



**Universität
Zürich** ^{UZH}

Department of Informatics
Department of Computational Linguistics

Leveraging Grammar Books to Create Synthetic Parallel Corpora for Low-Resource Languages

CLiC-it 2026 Conference Talk

Varun Ghat Ravikumar, Sina Ahmadi, Lena Jäger, Rico Sennrich

15 September 2026

Palermo, Italy

Motivation

Nearly **half of the world's 7,000+ languages are endangered**, and many may vanish within the next century if preservation efforts do not accelerate ([UNESCO, 2022](#)). Most low-resource languages don't have web data, except for **Descriptive Grammar Books** created by linguists.



Descriptive Grammar Books

- Contain rich linguistic rules
- Example parallel sentences from native speakers (less than 2k)
- Often contain wordlist or dictionary
- Unstructured prose in PDF format, not machine readable



Neural Machine translation

- Needs massive parallel corpora (10,000+ sentences) to generalize
- Requires clean, structured text (Source-Target pairs)
- Learns patterns and vocabulary from statistics, cannot apply explicit rules directly

UNESCO. 2022. A digital future for indigenous languages: Insights from partnerships forum.

Current Approaches & Limitations

1. Massively Multilingual Transfer (e.g., NLLB, Google Translate)

- Train one giant model on 200+ languages to encourage "positive transfer" (knowledge sharing).
- **Limitation:** Performance degrades sharply for languages completely unseen during pre-training. Finetuning requires parallel corpora.

2. Standard Data Augmentation (Back-Translation)

- Train a basic model to translate Target -> Source to create synthetic data
- **Limitation:** This requires an initial translation model. Lack of data to build even that first "seed" model.

3. LLM Zero-shot translation

- Prompting general-purpose LLMs (GPT-4, Gemini) to translate directly. Descriptive grammar and other resources could be in pre-training
- **Limitation:** LLMs produce fluent but hallucinated translations.

Grammar-Informed Approaches (Related Work)

[Tanzer et al. \(2024\)](#): Introduced "Machine Translation from One Book". Prompted LLMs with entire grammar book for Kalamang (Papuan language). Show that LLMs can learn from long context of grammar book. However, very large token cost per sentence translated.

[Aycock et al. \(2025\)](#): Investigated distinct book components (explanations vs. examples). Found that LLMs largely ignore abstract grammatical explanations, learning primarily from the embedded parallel sentences.

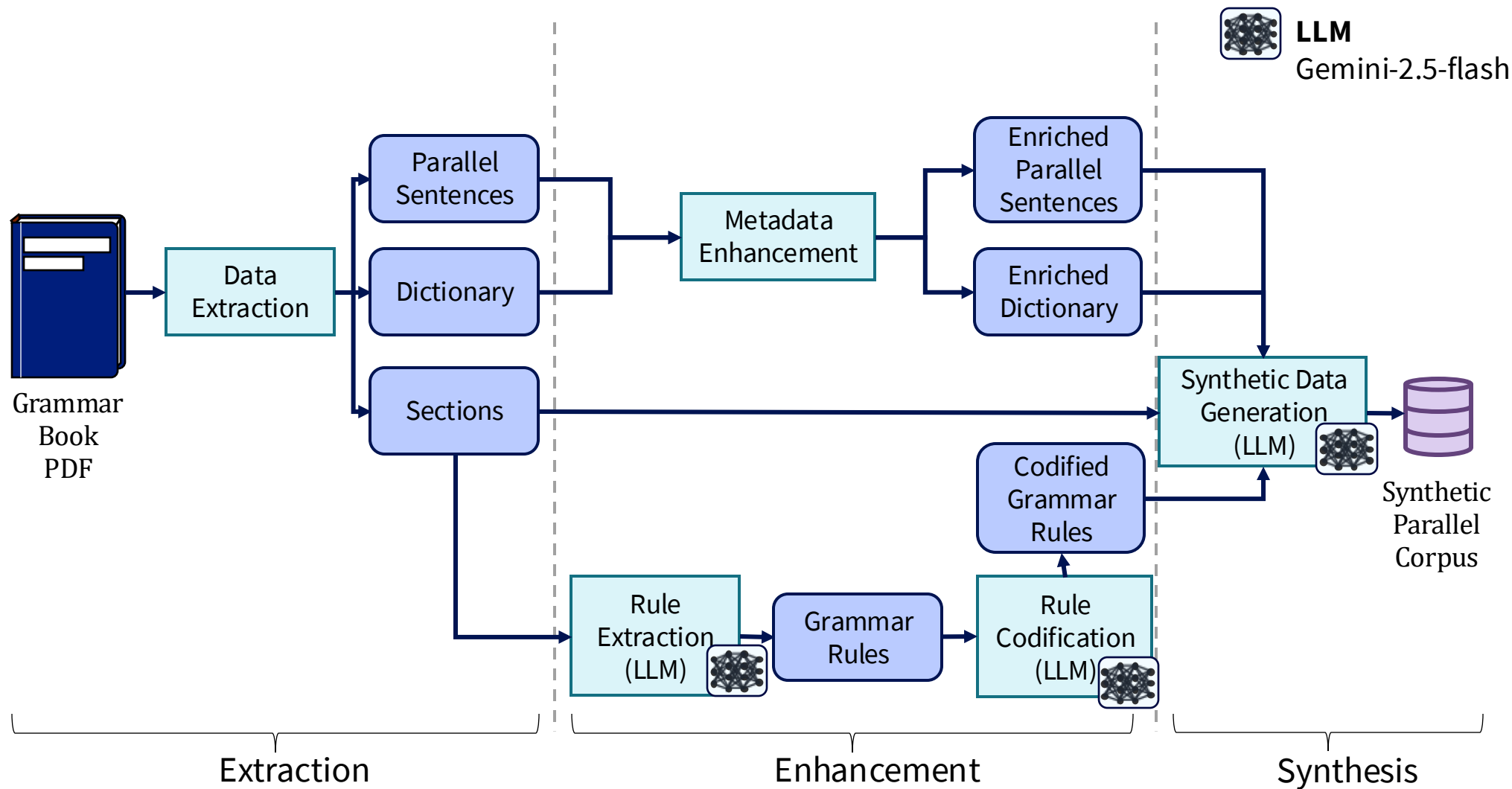
[Zhang et al. \(2025\)](#): Proposed two-stage rule selection, and converting natural language rules into executable code to guide translation for Zhuang -> Chinese. While effective, this relied on manual expert creation, creating a scalability bottleneck.

The Research Gap

Current methods lack a **fully automated pipeline** to extract linguistic rules and **synthesize parallel corpora** from static PDFs without human intervention.

Methodology

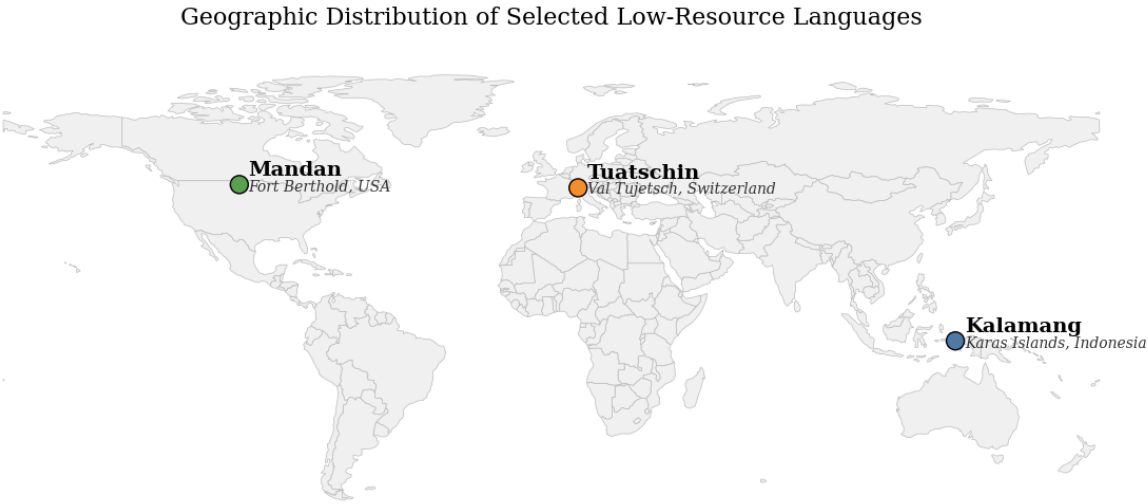
Pipeline Overview



Languages Selected and Data Sources

- **Kalamang (Papuan, Indonesia):** An endangered language from the Karas Islands with agglutinative morphology. Source: A Grammar of Kalamang (Visser, 2022).
- **Tuatschin (Romansh, Switzerland):** A dialect of Sursilvan Romansh exhibiting analytic structure. Source: A Grammar of Tuatschin (Maurer-Cecchini, 2021).
- **Mandan (Siouan, USA):** An extinct polysynthetic language with complex morphology. Being revitalized at Fort Berthold. Source: A Grammar of Mandan (Kasak, 2024).
 - Due to the lack of a comprehensive glossary in the grammar book, the lexicon was extracted from the external Comparative Siouan Dictionary (Rankin et al., 2015).

Language	Number of Speakers
Kalamang	~130
Tuatschin	~1500
Mandan	0 native speakers 6 semi-fluent learners and heritage speakers



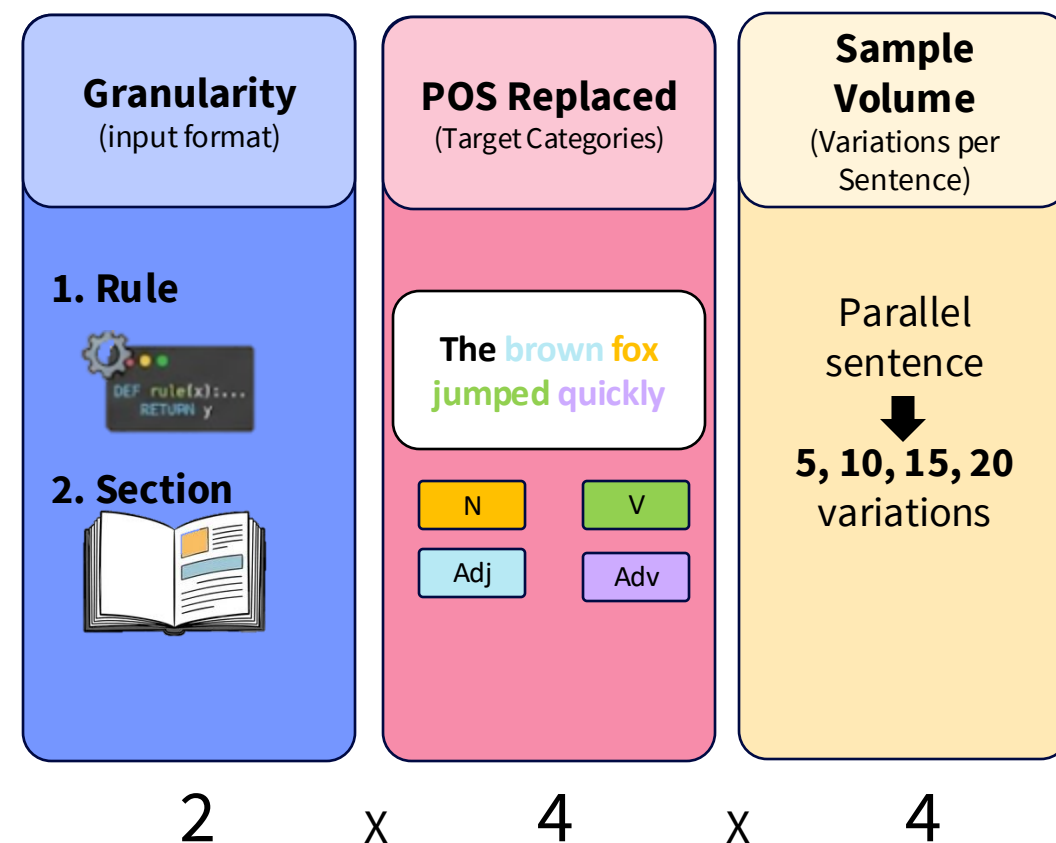
Step 4 – Evaluation

Evaluation metrics

- Finetuned Gemini-2.5-flash on generated synthetic data to translate from LRL -> English
- Test set used to evaluate model on BLEU, ChrF, ChrF++

Experimental Variables

- Context Granularity: Testing two input formats for the LLM:
 - Rule Granularity: Concise description + pseudo-code
 - Section Granularity: Raw paragraphs from the grammar book (broader context, higher noise).
- POS Replaced: Noun/ Verb / Adjective+Numeral / Adverb
- Sample Volume: Generating $k = 5, 10, 15, 20$ variations per sentence



Results

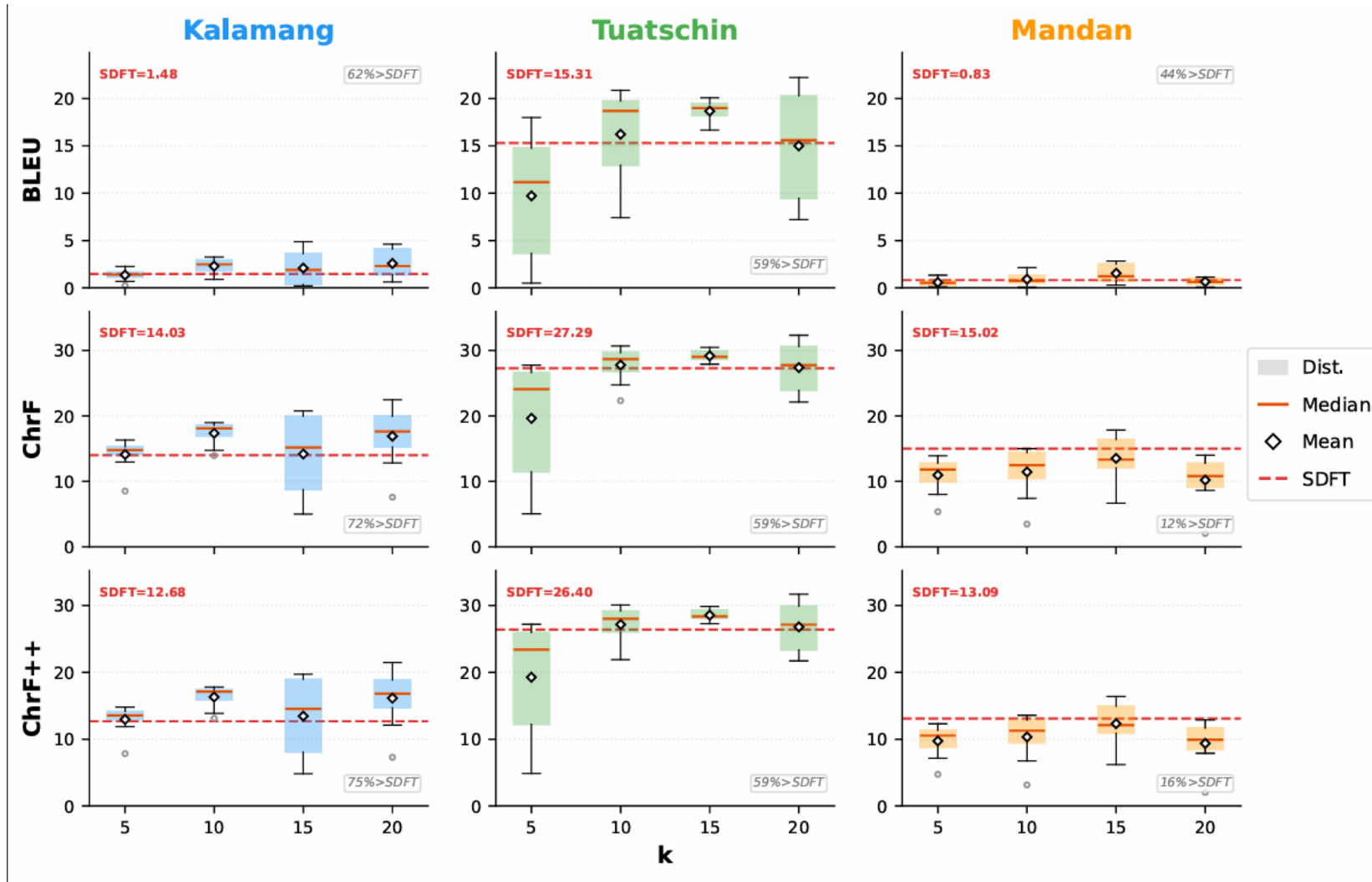
Results

Model	BLEU	ChrF	ChrF++
Language: Kalamang → English			
NLLB	1.12	10.69	9.23
Gemini (zero-shot)	0.77	12.60	11.27
GPT-4.1 (zero-shot)	1.65	13.01	11.34
GPT-5.2 (zero-shot)	1.89	13.25	12.03
ICL Gemini (rule)	3.77	26.91	25.32
ICL Gemini (section)	5.70	29.29	27.76
SDFT	1.48	14.03	12.68
Ours (mean ± SE)	2.09±0.25	15.65±0.78	14.71±0.75
Ours (best config)	4.61	22.47	21.48

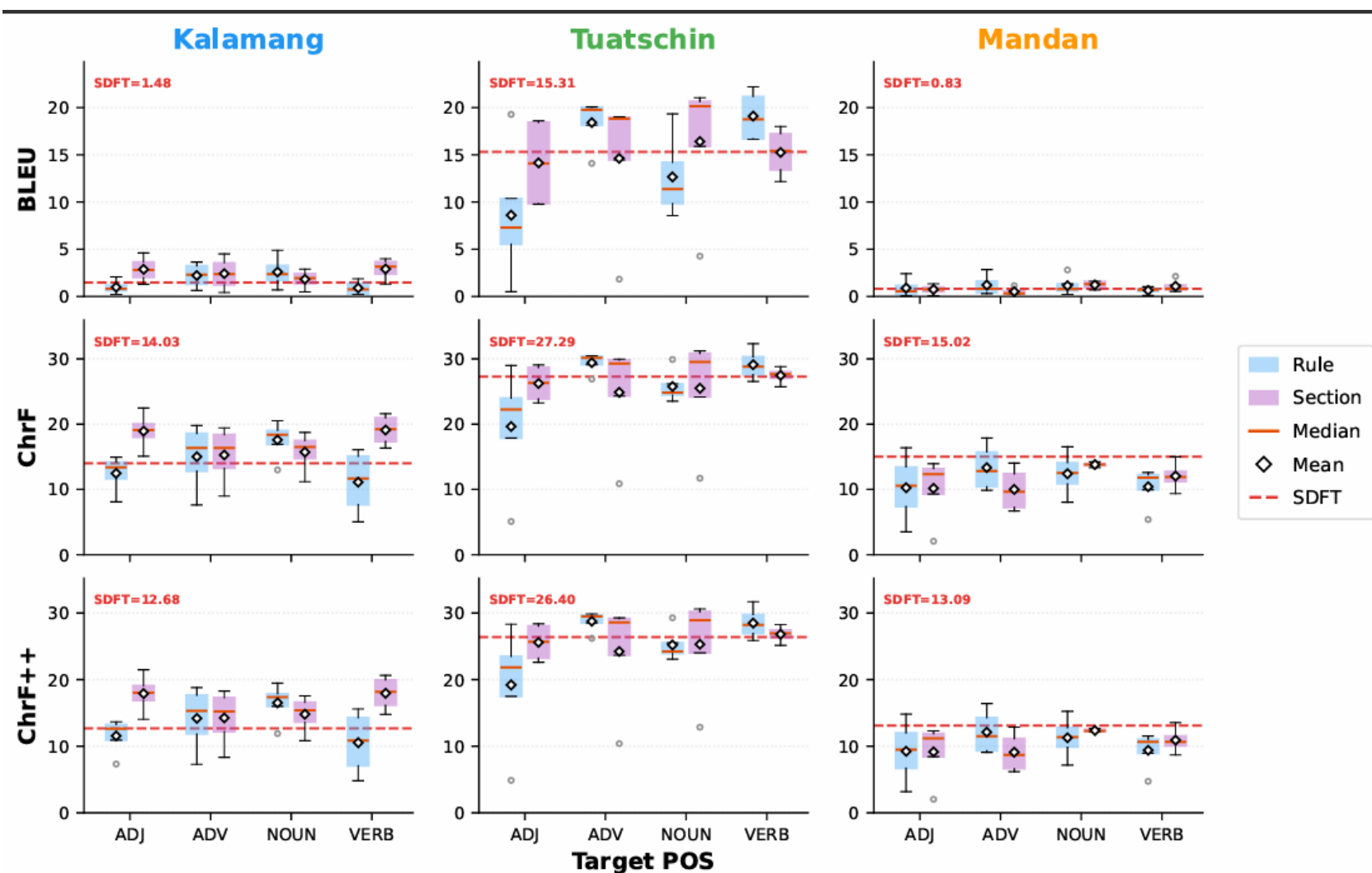
Model	BLEU	ChrF	ChrF++
Language: Tuatschin → English			
NLLB	7.92	13.07	12.20
Gemini (zero-shot)	8.32	23.37	22.28
GPT-4.1 (zero-shot)	12.81	23.48	22.64
GPT-5.2 (zero-shot)	10.39	20.47	19.58
ICL Gemini (rule)	7.82	22.64	21.49
ICL Gemini (section)	9.07	24.59	23.48
SDFT	15.31	27.29	26.39
Ours (mean ± SE)	14.90±1.06	26.00±1.09	25.44±1.05
Ours (best config)	22.21	32.33	31.68

Model	BLEU	ChrF	ChrF++
Language: Mandan → English			
NLLB	0.63	11.37	9.79
Gemini (zero-shot)	0.83	11.99	10.47
GPT-4.1 (zero-shot)	0.87	12.16	10.94
GPT-5.2 (zero-shot)	0.79	12.20	10.82
ICL Gemini (rule)	0.17	11.71	9.75
ICL Gemini (section)	0.76	13.34	11.24
SDFT	0.83	15.02	13.09
Ours (mean ± SE)	0.93±0.13	11.55±0.65	10.44±0.58
Ours (best config)	2.85	17.86	16.40

Analysis: Sample volume vs granularity



Analysis: Part-of-speech



Conclusion

- Developed and validated an end-to-end pipeline without manual expert intervention.
- Produces standalone corpora, not tied to any specific model: usable to train smaller, locally deployable systems that don't depend on commercial LLM APIs.
- Synthetic data performed better than SDFT, but cannot replace human-created text.
- Larger sample sizes generally lead to better output. but effects are configuration-dependent. No single POS or granularity is universally best
- LLMs still hallucinate even with grounding.

Limitations

- No native-speaker evaluation.
- One grammar-book per language: Patterns could be related to how the grammar book is written rather than the language
- Untested on tonal systems, non-concatenative morphology, or non-standard grammar book formats
- Lexical substitution lacks semantic co-occurrence constraints
- Pipeline depends on English-specific NLP tools throughout

Future Work

1. More languages, grammar books
2. Reverse-direction translation (English → low-resource language), validated with native speakers
3. Replace LLM-based extraction with smaller fine-tuned models

Thank you

Appendix

Appendix - Data Synthesis Examples

Kalamang → English: ADJ, section-level, $k=20$ <i>Original:</i> mu pas sem=ten=at koup <i>Translation:</i> “They hugged the scared woman.”	
tayuon	mu pas sem= tayuon =at koup → They hugged the not-good woman.
bes	mu pas sem= bes =at koup → They hugged the good woman.
siktaktak	mu pas sem= siktaktak =at koup → They hugged the slow woman.
Tuatschin → English: VERB, rule-level, $k=20$ <i>Original:</i> Nossadùna lèva dá li gjuven préjr... <i>Translation:</i> “The Holy Virgin wanted to give...”	
lèva	...lèva dá li gjuven préjr... → ...wanted to give the young priest...
saprèndev	...saprèndev ajn dá li gjuven... → ...took herself to give the young priest...
vulajèv	...vulajèv dá li gjuven préjr... → ...wanted to give the young priest...
Mandan → English: ADV, rule-level, $k=15$	
rəḥʃit pxinaateʔʃ → He sneezed in back of. wɨrɨ- wʔahka → I (acting on you) sneeze.	

Table 8
Representative synthesis outputs for each language’s best-performing configuration. Bold marks the substituted element.

#	Affix	Description
1	=at	OBJ case clitic on NOUN
2	=ko	LOC case on demonstrative ADJ
3	=ka	LAV case on interrogative ADJ
4	=saet	Superlative degree on ADJ
5	=ka	ALL case on demonstrative ADJ

Table 9
Candidate rules provided to the LLM for the Kalamang ADJ synthesis example above. Rules are retrieved via the scoring function (Eq. 1) and formatted as structured YAML.

Appendix - Qualitative Examples: Kalamang

Source	Reference	Model Output
[ror wa] me tabusik	This tree is short.	That tree is small.
Kapal cicauna kon	A small ship.	One is small, one is small.
som-kon oe-te kon~kon	Gave each person one, or gave each person one fruit.	(The child) has one hundred teeth, or (the child) has new one hundred teeth.
ma hukat= bon mia	He is coming with the net.	He/she hits the bottom net.
ma rap= et eba gier-un iriskap	When he laughs, his teeth are white.	He/she/it hits the whatsit, so his/her/its forehead is blue.
wa= nan im karuok weinun	[In] this [picture there are] three bananas too.	This is like a twenty net.
an tok tumun-an=at boubou=nin	I haven't yet bathed my child.	I haven't softened my child yet.
gier-ca me me ten	That tooth of yours is bad.	All the lobsters are dead.

Table 10
Translation examples for Kalamang → English (best model: ADJ-NUM, section-level, $k=20$).

Appendix - Qualitative Examples: Tuatschin

Source	Reference	Model Output
La mùma è vegnida dada in bi schénghèt看	Mother was given a beautiful present.	*The mother received a nice little present.
Quaj taur è bjals, ábar gròns sè 'l bét看	This bull is beautiful, but big he is not.	This bull is beautiful, but he is not big.
La scùla finéva . . . al mat看 . . . zarcladur	School ended in May ... June.	School ended ... in May ... weeder.
1920 [f4] vas schòn a paj	[that's what it takes you to] go on foot.	1920 [f4] you are already paying.
L' autar dé va ju gju la lubiantscha dad	I should go and make an appointment with those up there...	The original verb form is “gju”, identified as 1st Person Singular Past Tense...
Ju sùn dada gjù séla fatscha	I fell down on my face.	I am given down faces.
Las tgautschas fòn bét看 agl ùm	Trousers do not make a man.	My legs don't make the man.
Las nòtizjas sa ju bét看 danùndar als gjaniturs	I don't know where my parents had the news from.	The news I don't know where my parents had them from.

Table 11
Translation examples for Tuatschin → English (best model: VERB, rule-level, *k*=20).

Appendix - Qualitative Examples: Mandan

Source	Reference	Model Output
Paxirúukeena miní h̥i̥here	He was drinking water.	Cornsilk made him drink water.
Mí'shak máa'ək íwasek pshíiwahaani. . .	My land...	1sg.poss-pro land...
Kináatani máapte kaní'ni kasími réehoomako'sh	All around, there were very thick tall trees.	He got up again, climbed up the river bank, and he set off traveling.

Table 12
Translation examples for Mandan → English (best model: ADJ-NUM, rule-level, $k=15$).

Appendix: Corpus Statistics (Table 1)

Language	Dict	Parallel Sent.	Morph Rules	TTR	Sent. Len	Word Len
Kalamang	2,558	761	399	0.36	6.3	4.5
Tuatschin	1,788	981	144	0.24	11.6	3.8
Mandan	459	921	519	0.64	4.0	7.0

Table 1

Extracted information and corpus statistics based on type-token ratio (TTR), lengths measured in words and characters. Mandan’s high TTR (0.64) and word length (7.0) reflect on its polysynthetic morphology.

Appendix: Full Results - Kalamang (Table 5)

Language: Kalamang → English			Metrics (Value (Δ))		
POS	Granularity	Samples (k)	BLEU	ChrF	ChrF++
ADJ +NUM	rule	5	2.075 (+0.595)	14.943 (+0.918)	13.666 (+0.991)
		10	0.915 (-0.565)	13.956 (-0.069)	13.125 (+0.45)
		15	0.213 (-1.267)	8.098 (-5.927)	7.325 (-5.35)
		20	0.712 (-0.768)	12.818 (-1.207)	12.106 (-0.569)
	section	5	1.293 (-0.187)	15.095 (+1.07)	14.04 (+1.365)
		10	2.246 (+0.766)	18.992 (+4.967)	17.803 (+5.128)
		15	3.337 (+1.857)	19.196 (+5.171)	18.271 (+5.596)
		20	4.614 (+3.134)	22.466 (+8.441)	21.475 (+8.8)
ADV	rule	5	1.487 (+0.007)	14.523 (+0.498)	13.338 (+0.663)
		10	3.128 (+1.648)	18.178 (+4.153)	17.286 (+4.611)
		15	3.641 (+2.161)	19.777 (+5.752)	18.786 (+6.111)
		20	0.623 (-0.857)	7.601 (-6.424)	7.282 (-5.393)
	section	5	1.507 (+0.027)	14.66 (+0.635)	13.46 (+0.785)
		10	3.252 (+1.772)	18.041 (+4.016)	16.952 (+4.277)
		15	0.419 (-1.061)	8.982 (-5.043)	8.332 (-4.343)
		20	4.505 (+3.025)	19.419 (+5.394)	18.28 (+5.605)
NOUN	rule	5	0.694 (-0.786)	12.984 (-1.041)	11.892 (-0.783)
		10	2.032 (+0.552)	18.546 (+4.521)	17.343 (+4.668)
		15	4.891 (+3.411)	20.524 (+6.499)	19.467 (+6.792)
		20	2.739 (+1.259)	18.223 (+4.198)	17.353 (+4.678)
	section	5	2.267 (+0.787)	15.939 (+1.914)	14.538 (+1.863)
		10	2.897 (+1.417)	18.729 (+4.704)	17.55 (+4.875)
		15	0.484 (-0.996)	11.177 (-2.848)	10.816 (-1.859)
		20	1.635 (+0.155)	17.052 (+3.027)	16.276 (+3.601)
VERB	rule	5	0.268 (-1.212)	8.538 (-5.487)	7.835 (-4.84)
		10	1.251 (-0.229)	14.747 (+0.722)	13.838 (+1.163)
		15	0.215 (-1.265)	5.034 (-8.991)	4.822 (-7.853)
		20	1.87 (+0.39)	16.101 (+2.076)	15.591 (+2.916)
	section	5	1.311 (-0.169)	16.319 (+2.294)	14.779 (+2.104)
		10	2.725 (+1.245)	17.66 (+3.635)	16.613 (+3.938)
		15	3.601 (+2.121)	20.785 (+6.76)	19.729 (+7.054)
		20	3.987 (+2.507)	21.612 (+7.587)	20.63 (+7.955)

Table 5
Translation performance on Kalamang test sets across all experimental configurations. Results are organized by target POS, retrieval granularity, and sample volume (*k*). Values in parentheses show absolute improvement (Δ) over SDFT baseline. Bold values indicate best performance per metric.

Appendix: Full Results - Tuatschin (Table 6)

Language: Tuatschin → English			Metrics (Value (Δ))		
POS	Granularity	Samples (k)	BLEU	ChrF	ChrF++
ADJ +NUM	rule	5	0.505 (-14.804)	5.08 (-22.211)	4.875 (-21.522)
		10	7.418 (-7.891)	22.349 (-4.942)	21.909 (-4.488)
		15	19.29 (+3.981)	28.993 (+1.702)	28.305 (+1.908)
		20	7.2 (-8.109)	22.139 (-5.152)	21.705 (-4.692)
	section	5	9.834 (-5.475)	23.248 (-4.043)	22.59 (-3.807)
		10	18.382 (+3.073)	28.584 (+1.293)	27.928 (+1.531)
		15	18.617 (+3.308)	29.06 (+1.769)	28.382 (+1.985)
		20	9.764 (-5.545)	24.063 (-3.228)	23.437 (-2.96)
ADV	rule	5	14.096 (-1.213)	26.904 (-0.387)	26.196 (-0.201)
		10	19.478 (+4.169)	29.881 (+2.59)	29.278 (+2.881)
		15	20.075 (+4.766)	30.46 (+3.169)	29.851 (+3.454)
		20	20.029 (+4.72)	30.395 (+3.104)	29.665 (+3.268)
	section	5	1.831 (-13.478)	10.87 (-16.421)	10.392 (-16.005)
		10	18.994 (+3.685)	28.783 (+1.492)	28.073 (+1.676)
		15	18.675 (+3.366)	29.96 (+2.669)	29.27 (+2.873)
		20	19.023 (+3.714)	29.819 (+2.528)	29.142 (+2.745)
NOUN	rule	5	12.485 (-2.824)	24.925 (-2.366)	24.198 (-2.199)
		10	10.267 (-5.042)	24.759 (-2.532)	24.252 (-2.145)
		15	19.346 (+4.037)	29.901 (+2.61)	29.281 (+2.884)
		20	8.568 (-6.741)	23.521 (-3.77)	23.058 (-3.339)
	section	5	4.27 (-11.039)	11.705 (-15.586)	12.855 (-13.542)
		10	20.545 (+5.236)	30.709 (+3.418)	30.072 (+3.675)
		15	19.763 (+4.454)	28.314 (+1.023)	27.714 (+1.317)
		20	21.048 (+5.739)	31.226 (+3.935)	30.605 (+4.208)
VERB	rule	5	16.698 (+1.389)	26.544 (-0.747)	25.863 (-0.534)
		10	20.847 (+5.538)	29.657 (+2.366)	29.122 (+2.725)
		15	16.65 (+1.341)	27.924 (+0.633)	27.275 (+0.878)
		20	22.209 (+6.9)	32.334 (+5.043)	31.677 (+5.28)
	section	5	18.002 (+2.693)	27.774 (+0.483)	27.183 (+0.786)
		10	13.832 (-1.477)	27.51 (+0.219)	26.666 (+0.269)
		15	16.981 (+1.672)	28.814 (+1.523)	28.266 (+1.869)
		20	12.176 (-3.133)	25.727 (-1.564)	25.136 (-1.261)

Table 6
Translation performance on Tuatschin test sets across all experimental configurations. Results are organized by target POS, retrieval granularity, and sample volume (*k*). Values in parentheses show absolute improvement (Δ) over SDF baseline. Bold values indicate best performance per metric.

Appendix: Full Results - Mandan (Table 7)

Language: Mandan → English			Metrics (Value (Δ))		
POS	Granularity	Samples (k)	BLEU	ChrF	ChrF++
ADJ +NUM	rule	5	0.746 (-0.081)	12.453 (-2.569)	11.075 (-2.014)
		10	0.074 (-0.753)	3.499 (-11.523)	3.16 (-9.929)
		15	2.42 (+1.593)	16.387 (+1.365)	14.819 (+1.73)
		20	0.324 (-0.503)	8.659 (-6.363)	7.889 (-5.2)
	section	5	1.34 (+0.513)	13.942 (-1.08)	12.296 (-0.793)
		10	0.62 (-0.207)	11.693 (-3.329)	10.579 (-2.51)
		15	0.885 (+0.058)	12.933 (-2.089)	11.753 (-1.336)
		20	0.083 (-0.744)	2.053 (-12.969)	2.033 (-11.056)
ADV	rule	5	0.307 (-0.52)	10.593 (-4.429)	9.387 (-3.702)
		10	1.223 (+0.396)	15.031 (+0.009)	13.576 (+0.487)
		15	2.847 (+2.02)	17.856 (+2.834)	16.395 (+3.306)
		20	0.425 (-0.402)	9.853 (-5.169)	9.094 (-3.995)
	section	5	0.363 (-0.464)	11.842 (-3.18)	10.57 (-2.519)
		10	0.259 (-0.568)	7.409 (-7.613)	6.767 (-6.322)
		15	0.299 (-0.528)	6.69 (-8.332)	6.186 (-6.903)
		20	1.144 (+0.317)	14.028 (-0.994)	12.86 (-0.229)
NOUN	rule	5	0.22 (-0.607)	8.032 (-6.99)	7.152 (-5.937)
		10	0.835 (+0.008)	13.281 (-1.741)	11.958 (-1.131)
		15	2.806 (+1.979)	16.529 (+1.507)	15.237 (+2.148)
		20	0.727 (-0.1)	11.776 (-3.246)	10.764 (-2.325)
	section	5	0.704 (-0.123)	13.86 (-1.162)	12.129 (-0.96)
		10	1.585 (+0.758)	14.221 (-0.801)	12.642 (-0.447)
		15	1.602 (+0.775)	13.741 (-1.281)	12.46 (-0.629)
		20	1.037 (+0.21)	13.372 (-1.65)	12.18 (-0.909)
VERB	rule	5	0.109 (-0.718)	5.382 (-9.64)	4.733 (-8.356)
		10	0.665 (-0.162)	11.489 (-3.533)	10.369 (-2.72)
		15	0.76 (-0.067)	12.133 (-2.889)	10.907 (-2.182)
		20	1.024 (+0.197)	12.607 (-2.415)	11.519 (-1.57)
	section	5	0.893 (+0.066)	11.815 (-3.207)	10.515 (-2.574)
		10	2.132 (+1.305)	14.979 (-0.043)	13.557 (+0.468)
		15	0.759 (-0.068)	11.996 (-3.026)	10.846 (-2.243)
		20	0.534 (-0.293)	9.361 (-5.661)	8.692 (-4.397)

Table 7
Translation performance on Mandan test sets across all experimental configurations. Results are organized by target POS, retrieval granularity, and sample volume (*k*). Values in parentheses show absolute improvement (Δ) over SDFt baseline. Bold values indicate best performance per metric.

Appendix: Official Pipeline Diagram (Figure 1)

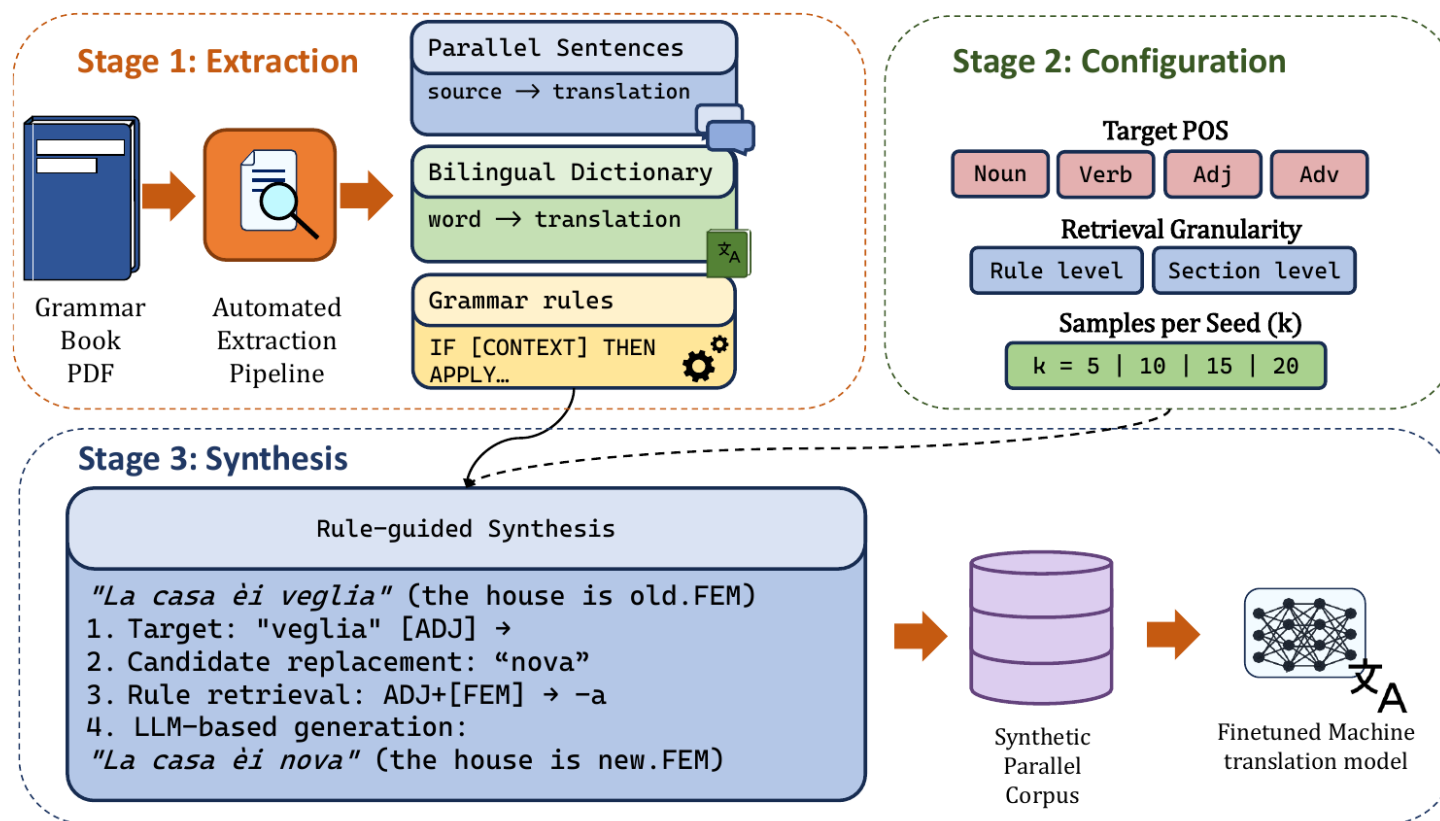


Figure 1: Overview of our methodology. Stage 1 extracts linguistic information (parallel sentences, dictionaries, rules). Stage 2 shows the combinatorial experimental configuration. Stage 3 generates synthetic variants via rule-guided lexical substitution (Tuatschin example shown).

Appendix: Official Experimental Design (Figure 2)

substitution considering three factors illustrated in Figure 2.

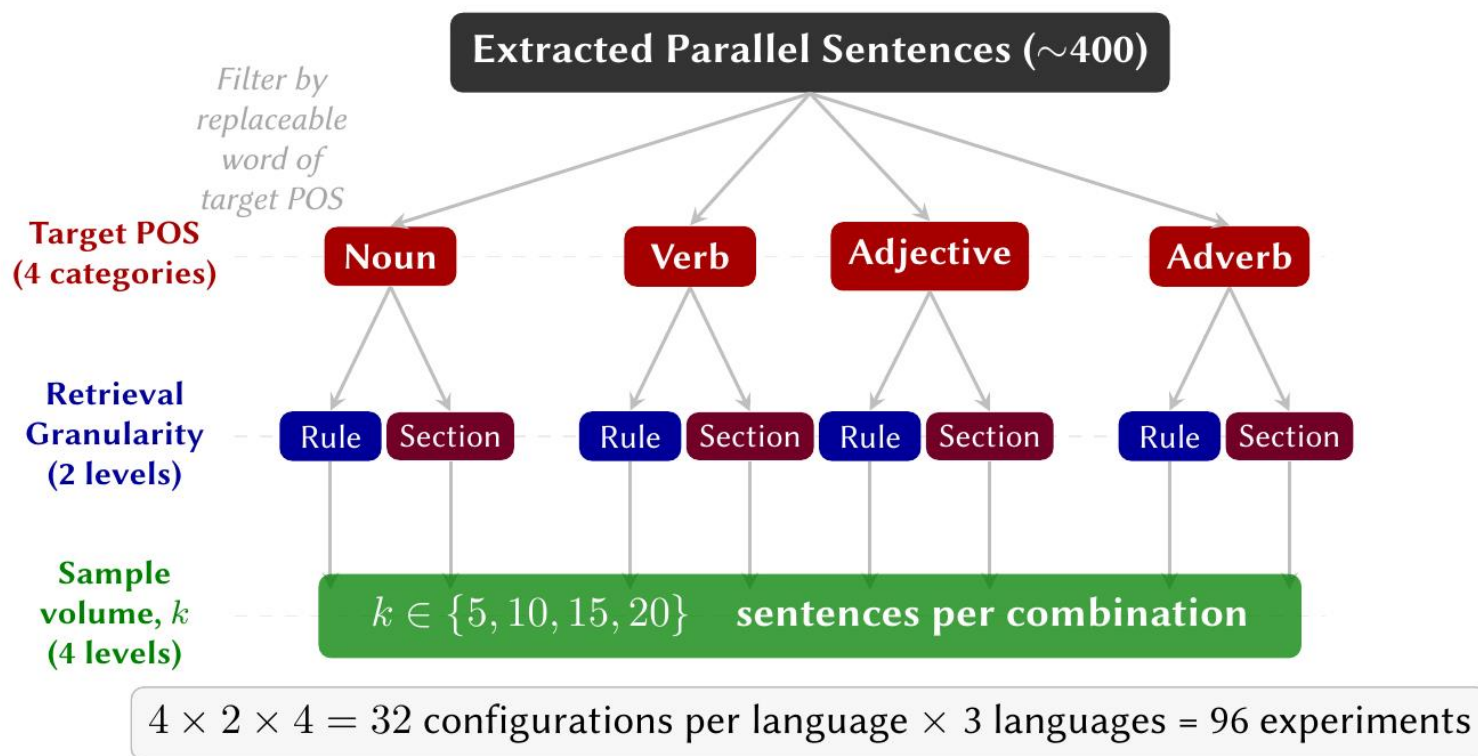


Figure 2: Experimental design. The seed corpus is filtered by target part-of-speech to select sentences containing a replaceable word of that category. Each subset is varied along retrieval granularity and sample volume k . All 32 configurations are evaluated independently per language.

Appendix: Results by Sample Size, Official (Figure 3)

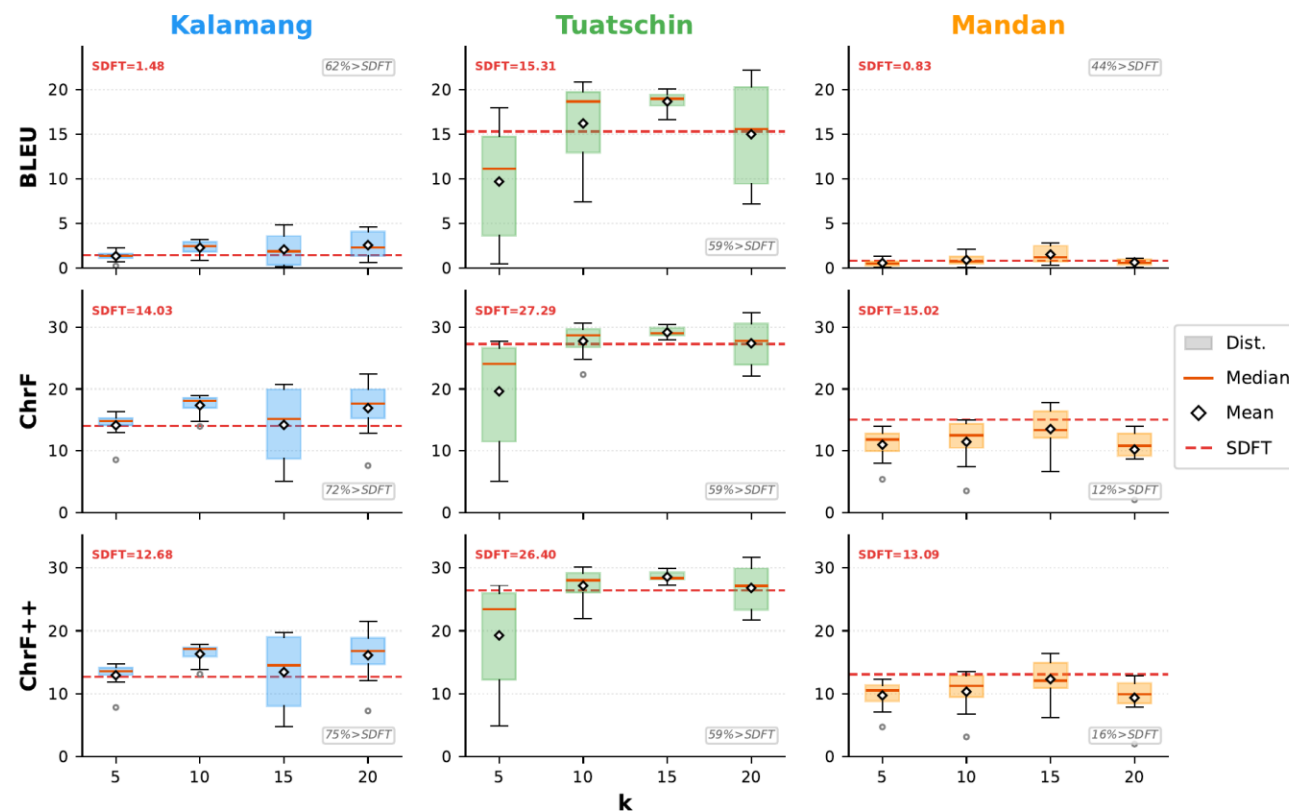


Figure 3: Distribution of translation quality across all 32 configurations per language, grouped by sample volume k . Red dashed lines mark the SDFT baseline. Kalamang exceeds the baseline at all k values in most configurations. Tuatschin fails at $k=5$ but recovers at higher volumes. Mandan rarely exceeds the baseline, with performance peaking at $k=15$ before degrading sharply at $k=20$.

Appendix: Results by POS, Official (Figure 4)

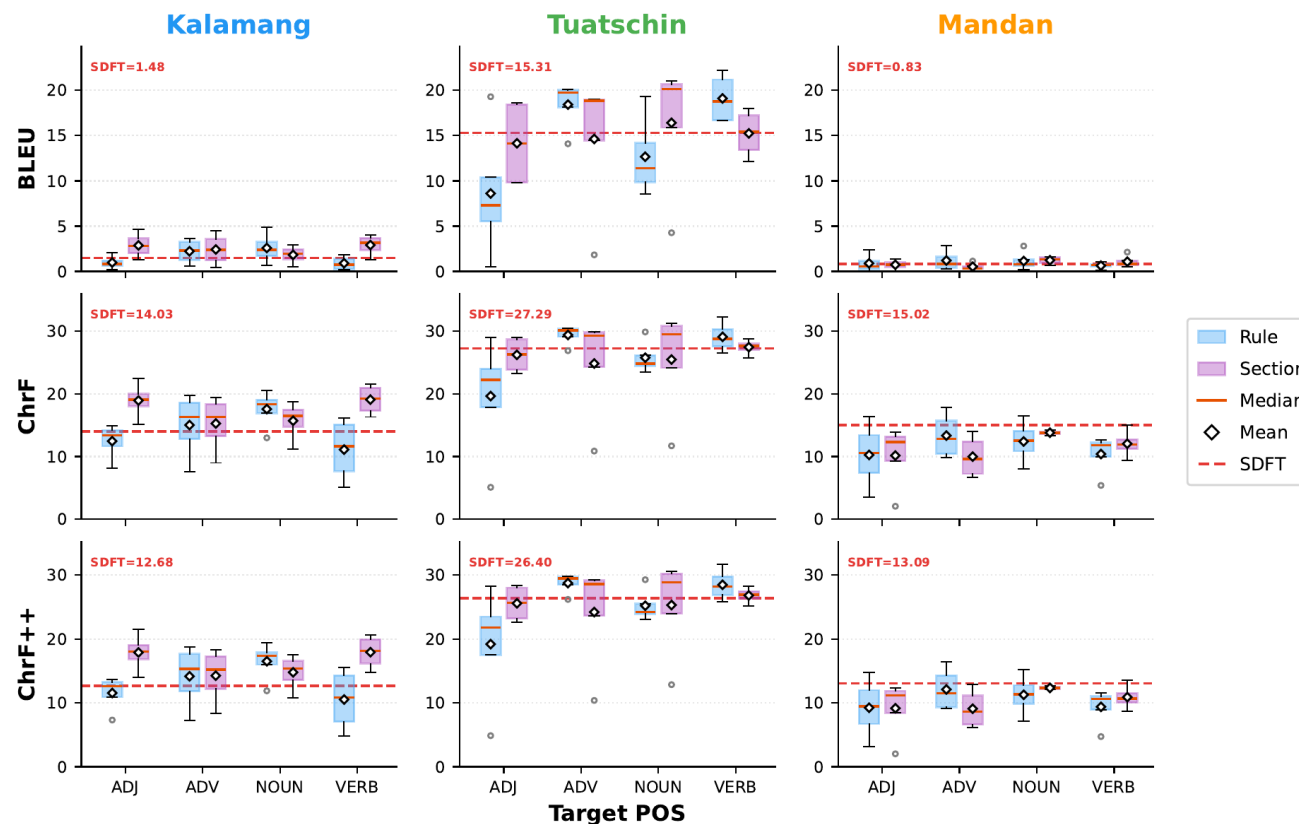


Figure 4: Translation quality by target POS and retrieval granularity, aggregated across all sample volumes. For Kalamang, section-level retrieval produces higher medians across most POS categories. For Tuatschin, adjective configurations show high variance while verbs are most stable. Mandan rarely exceeds the SDFT baseline (red dashed lines) regardless of POS or granularity

Appendix: Prompt - Rule Extraction (Figure 5)

You are an expert descriptive linguist and data structurer. Your core task is to **extract morphological and syntactic rules** from a grammar book paragraph for a low-resource language. These rules will be used by a downstream AI agent to synthesize valid sentences.

Selection Criteria

Extract a rule IF AND ONLY IF:

1. It involves an **overt surface change** (suffix, prefix, clitic, particle, or mutation).
2. It is **productive** (applies generally to a class of words, not one specific exception).
3. It has a clear **grammatical function** (Case, Tense, Mood, Aspect, Agreement).

Do NOT extract: purely phonological rules, typological trivia, examples, or footnotes.

Output Schema (YAML)

```
category: surface_rule
description: <POS, morpheme, and grammatical function>
target_pos: <NOUN | VERB | ADJ | PRON>
affix_type: <SUFFIX | PREFIX | CLITIC | PARTICLE | ...>
morpheme: <literal string, e.g., "-at", "ko=">
application_string: <e.g., STEM + "-an">
unimorph_feature: <e.g., CASE, TENSE, AGR, NUM>
unimorph_value: <e.g., ACC, PST, PL, 3SG>
context_dependency: <phonological/agreement environment, or N/A>
semantic_trigger: <when to apply this rule>
```

Output Handling

If no rule is found: output category: N/A and description: "No applicable rule found."

Task

Process the following paragraph: {input_paragraph}

Figure 5: Prompt template for grammar rule extraction. The model outputs one YAML block per extracted surface rule, or a fixed N/A block if no applicable rule is found. The {input_paragraph} placeholder is filled with a single paragraph from the parsed grammar book.

Appendix: Prompt - Rule Codification (Figure 6)

```
You are a computational linguist.
Task: Translate the single input rule into a concise pseudo-code function, ensuring the
output YAML is perfectly parsable.

Output Schema (YAML)
canonical_description: <single sentence synthesizing
                        meaning, usage, and all conditions>
lrl_code: <pseudo-code function applying the rule>

Codification Rules
1. Function MUST be named ApplyRule(STEM, POS).
2. Affixation must use the application_string field.
3. Implement all conditions with IF/ELSE IF checks
   (e.g., IF STEM_ENDS_IN('x'), IF POS == 'Y').
4. If conditions are not met, MUST return the original
   STEM.

Input Rule (YAML)
Process this rule.
-
{input_rule}
-
```

Figure 6: Prompt template for rule codification. Each extracted surface rule (in YAML) is converted into a canonical_description and a deterministic ApplyRule(STEM, POS) pseudo-code function. The codified rules are indexed for semantic retrieval and serve as the grammar representation in the *rule*-granularity generation condition.

Appendix: Prompt - Synthetic Generation (Figure 7)

```
You are a professional linguist for {lang}.
Task: Replace the {pos} in the sentence using provided rules. Output must be concise.

Constraints
- Deduce existing clitics by comparing the original sentence to the provided stem.
- Allomorphy: Apply the correct variant based on the new stem's ending.
- You MUST cite Rule #s for all changes.

Inputs
- {granularity_statement}: {rules_data}
- Sentence: {sentence_text}
- Translation: {translation}
- Target:      Replace {stem_to_replace} ({original_gloss_features}) with {word}
               ("{word_translation}") [POS: {pos}]

Step-by-step reasoning
1. Context: [Target word] + [Deduced clitics]
2. Rule Selection: [Rule #s] + [Justification]
3. Transformation: [New stem] + [Applied changes]
4. Final Sentence: [{lang} result]
5. Translation: [English result]

Output format
final_sentence: "...
english_translation: "...
```

Figure 7: Prompt template for grammar-aware synthetic generation. Placeholders (in braces) are filled per language and configuration. The granularity_statement field contains either codified rules or full grammar book sections depending on the retrieval granularity condition.

Appendix: Feature Extraction Reference (Tables 3-4)

Stage	Feature	Usage
Dict. filtering	head_pos	Assigns entries to POS pools
Rule selection	tense	Selects tense-related rules
	number	Selects agreement rules
	gender	Selects gender-related rules
	Case	Selects case-specific rules

Table 3
Morphosyntactic features actively used in the pipeline. All features are extracted by spaCy from English translations.

Lang.	Entry	Definition	Head	POS	Morph.
<i>Single-word definitions</i>					
KLM	<i>alar</i>	fish	fish	NOUN	Num=Sg
TUA	<i>aua</i>	water	water	NOUN	Num=Sg
<i>Multi-word definitions</i>					
KLM	<i>emnem</i>	old woman	woman	NOUN	Num=Sg
TUA	<i>cavagl</i>	horse	horse	NOUN	Num=Sg

Table 4
spaCy analysis of representative dictionary entries. HEAD indicates the syntactic head used for POS assignment. Single-word entries are tagged in isolation; multi-word entries are parsed for head identification.

Appendix: Prompts - ICL Baseline & Fine-tuning (Figures 8-9)

```
You are a professional linguist for {lang}.
Task: Translate the sentence into English using the provided grammatical rules. Output
must be concise. Use fragments, not full paragraphs.

Constraints
- Use the provided rules to interpret morphological forms in the sentence.
- Use the dictionary entries to identify known word meanings.
- Citation: You MUST cite Rule #s used in your reasoning.

Inputs
- {granularity_statement}:
{rules_data}
- Sentence: {sentence_text}
- Dictionary (source words found in lexicon):
{dictionary_entries}

Step-by-step reasoning (Direct & Brief)
1. Segmentation: [Words + clitics/morphology identified].
2. Rule Selection: [Rule #s] used + [1-phrase justification].
3. Word Meanings: [Known words from dictionary].
4. Translation: [English result].

Output format
english_translation: "..."
```

Figure 8: Prompt template for the grammar-aware ICL translation baseline. Unlike the generation prompt (Figure 7), there is no target word to substitute; instead the model translates the full source sentence directly. The `granularity_statement` field selects between codified rules (*rule* condition) and raw grammar book paragraphs (*section* condition). Dictionary entries are retrieved by tokenising the source sentence and looking up sub-morphemes in the bilingual lexicon.

User turn

Translate to english from {language_name}: {source_sentence}

Model turn (training target only)

{english_translation}

Figure 9: Prompt template used for Gemini fine-tuning and baseline batch inference. Each example is formatted as a two-turn conversation: the user provides the source sentence prefixed with a language identifier, and the model response is the English translation. At inference time only the user turn is sent; the model turn is omitted.