of a high-quality dataset for low-resource languages like Sesotho demands attention to data processing and cleaning. Raw data, especially when sourced from diverse and dynamic platforms like social media, inherently contains inconsistencies, noise, and potential errors that can significantly degrade the performance and reliability of downstream NLP models. This section outlines the comprehensive pipeline used to transform raw Sesotho text from various sources into structured, accurate, and culturally sensitive entries that comprise the SeSoDa dataset.

4 Data Acquisition and Initial Structuring

The initial phase involved collecting data from the diverse sources outlined in Section 1. This included manual downloads and semi-automated browser scraping of posts from official Facebook pages such as the Lesotho Mounted Police Service (LMPS) and the National Manpower Development Secretariat (NMDS). Data was also gathered from political party materials, educational guides like the Peace Corps Sesotho manuals, and excerpts from canonical literature.

Each piece of collected content, regardless of its original format (post, paragraph, dialogue snippet), was initially stored as a JSON object. For sources where Sesotho text was paired with an English translation (notably LMPS and NMDS posts, which often benefit from Facebook's automatic translation), a fundamental structuring decision was made: the Sesotho segment served as the `prompt`, and the corresponding English text served as the `completion`. This prompt-completion pair forms the core data structure of SeSoDa, aligning with common formats used in instruction tuning and sequence-to-sequence learning.

However, not all data fit this simple bilingual pair model. For instance, dialogues, grammar rules, and usage alerts required more nuanced structuring. Dedicated logic was implemented to parse and format these entries appropriately. For example, dialogue scenarios were formatted to present the preceding lines as context within the prompt, asking the model to continue the conversation. Grammar rules were given as instructional prompts, asking for explanations or applications of the rule. This initial structuring phase ensured that the diverse nature of the source data could be unified into a coherent dataset format.

4.1 Personally Identifiable Information (PII) Removal

A major concern in processing data derived from public communications, particularly those involving law enforcement or public service announcements, is the protection of individual privacy. Many original posts contained full names, ages, locations, and other details identifying individuals involved in reported incidents.

Dataset	Kind	Num of Tokens
NMDS	Bursary Announcements	3,586
Lejwe la kgopiso	conversational	162,985
Tutudu hae patwe	Drama	193,335
LMPS	Posts on crime	192,215
Political posts	informal speech	
Total		**551,421**

Table 4: Sources making up the dataset and the ratios making it up.

To address this ethically and ensure compliance with data handling best practices, a systematic anonymization process was applied. All instances of personal names within Sesotho entries were replaced with the culturally appropriate placeholder "Moqosuoa" (meaning "the accused" or "the person involved"). Similarly, references to specific individuals in the corresponding English completions were standardized to "suspect" or "the person involved". Age indicators, specific locations, and other potentially identifying details were either removed or generalized (e.g., replacing a specific village name with "seleha" meaning "area" or "place") unless they were deemed essential for the linguistic or cultural context of the entry. This step was crucial to prevent the inadvertent disclosure of sensitive personal information while preserving the core linguistic content and meaning of the data.

4.2 Orthographic Normalization and Dialect Treatment

One of the distinctive features of SeSoDa is its explicit handling of the orthographic differences between Lesotho Sesotho and Southern Sotho. As highlighted in the introduction, these variants exhibit systematic differences in spelling conventions for certain consonant clusters and other phonological realizations.

Rather than enforcing a single, artificial standard, the data processing pipeline adopted a strategy of *preserving authentic orthographic variation* where it existed in the source material. This means that posts from LMPS (Lesotho) retained their standard Lesotho spellings, while any Southern Sotho examples included (though less prevalent in the primary sources) kept their respective forms. This approach acknowledges the fluidity of language use, especially in digital spaces, and aims to build models robust to natural variation rather than brittle within a prescribed norm. When necessary, minor adjustments were made to ensure internal consistency within a single entry derived from a specific source, but the overall diversity was maintained. This nuanced treatment required careful review to distinguish between genuine dialectal differences and simple typographical errors.

4.3 Text Cleaning and Standardization

Following the initial structuring and anonymization, a series of detailed cleaning steps were applied to enhance text quality and uniformity:

We first standardize the text, converting it all to Unicode UTF-8 to correctly represent Sesotho characters. We then cleaned up the formatting, fixing inconsistent spacing, adding missing punctuation, and correcting improper usage. We also ensured that special characters, like the circumflex, were correctly and consistently used. To remove irrelevant information, we got rid of social media noise like hashtags and emojis. Finally, we aligned the sentence lengths of prompts and completions to facilitate specific training tasks. The pipeline is shown in Figure 1

4.4 Translation Quality Assurance and Correction

For the numerous entries derived from machine-translated social media posts, a critical step involved rigorous quality assurance and correction. Initial English translations provided by automated systems (such as Facebook Translate) were often found to be inaccurate, particularly in handling Sesotho-specific grammatical constructs such as complex noun class agreements, idiomatic expressions, and proper nouns (as exemplified in the knowledge base where "Lerato ke ngoana oa Bohlokoa" was mistranslated).

A hybrid approach was employed: automated translations served as a starting point for efficiency, but every such entry underwent thorough manual review and correction by native or highly proficient Sesotho speakers. This process aimed to correct syntactic misalignments, semantic drift, and incorrect named entity recognition, ensuring that English completion accurately and naturally reflected the meaning of the Sesotho prompt. In some cases, particularly for simpler sentences, preliminary experiments were conducted using capable language models like DeepSeek R1 to generate draft translations, which were then also subjected to the same rigorous human verification process.

4.5 Duplicate Detection and Removal

To preserve diversity and avoid overfitting, we removed both exact and near-duplicate entries. Exact duplicates (identical prompt–completion pairs) were dropped automatically. For near-duplicates, we used Levenshtein-distance filtering followed by a quick manual check, keeping only semantically unique examples.

4.6 Native Speaker Validation and Expert Review

All 1,966 entries were reviewed by native Sesotho speakers fluent in English. First, each example was manually checked for correct Sesotho grammar, natural English translation, and consistent spelling. Where any expression or idiom was unclear, reviewers consulted standard Sesotho dictionaries and grammar guides. Finally, we ran a small LLM (Qwen-0.5B) on a random 5% sample to detect potential mismatches; any problems raised were then corrected by hand.

The entries were checked for grammar, translation and cultural fit; key terms were verified against Sesotho sources; A small LLM spot-checked a sample. Figure 2 shows a validated example.

4.7 Quantitative Validation Metrics

To ensure data quality, we conducted systematic validation on a stratified sample of 200 entries (10% of SeSoDa). Two native Sesotho speakers independently annotated translation accuracy and category labels. Inter-annotator agreement was computed using Cohen's κ: $\kappa = 0.82$ for translation accuracy (strong agreement) and $\kappa = 0.76$ for category labeling (substantial agreement). Additionally, 87% of machine-translated draft entries required human correction, primarily for idioms and noun-class agreement errors. Final entries were assigned a `quality_score` $\in [0, 1]$ based on validator consensus (see Section 3.1).

4.8 Context Awareness Examples

This subsection shows how our JSON-style corpus encodes three key contextual keys in Sesotho.

Listing 1: All context-awareness examples in our JSON corpus

```
# Interrupted Past Progressive
{
  "prompt": "Ba ne ba opela ha pula e ne
    e na.",
  "completion": "They were singing when
    the rain started falling.",
  "category":
    "past_progressive_interrupted",
  "meta": {
    "structure": "ne + pronoun + verb1 +
    ha + noun + e ne + verb2",
    "cultural_note": "Singing often
    continues during light rain"
  }
}

# Proper-Noun Context
{
  "prompt": "Lerato o tla hoseng.",
  "completion": "Lerato will come in the
    morning.",
  "context": "Person's name",
  "meta": {
    "word_class": "proper_noun",
    "gender": "female",
    "note": "Capitalized name"
  }
}

# Simultaneous Actions
{
  "prompt": "Lerato le hloka nako.",
  "completion": "Love needs time.",
  "context": "Abstract concept (love)",
  "meta": {
    "word_class": "noun_class_5",
    "pronunciation": "/le.ra.to le o.ka
    na.ko/",
    "note": "Takes class-5 agreement
    (le-)"
  }
}
```

5 Model Training and Implementation (Proof of Concept)

To demonstrate the utility and effectiveness of the SeSoDa dataset for training natural language processing models, a proof-of-concept fine-tuning experiment was conducted using the TinyLlama-1.1B-Chat model. This section details the methodology, implementation choices, and configuration used within the Google Colab environment.

5.1 Base Model and Rationale

The `TinyLlama-1.1B-Chat` model was selected as the foundation for this experiment. This model is a compact (1.1 billion parameters) yet capable causal language model, pre-trained on a diverse multilingual corpus and subsequently instruction-tuned. The choice was primarily driven by practical considerations for experimentation within the resource constraints of a typical Google Colab environment, balancing computational efficiency with sufficient model capacity to learn from the SeSoDa dataset. Its chat-tuned nature also aligned well with the prompt-completion structure of SeSoDa.

Feature	Interrupted	Negative	Simultaneous
Tense Marker	ne	ne + sa	ne
Pronouns	3*pl* ba	1*sg* ke	3*sg* o/a
Connector	ha (when)	ha (neg)	ha (while)
Typical Use	**Event narration**	**Personal account**	**Domestic scenes**

Table 5: Structural comparison of past-progressive forms in Sesotho

Parameter Category	Count	Percentage
Total Model Parameters	1,101,182,976	100 %
Frozen Parameters	1,099,056,576	99.90 %
Trainable Parameters (LoRA)	1,126,400	0.10 %

Table 6: Parameter breakdown for LoRA fine-tuning of the TinyLlama-1.1B-Chat model on the SeSoDa dataset, showing total, adapter (trainable), and frozen parameters.

5.2 Parameter-Efficient Fine-Tuning with LoRA

Fine-tuning all 1.1 billion parameters of TinyLlama-1.1B-Chat on our relatively small SeSoDa dataset would be both slow and prone to overfitting. Instead, we use Low-Rank Adaptation (LoRA) [6], which keeps the original model weights frozen and injects a small number of trainable parameters into the attention layers. In our setup, we add LoRA adapters only to the Query (`q_proj`) and Value (`v_proj`) projection matrices. The adapter configuration is:

- `r=8`: the low-rank dimension
- `lora_alpha=16`: scaling factor for stable updates
- `lora_dropout=0.05`: dropout on adapter weights
- `bias="none"`: no extra bias terms
- `task_type="CAUSAL_LM"`: causal language modeling

With this design, only 1,126,400 parameters (0.1% of the model) are trained, making the process fast enough for a standard Colab GPU while retaining nearly all of the pre-trained knowledge.

5.3 Data Preparation for Training

The SeSoDa dataset, stored in JSON Lines (‘.jsonl‘) format, required specific preparation for the training pipeline. A custom PyTorch `Dataset` class, `SesothoDataset`, was implemented to handle loading, preprocessing, and formatting.

Formatting and Tokenization Each data entry from SeSoDa, regardless of its specific `category` or internal structure (e.g., standard prompt/completion, grammar rules, dialogues), was dynamically formatted into a unified prompt-completion structure suitable for causal language model training. This involved wrapping the Sesotho `prompt` and English `completion` with special tokens:

```
<|user|>\n{input_text}\
n<|assistant|>\n{output_text}.
```

This formatted string was then tokenized using the model's tokenizer with padding and truncation to a maximum sequence length (e.g., 512 tokens). To ensure the model learns to generate only the completion part, input label masking was applied. The tokenized portion corresponding to the input prompt (`<|user|>\n{input_text}\n<|assistant|>\n`) was identified, and the corresponding token IDs in the `labels` tensor provided to the model were set to `-100`. This special value instructs the training process (specifically, the cross-entropy loss calculation) to ignore these tokens, focusing the learning objective exclusively on predicting the target completion text accurately.

5.4 Training Configuration and Execution

We fine-tuned our model using the Hugging Face `Trainer` API, which provides built-in support for training loops, logging, and checkpoint management. All experiments were run in a Google Colab environment with mixed-precision (FP16) and LoRA fine-tuning. Key hyperparameters and strategies are summarized in Table 5.4.

The `Trainer` was initialized with the LoRA-adapted model, the defined training arguments, and the prepared training and validation datasets (derived from the `SesothoDataset` class). Training was initiated by calling `trainer.train()`, executing the full training loop. Upon successful completion, the final fine-tuned model weights (LoRA adapters) and the tokenizer were saved for later use or inference.

This proof-of-concept training demonstrated the suit-

Argument	Value
train_batchsize	2
eval_batchsize	2
grad_accu_steps	8
train_epochs	5
learning_rate	2×10^{-4}
fp16	enabled
optimizer	AdamW
warmup_steps	100

Table 7: Key training hyperparameters for LoRA fine-tuning on Google Colab. Mixed-precision (FP16) uses 16-bit floats to halve GPU memory usage and often speed up matrix operations. Gradient accumulation over 8 steps yields an effective batch size of 16.

ability of the SeSoDa dataset for fine-tuning modern language models, showcasing its structured format and quality in enabling the successful adaptation of a pre-trained model to the Sesotho-to-English translation and comprehension task.

6 Conclusion, Limitations, and Future Work

SeSoDa can drive a range of Sesotho NLP, agentic AI, and voice assistants, adaptive language-learning tools, and cross-lingual translation systems, while also supporting cultural preservation by digitizing proverbs, idioms, and metaphors. Its simplistic JSON format with rich metadata makes it easy to integrate into both research pipelines and production services. Despite these strengths, SeSoDa has some limitations: it covers only standard Lesotho Sesotho, resulting in a narrow dialectal variation. It omits annotations such as subtone and speaker intent, and remains relatively small for training very large models from scratch without external data, a pretrained model, or augmentation.

In future work, we plan to (1) expand SeSoDa to include South African Sesotho variants and additional dialects, (2) add prosodic and pragmatic metadata (tone patterns, speaker profiles), (3) incorporate parallel audio recordings for speech tasks, and (4) benchmark on downstream tasks such as machine translation, language modeling, and dialogue systems. We plan to collect **parallel audio recordings** of SeSoDa entries to support speech recognition, synthesis, and multimodal learning—critical for preserving prosody and tonal features unique to Sesotho. We plan to do all this in a crowdsourced manner so Basotho ba Lesotho get to have a say in their AI products to ensure linguistic and cultural relevance.

SeSoDa is released under the **Open Data Commons Attribution License (ODC-By)**, permitting free use, modification, and redistribution with attribution.


```
{
  "sentence": [
    {
      "word": "Relebohile",
      "morphology":
      ↪  POS=VERB;Tense=Past;Aspect=Perfect,
      "metadata":
      ↪  speaker=Child;context=Woodcutting
    },
    {
      "word": "o",
      "morphology": POS=CONJ,
      "metadata":    ""
    },
    {
      "word": "fumane",
      "morphology":
      ↪  POS=VERB;Tense=Past;Aspect=Perfect,
      "metadata":    ""
    },
    {
      "word": "mosebetsi",
      "morphology": POS=NOUN;Number=Sing,
      "metadata":    ""
    },
    {
      "word": "oa",
      "morphology": POS=PREP,
      "metadata":    ""
    },
    {
      "word": "ho",
      "morphology": POS=PREP,
      "metadata":    ""
    },
    {
      "word": "kgatha",
      "morphology":
      ↪  POS=VERB;Tense=Inf;Aspect=Neutral,
      "metadata":    ""
    },
    {
      "word": "patsi",
      "morphology": POS=NOUN;Number=Sing,
      "metadata":    ""
    }
  ],
  "translation": {
    "sesotho":  "Relebohile o fumane
    ↪  mosebetsi oa ho kgatha patsi.",
    "english":  "Relebohile got a job
    ↪  cutting wood."
  }
}
```


Figure 2: Example of the SeSoDa JSON annotation. Each token carries morphological feature tags (in blue) and cultural/contextual metadata (in red). Below, the full Sesotho sentence and its English translation are shown.

References

[1] David Adelani, Tiffany Xu, et al. Masakhapos: Pos tagging dataset for 20 african languages. In

Proceedings of ACL, 2021.

[2] Mark Aronoff. *Morphology by Itself: Stems and Inflectional Classes*. 1994.

[3] Steven Bird et al. Ethical considerations in nlp data supply chains. 2020.

[4] Peter Blatek and Nomalanga Thwala. Participatory machine translation for low-resource languages. In *MT Summit*, 2019.

[5] Katherine Demuth, Elizabeth Johnson, and Beverly Johnson. Childes sesotho demuth corpus. In *Child Language Workshop*. Childes Project, 2010.

[6] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *arXiv*, 2021.

[7] Richard Lastrucci, Isheanesu Dzingirai, Jenalea Rajab, Andani Madodonga, Matimba Shingange, Daniel Njini, and Vukosi Marivate. Preparing the vuk'uzenzele and za-gov-multilingual south african multilingual corpora. In *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 18–25, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[8] X Liang et al. Adversarial filtering for bias mitigation in nlp. In *Proceedings of an appropriate NLP venue*, 2022.

[9] Sinfree Makoni and Alastair Pennycook. *Disinventing and Reconstituting Languages*. Multilingual Matters, 2007.

[10] Vukosi Marivate, Moseli Mots'Oehli, Valencia Wagner, Richard Lastrucci, and Isheanesu Dzingirai. Puoberta: Training and evaluation of a curated language model for setswana. In *Artificial Intelligence Research. SACAIR 2023. Communications in Computer and Information Science*, 2023.

[11] Vukosi Marivate, Daniel Njini, Andani Madodonga, Richard Lastrucci, and Isheanesu Dzingirai. The vuk'uzenzele south african multilingual corpus, 2023.

[12] Phi Mots'o and Thabo Khoza. Sesotho news headlines sentiment dataset. *African NLP Journal*, 1(2):45–52, 2021.

[13] Salikoko S. Mufwene. *Language Evolution: Contact, Competition and Change*. Continuum, 2008.

[14] Xhosa Navy and Sibusiso Mkhize. Xhosanavy: A parallel corpus for isixhosa–english. In *Proceedings of LREC*, 2021.

[15] NCHLT Consortium. Nchlt south african bantu language resources. `http://nchlt.org.za`, 2020.

[16] Linda Ngqondi and Siyabonga Khumalo. Vixsd: Vuk'uzenzele isixhosa speech dataset. In *Proceedings of Interspeech*, 2023.

[17] Tolulope Ogunleye and Funmi Adetoro. Masakhanews: Topic classification benchmark for african languages. In *Proceedings of EMNLP*, 2022.

[18] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Gholollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. Masakhane—machine translation for africa. *arXiv preprint arXiv:2003.11529*, 2020.

[19] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, 2018.

[20] Jonathan Rosa and Nelson Flores. Title of the article. *Name of the Journal*, Volume number(Issue number):Page numbers, 2019.

[21] John Smith and Lethabo Mokoena. Speechreporting corpus for discourse phenomena. *Language Resources and Evaluation*, 54(3):123–138, 2020.

[22] C. Van Heerden and H. Pretorius. Safrisenti: A multilingual sentiment dataset for south african languages. In *Proceedings of LREC*, 2022.

[23] Lin Wang, Matthew Roberts, and Ling Zhao. Lora+: Enhanced low-rank adaptation for parameter-efficient tuning. In *Proceedings of ACL*, 2023.

[24] Xiaoyu Zhang, Anand Patel, Thu Nguyen, and Siddharth Kumar. Dylora: Dynamic low-rank adaptation for efficient fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.