Can LLMs Help Create Grammar?: Automating Grammar Creation for Endangered Languages with In-Context Learning

Piyapath T Spencer[1,2] and **Nanthipat Kongborrirak**[1]
[1]Language and Information Technology Programme, Faculty of Arts, CU, Thailand
[2]Center for Information and Language Processing (CIS), LMU Munich, Germany
linguistics@piyapath.uk prompt.k10@gmail.com

Abstract

In the present-day documenting and preserving endangered languages, the application of Large Language Models (LLMs) presents a promising approach. This paper explores how LLMs, particularly through in-context learning, can assist in generating grammatical information for low-resource languages with a limited amount of data. We take Moklen as a case study to evaluate the efficacy of LLMs in producing coherent grammatical rules and lexical entries using only bilingual dictionaries and parallel sentences of the unknown language without building the model from scratch. Our methodology involves organising the existing linguistic data and prompting to efficiently enable to generate formal XLE grammar. Our results demonstrate that LLMs can successfully capture key grammatical structures and lexical information, although challenges such as the potential for English grammatical biases remain. This study highlights the potential of LLMs to enhance language documentation efforts, providing a cost-effective solution to generate linguistic data and contributing to the preservation of endangered languages.

1 Introduction

In linguistics, the quest for universality (Chomsky and Berwick, 2016), despite its criticism, motivates the comparison of languages to identify common patterns and structures. Many linguists investigate how different languages handle concepts like tense, number, or syntactic structure to identify universal features or constraints.

Within this broader discourse, Lexical-Functional Grammar (LFG) offers a model for understanding the complexities of different information within and across languages. In its different structures, it assumes both variability and universality. For instance, the analysis of a language with a free word order sequence is valid within the LFG framework (Simpson, 1991).

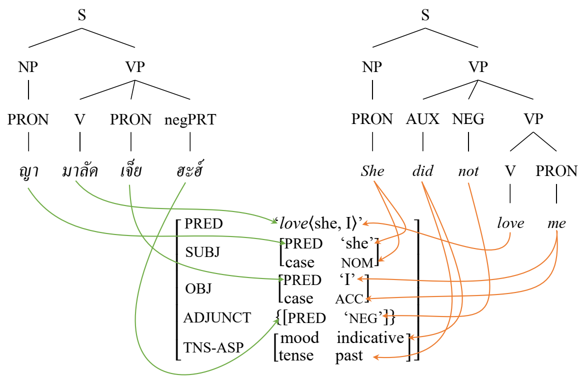

Figure 1: The ϕ mapping (Dalrymple, 2023) from different c(onstituent)-structures of Moklen and English to the same f(unctional)-structure

According to Butt et al. (1999), "a common set of linguistic principles and a commonly agreed upon set of grammatical analyses and features" unify linguistic insights across languages whilst acknowledge the uniqueness of individual languages.

The XLE (Xerox Linguistic Environment), developed on the basis of LFG (Maxwell and Kaplan, 1993; Crouch et al., 2011), serves as a powerful tool for the computational implementation of multilingual grammars (grammars across a wide range of languages). However, the 'grammars' are usually manually constructed since it requires the understanding of nuances within a language, which is expensive and time-consuming to produce. This fact makes it difficult to start developing deep grammar for low-resource languages, especially endangered languages.

The recent development of large language models (LLM) has shown notable capabilities for 'adaptation' (Brown et al., 2020; Wei et al., 2022) to various tasks and understanding natural language instructions through in-context learning and understanding natural language instructions with in-context learning (Brown et al., 2020; OpenAI, 2024; Team, 2024). Their linguistic performance

Proceedings of the 31st International Conference on Computational Linguistics, pages 10214–10227
January 19–24, 2025.

and metalinguistic abilities have also been recognised (Beguš et al., 2023). This presents an opportunity to use their capabilities to generate grammar and linguistic information for endangered languages, potentially overcoming the need for exhaustive resources and the limitations of manual grammar construction.

This paper proposes a novel approach that exploits LLMs' linguistic competences in English language to generate such coherent formal language as XLE grammar for natural languages that have not been encountered during the model's pre-training, as well as previously undocumented ones. This approach relies solely on bilingual dictionaries and a limited number of parallel sentence pairs, reflecting the typical data available to linguists during the initial documentation of endangered languages (Spencer, 2024). Importantly, this method is not intended to replace traditional workflows; rather, it aims to serve as a complementary tool to assist linguists in both fieldwork and theoretical analysis.

2 Background and Related Work

Despite the significant advancements, most existing studies focus on high-resource languages, with limited exploration of how LLMs can be effectively applied to lower-resource or *truly* low-resource languages. This gap is particularly concerning given the vast number of languages spoken worldwide, many of which lack sufficient digital representation and the risk of extinction.

On Large Multilingual Grammar Development
Large multilingual grammar development has seen significant advancements through theoretical grammar models such as Head-Driven Phrase Structure Grammar (HPSG) and Lexical-Functional Grammar (LFG). Such projects include LinGO (Oepen et al., 2002), Matrix (Zamaraeva et al., 2022), and DELPH-IN (Copestake, 2002) for HPSG, and the ParGram project (Butt et al., 2002) for LFG. These frameworks often rely on rule-based systems to capture the syntactic and semantic nuances of multiple languages. Their common objective is to create comprehensive grammatical analyses that can be applied across diverse linguistic contexts. However, the manual construction of these grammars demands extensive linguistic expertise and resources, which can be considered as a barrier for low-resource languages.

On '*Low-Resource*'
In the fields of artificial intelligence (AI), natural language processing (NLP), and LLMs, the term "low-resource" is surprisingly is a broad. This classification even includes such languages as German, Filipino and some institutional languages that have their own pretrained models, machine translation (MT) systems, and various NLP applications. This disparity arises from the uneven distribution of linguistic resources, with only 14 languages comprising over 90% of internet content, and English alone accounting for half of all data (W3Techs). In contrast, there are over 7,000 known languages, with approximately half classified as endangered according to Ethnologue. Many of these endangered languages are likely to disappear by the end of this century.

When considering endangered languages alongside the NLP concept of low-resource languages, it becomes evident that many of these languages even lacks native speakers in their own society, not to mention its minimal or no digital presence. Often, they are primarily spoken languages without a writing system, meaning there may be no written documentation or corpus available. Despite these challenges, linguists have made efforts to document these sorts of language by, firstly, collecting vocabulary, creating grammar books and dictionaries to preserve their existence of language in the world.

On Competence and Performance
Psycholinguistically, a distinction exists between linguistic competence and performance (Pritchett, 1992). Competence refers to the innate ability to control all aspects of a language's structure, ranging from the intricate array of grammatical rules to pragmatic nuances of usage. Performance, on the other hand, pertains to the actual production of language in real-world contexts. This distinction differentiates a critical difference between human linguistic abilities and those of AI, as LLMs possess extensive knowledge of a language due to their rigorous pretraining on large datasets.

When discussing generative AI, the focus is often on how language is created or generated, as suggested by the term itself. Generative AI, including LLMs, thus simulates human linguistic performance (Miranda-Saavedra, 2024). In the context of linguistics, this raises questions about how language is produced and understood. Over time, LLM benchmarks have shown saturation, indicating that models are becoming increasingly powerful and capable of performing complex tasks, as exemplified by models like GPT-4. This leads to comparisons between AI capabilities and human language abilities, prompting inquiries into whether LLMs truly understand language.

Generally, the performance of LLMs is largely determined by both quality and, but much more important, quantity of the data on which they are trained. Whilst LLMs may demonstrate impressive capabilities across various languages, this does not necessarily guarantee a true understanding of those languages. This distinction is particularly evident in the case of English, where LLMs benefit from extensive training in vast and diverse datasets. As a result, their performance in English is often more robust and nuanced compared to their performance in lower-resource languages, which may not have the same level of data availability and linguistic representation.

On In-Context Learning
Recent advances in LLMs have improved their performance not by training from scratch or requiring extensive datasets, but through prompt engineering that enables in-context learning. This approach allows models to adopt information presented within the context of specific inputs to generate relevant and coherent responses. Various techniques, such as zero-shot, few-shot, and chain-of-thought prompting, have emerged as effective strategies for enhancing LLM capabilities.

In the context of low-resource languages, in-context learning has been employed to refine LLM performance, particularly in machine translation. For instance, Tanzer et al. (2024) utilised dictionaries and grammar books to translate endangered languages with LLMs, establishing benchmarks for evaluation. Zhang et al. (2024), whose work this paper aims to build upon, extended this approach to cover multiple NLP tasks, evaluating their LINGOLLM methodology in different endangered languages. In contrast to these studies, we seek to advance the application of LLMs in a more radical linguistic task: analysing the grammar of languages under extremely limited conditions where prior data is not applicable.

3 Language under Study: Moklen Language

Moklen (ISO 639-3: `mkm`) is an endangered language estimated to be spoken by fewer than 1,000 people, which constitutes only a quarter of the total population, most of whom are over 50 years old. The speakers are predominantly scattered across Phang Nga and Phuket in Southern Thailand. Moklen has been influenced by Malay and Thai, the latter being the national language of Thailand. It is classified as one of the Austronesian languages.

Despite attempts by the community to use the Thai alphabet to record and teach Moklen to children, the language remains primarily oral, with no formal written tradition or administrative use. Consequently, there is little to no evidence of the language being used on the internet. Moklen is analysed to have a subject-verb-object (SVO) word order and a nominative-accusative alignment. While this word order is prominent, Moklen also features topicalisation to emphasise certain parts of an utterance. Additionally, it lacks inflectional morphology, meaning that words can be combined in various ways to convey meaning. Grammatical features such as tense, aspect, and number are expressed through content words.

Thorough documentation of the language began in late 2017, leading to the development of a pilot version of a dictionary (Pittayaporn et al., 2022), following the establishment of a Thai-based orthography system (Pittayaporn and Choemprayong, 2024). Nevertheless, there are only a few samples of Moklen sentences, primarily derived from field notes, and limited work has been done on the language.

Taking into account all these aspects, Moklen is well-suited for the current study. Firstly, it meets the requirements of the paper's approach, which necessitates bitext and a dictionary with the source language (Moklen) and the target language (English). Secondly, the dictionary is relatively comprehensive and contains words that can be used to express a wide range of concepts. Finally, the isolating nature of the Moklen syntax allows for the application of LLMs without the complications of inflectional morphology and complex tokenisation.

4 Methodology

The proposed methodology aims to simulate the steps linguists use to analyse the grammar of a language. It closely mirrors the interlinear glossing method employed by linguists. The process begins with a sentence from the source language, which is then tokenised based on available observations or dictionary entries. Simultaneously, each tokenised word is mapped to its corresponding translation in the target language, creating an interlinear gloss. Following this, an attempt is made to translate the

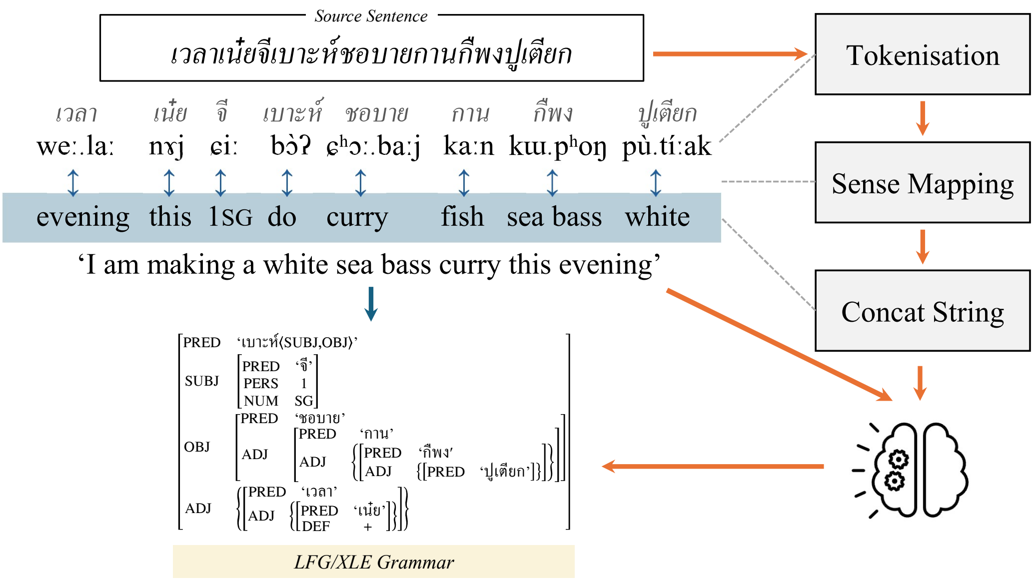

Figure 2: Our approach reflects the approach linguists take when analysing grammar and the cognitive aspects of human language. To implement this, each pair of sentences from the bilingual text (bitext) must undergo tokenisation, sense mapping, and sentence concatenation before being included in the prompt for analysis.

entire sentence. This allows for a comparison of the structures of both languages, facilitating the development of grammatical rules for the source language.

When generating grammar using an LLM, the approach largely follows these steps but also requires a complete translation of the sentence. This is essential for providing the LLM with knowledge of the target language as to serve as a reference for comparison. To be precise, the process comprises five major steps, as shown in Figure 2:

(1) Given bitext, a dictionary-based tokeniser is used to segment the Moklen text into individual words.

(2) For every word in a sentence, we search for the closest match from the dictionary to map the meaning of English to the corresponding words in the source language.

(3) We concatenate the meanings of each word in each sentence and pair them side-by-side with the original sentences.

(4) We prompt a large language model (LLM) with the mapped, tokenised bitext and materials related to creating grammar in XLE format.

(5) We then use the generated grammar rules to guide the LLM in generating lexical entries based on the words in the dictionary.

The primary data sources for this study include bilingual dictionaries and parallel sentences that pair Moklen with English. The dictionary in this paper developed from (Pittayaporn et al., 2022) serves as a foundational resource, containing approximately 1,000 basic vocabulary items in Moklen along with their English translations. In addition, a collection of parallel sentences, derived from field notes and existing documentation, will be used to provide contextual examples of grammatical structures in both languages.

4.1 Tokenisation

In this step, we create a dictionary-based tokeniser to segment the Moklen text into individual words. This approach is feasible since all the words in the Moklen sentences from the bitext exist in the dictionary. However, tokenising Moklen presents unique challenges due to its high frequency of compound words. To address this, a longest-match strategy is employed, where the tokeniser first looks for the closest and largest item in the dictionary before considering smaller units. This prevents over-tokenisation and ensures that the true meaning of each sentence is accurately represented.

For example, the Moklen word *maklaw* 'to speak', is a compound form derived from the root word *klaw*. If tokenised incorrectly, it could be split into two different words: *ma* 'horse' and *klaw* 'to speak'. This misinterpretation could lead to a sentence being incorrectly understood as 'The horse is speaking.' Accordingly, the tokeniser will recognise maklaw as a single unit, thereby preserving the correct meaning and enhancing the overall accuracy of the analysis.

4.2 Sense Mapping

Once the text is tokenised, we perform sense mapping by searching for the corresponding meaning of each token in the Moklen sentences. This step is crucial to ensure that the LLM does not attempt to match each token independently, which often results in inaccuracies. Instead, our aim is to provide the LLM with contextually relevant mappings.

Each word in Moklen may have multiple meanings, and it is essential for the LLM to select the appropriate meaning based on the context provided. This process assumes that the LLM can consider the various meanings of words and choose the one that best fits the context of the sentence. For example, in the sentence *tichum boh pong*, which translates to 'The bird is building the nest', the word *pong* has two different senses: 'nest' and 'father'. The LLM should correctly interpret the sentence as 'The bird is building the nest' rather than 'The bird is building/making the father', demonstrating its ability to disambiguate based on context.

4.3 Generation of Grammar

After mapping the meanings, we concatenate the meanings of each word and prepare the data for input into the LLM. The underlying idea is that even though the LLM may have no prior knowledge of Moklen, it can employ its logical reasoning capabilities to analyse the provided data. By comparing the shared structures of Moklen and English, the LLM can induce grammatical rules from the context and information given.

For model generation, there are two approaches to creating the Moklen grammar. The first approach involves describing the language using natural language, detailing information such as word order, grammatical features, and functions. The second approach is to generate the grammar in a formal language, specifically XLE, by creating rules and lexical entries using a specific template and syntax[1]. Whilst we favour the second approach as the end product of this paper, the natural language description can be a good indicator to understand how the LLM interprets and analyses the language data.

5 Experiment

5.1 Experimental Setup

We ran all experiments on OpenAI's API-based model `gpt-4o-mini-2024-07-18` and keep a rather low temperature at 0.1 across tasks. The benchmark data, code, and model generations can be found in the supplementary material.

Despite being a single task, we also perform the task under different experimental settings, varying the kind of retrieved contexts that are provided in the prompt. See Appendix A for full details. The types of context include:

No context (-): Apart from Moklen sentences, the model is told only that Moklen is a language spoken around Southern Thailand and given no reference material. This measures the base model's zero-shot capabilities and validates that they have effectively learned zero Moklen during pre-training.

Bitext Context (B): For each sentence, we provide an English translation into the prompt.

Tokenised Context (T): We experiment with two different ways. First, for each sentence, we provide tokens (T^0). Second, for each word in each Moklen sentence, we map them with the English definition no matter how many senses they are in the prompt (T^D).

Concatenated Sentence Context (C): For each sentence, we take the translation of each word to concatenate as one string, though they may be nonsensical or ungrammatical (since we need the incorrect ones on purpose!). It is assumed that this type of string will reflect how word order in Moklen works as opposed to English.

Example Context (E): The model is provided with an example of how the English language is implemented with XLE.

Self-Explanation Context (S): Prior to the generation, the model is given the relevant data to analyse and describe in natural language, and take it as a guideline when generating the grammar.

For all contexts, the model is retrieval-augmented by providing with the XLE documentation that describes how to create a grammar and a Moklen dictionary except for no context condition, storing in a separate datastore (Asai et al., 2024).

[1]Henceforth, *grammar* refers to formal rules and lexical entries in XLE.

5.2 Evaluation Measures

Evaluating the quality of generated grammars is inherently an complex task, as there is no single correct approach to grammar construction. Consequently, the evaluation must consider both quantitative and qualitative aspects of the model's performance and the output.

Translation. We follow Tanzer et al. (2024) and Zhang et al. (2024) in that the translation performance is used as an indirect indicator of the model's understanding of linguistic structures, albeit not explicitly. Thus, this task is the starting point to selecting the best combination of contexts to proceed to the more sophisticated step. We will evaluate the model's performance on Moklen-to-English translation tasks, using an additional set of 40 parallel sentences. The quality of translations will be assessed using several metrics including a traditional BLEU (Papineni et al., 2002), ROUGE (Ganesan, 2018), METEOR (Banerjee and Lavie, 2005), chrF (Popović, 2015), and BERTScore (Zhang et al., 2020). This very task will serve as an experimental ablation to determine the most effective context settings for further steps. Apart from this, English-to-Moklen translation will be conducted both before and after generating the grammar to determine the impact of the generated grammar on translation performance.

Grammar Rule Accuracy. Never before had Moklen been developed its grammar using XLE. Hence, there is no XLE grammar for Moklen; we attempt to create one based on Spencer (2022) and served as a gold standard. The model will be prompted to generate the generated grammar rules, which will be qualitatively evaluated. Ideally, the rules must demonstrate the capacity to accurately model all possible sentence structures in Moklen, including the full range of parts-of-speech categories. Specifically, the grammar should cover syntactic constructions, word order, and morphological agreement, ensuring that all the core grammatical elements of Moklen are represented.

Lexical Entry and Schemata Accuracy. The model will generate lexical entries for 100 words from the Moklen dictionary that do not appear in the bitext, primarily based on the grammar rules produced by the model and the information provided in the dictionary. For each lexical entry, there are differences between how a different element differently exhibits their functional schemata. This accuracy is also assumed to be the result of the accuracy of the grammar rules, as well as the understanding of both grammar of the language and XLE architecture. We will manually assess the accuracy of these lexical entries by comparing them with existing dictionary definitions and evaluating their coherence across generations. We will consider the completeness of the entries, examining whether the model captures various meanings and usages of each word. Ultimately, these lexical entries will be integrated into the XLE system.

6 Results

6.1 Translation

Ablation. We explored 48 possible context combinations (see Appendix B) and identified 10 particularly noteworthy ones based on representativeness, not necessarily the best performance. These combinations are: **-**, **B**, $\mathbf{T}^0$, $\mathbf{B}$+$\mathbf{T}^0$+$\mathbf{S}$, $\mathbf{B}$+$\mathbf{T}^D$, $\mathbf{B}$+$\mathbf{T}^D$+$\mathbf{C}$, $\mathbf{T}^D$+$\mathbf{C}$+$\mathbf{S}$, $\mathbf{B}$+$\mathbf{T}^D$+$\mathbf{C}$+$\mathbf{E}$, $\mathbf{T}^D$+$\mathbf{C}$+$\mathbf{E}$+$\mathbf{S}$, and $\mathbf{B}$+$\mathbf{T}^D$+$\mathbf{C}$+$\mathbf{E}$+$\mathbf{S}$. Further translation tasks were conducted on these combinations.

Surprisingly, the combination with all contexts did not yield the best performance. Instead, the $\mathbf{T}^D$ context alone achieved the highest score, with a BERTScore F1 of 0.4904, only slightly better than the second-best. This suggests that $\mathbf{T}^D$ is more effective when combined with other contexts, outperforming $\mathbf{T}^0$ in all scenarios. The impact of contexts **E** and **S** is relatively equal when combined with other contexts.

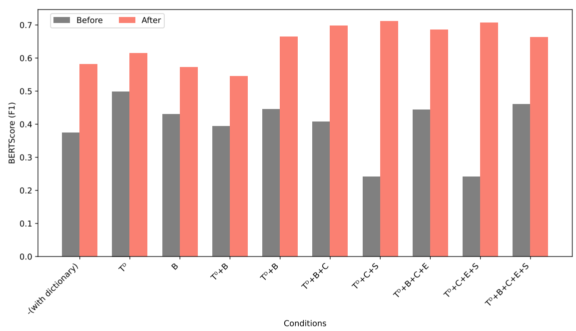

Figure 3: When instructing with grammar, the model performs better across various combinations of context

English-to-Moklen with Grammar. When incorporating the XLE gold standard grammar into the prompt across all models, we observed an overall improvement in translation performance. This suggests that the inclusion of grammar plays a crucial role in improving the quality of the model output.

This finding is consistent with our previous observations. Specifically, the context combination $\mathbf{T}^D$**+C+S** emerged as the most effective, achieving the highest score in nearly all benchmarks with an estimated BERTScore F1 of 0.7110. This result highlights that certain combinations of contexts can significantly enhance translation accuracy.

English-to-Moklen with Grammar. Interestingly, translating from Moklen to English demonstrated better performance compared to the reverse translation (English-to-Moklen). This suggests that the model's ability to generate English sentences from Moklen input is more refined or effective than generating Moklen sentences from English. The implication behind this will be discussed later in the next section.

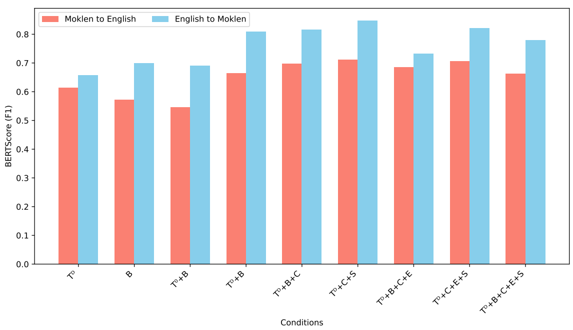

Figure 4: The improvement in scores across contexts leading to the suspicion of the reliability of the metric

6.2 Grammar Rules: Accuracy and Completeness

To assess the completeness of the grammar rules, we evaluated whether they could accurately capture all parts of speech in Moklen without overly relying on English examples.

Using the context $\mathbf{T}^D$**+C+S**, the model effectively identified common parts of speech such as nouns, verbs, adjectives, and adverbs. However, some unique English parts of speech, particularly determiners, were incorrectly transferred to Moklen grammar, despite the language not having equivalent features. This discrepancy raises questions that will be explored in the following discussion. Whilst these errors highlight the influence of English-centric training, the XLE implementation successfully captured Moklen's essential syntactic structures, including word order and topicalisation. These results suggest that LLMs can be a valuable starting point, but careful post-editing by linguists remains essential.

6.3 Lexical Entries: Accuracy and Coherence

We generated lexical entries for 100 Moklen words that are not present in the bitext and assessed their accuracy by comparing them to existing dictionary definitions. Out of these, 86 entries were deemed accurate and coherent. However, some incomplete lexical entries captured only the more prominent senses of words. For example, the word *data* was defined as a preposition meaning 'on' and a noun meaning 'the top part', but the latter sense was not captured by the model.

7 Analysis & Discussion

There are several critical aspects emerging from the experimental results that deserve further discussion.

Dictionary Integration. The inclusion of a dictionary in the prompt significantly impacts the performance of the model, even without additional contextual modifications. This simple integration yields results that are comparable to or better than the combination of the dictionary with other contexts. Specifically, the dictionary enhances the model's lexical understanding by grounding it in a structured reference, which is essential when working with languages with minimal resources. For Moklen, a language that lacks morphological inflections and instead uses content words to express grammatical relations (e.g., tense or aspect), dictionary-based lexical grounding proves effective. The performance improvements can be seen in the BERTscores, where the model achieved translation quality approaching that of more morphologically complex languages, despite Moklen's isolating structure. This supports the idea that even minimal data, when properly used, can provide significant gains in language processing tasks.

Evaluation of Translation Metrics. A key challenge with existing translation metrics such as BLEU, ROUGE-L, METEOR, and chrF is their heavy reliance on word forms and syntactic similarity, which aligns poorly with Moklen's structural characteristics. For instance, these metrics work effectively in English, where grammatical markers such as tense are explicitly coded into verb forms. However, Moklen's lack of grammatical tense markers presents a problem as a single word can express multiple temporal aspects. This reduces the validity of wordform-based metrics for this language, leading to discrepancies between numerical scores and actual translation quality. As an

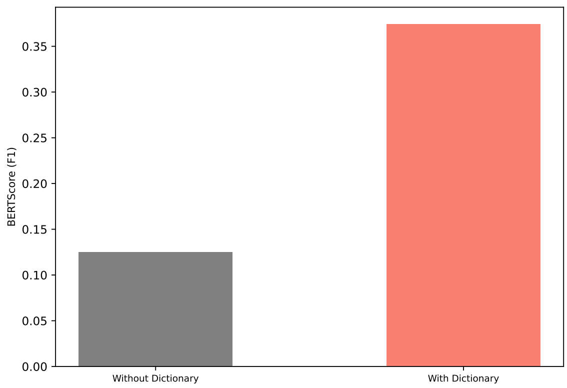

Figure 5: The clear improvement after intergrating dictionary into the prompt in Moklen-to-English translation tasks

alternative, BERTScore, which prioritizes semantic similarity over word forms, emerges as a better choice for evaluating translations of languages like Moklen. However, even BERTScore falls short in fully capturing the subtleties of Moklen's grammar. To address this, future work should explore the development of a new evaluation metric tailored to languages with similar grammatical features, focusing on meaning representation rather than form.

Model Size Considerations. One critical design decision in this study was the intentional use of a small model. The choice was made to work with a smaller model to demonstrate the potential for generating usable grammar and lexical information with limited computational resources. The success of the small model implies that larger models could yield even better performance, especially in capturing more subtle linguistic nuances or handling rare linguistic phenomena. Given the scalability of LLMs, this opens a pathway for applying similar methods to larger models for endangered languages with more complex syntactic structures or larger datasets.

Hallucinations and Model Inaccuracies. In low-resource settings, model hallucinations, particularly in zero-shot translation tasks, were evident. Hallucinations occurred primarily when the model was forced to translate or generate sentences with little or no exposure to similar data during training. This was especially pronounced with Moklen, where the use of Thai script caused confusion. Due to overlap in the lexicons of Thai and Moklen, the model occasionally produced hybridized translations, mistaking Thai words for Moklen ones. Although these inaccuracies were rare, they underscore the importance of refining training data and prompt construction to mitigate such issues. Moving forward, the inclusion of more diverse Moklen data or deliberate disambiguation in the training process may reduce hallucinations and improve model robustness.

However, while these deviations can be problematic, they can also offer unexpected insights to linguists. In some cases, the model's introduction of 'non-Moklen' parts of speech or grammar elements not traditionally associated with the language might suggest an overlooked pattern or nuance. These hallucinations, though initially appearing as errors, can prompt linguists to reconsider their assumptions and explore whether the model has captured subtle relationships or features that were not immediately apparent from a human perspective. For instance, the model's introduction of categories that do not exist formally in Moklen may highlight potential areas of linguistic overlap or underlying structures that deserve further investigation. This kind of speculative output, while not always accurate, provides a thought-provoking dimension to the analysis, encouraging a holistic reexamination of linguistic data.

Potential for XLE Parsing and Beyond. this approach should not be overlooked. XLE's highly sophisticated nature for parsing natural language makes it an ideal framework for documenting endangered languages. However, given the success of this methodology, it is reasonable to assume that the generated grammar could be adapted to other formal frameworks, such as HPSG or dependency grammars, making it widely applicable across different linguistic projects. This flexibility suggests that LLMs, when properly guided, can not only assist in XLE parsing but also contribute to the broader field of computational linguistics by providing insights into the underlying structure of under-documented languages.

8 Conclusion

It is evident that an LLM, if sufficient information is provided, can assist linguists in generating linguistic data for such complex and, moreover, tedious tasks. This methodology with careful prompting techniques offers a cost- and time-effective means of obtaining grammatical information for the endangered Moklen language by means of minimal linguistic resources and analogy from English language. By using as little as bilingual dictionaries and in-context learning, we successfully extracted coherent grammatical rules and lexical entries, thereby contributing to the documentation and preservation of Moklen or, at least, isolating languages in general.

Limitation

While this study demonstrates promising results, it is essential to acknowledge its limitations. We focused on a single language, Moklen, an isolating language with no grammatical inflections, to evaluate the LLM's capabilities under resource-scarce conditions. This typological characteristic may have facilitated the LLM's ability to decipher Moklen from a minimal amount of information compared to more complex synthetic or agglutinative languages. To enhance the robustness of our findings, we plan to extend this methodology to include various languages from different typological and morphological backgrounds, assessing the method's broader applicability. Furthermore, the potential for English grammatical biases in the model's outputs indicates a need for further research to develop techniques that minimize cross-linguistic interference in low-resource language modelling. Addressing these limitations through multilingual training and iterative feedback from linguists will be crucial for expanding the utility of this approach.

Acknowledgments

We would like to extend our heartfelt thanks to the people at the CIS at LMU Munich, especially Prof. Hinrich Schütze for inspiring this project during the exchange period of PS. We are also grateful to Yihong Liu and Haotian Ye for their valuable comments on the first draft, which significantly enhanced the quality of our work. Special appreciation goes to Prof. Pittayawat Pittayaporn, the PI of the Moklen project at Chulalongkorn University, for his ongoing support in providing both resources and insights into the language. Most importantly, we acknowledge the Moklen community for their collaboration and support, which is indeed crucial for the success of this research.

References

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. *Preprint*, arXiv:2403.03187.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Gašper Beguš, Maksymilian Dąbkowski, and Ryan Rhodes. 2023. Large linguistic models: Analyzing theoretical linguistic abilities of llms. *Preprint*, arXiv:2305.00948.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. In *COLING-02: Grammar Engineering and Evaluation*.

Miriam Butt, Tracy Holloway King, Maria-Eugenia Niño, and Frederique Segond. 1999. *A Grammar Writer's Cookbook*. Stanford University Press.

Noam Chomsky and Robert Berwick. 2016. *Why Only Us*. MIT Press.

Ann Copestake. 2002. Definitions of typed feature structures. In Stephan Oepen, Dan Flickinger, Jun-ichi Tsujii, and Hans Uszkoreit, editors, *Collaborative Language Engineering*, pages 227–230. CSLI Publications, Stanford, CA.

Richard Crouch, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. III Maxwell, and Paula S. Newman. 2011. *XLE Documentation*.

Mary Dalrymple, editor. 2023. *Handbook of Lexical Functional Grammar*. Number 13 in Empirically Oriented Theoretical Morphology and Syntax. Language Science Press, Berlin.

Ethnologue. 2024. How many languages are endangered?

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *Preprint*, arXiv:1803.01937.

John T. Maxwell and Ronald M. Kaplan. 1993. The interface between phrasal and functional constraints. *Computational Linguistics*, 19(4):571–590.

Diego Miranda-Saavedra. 2024. Generative ai models and the quest for human-level artificial intelligence. *Real World Data Science*.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO redwoods treebank: Motivation and preliminary applications. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pittayawat Pittayaporn and Songphan Choemprayong. 2024. A proposal for a thai-based moklen orthography. *Language Documentation and Conservation*.

Pittayawat Pittayaporn, Warunsiri Pornpottanamas, and Daniel Loss. 2022. *Moklen-Thai-English Dictionary: A Pilot Version*. Chulalongkorn University Press.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Bradley L. Pritchett. 1992. *Grammatical Competence and Parsing Performance*. The University of Chicago Press.

Jane Simpson. 1991. *Warlpiri Morpho-Syntax: A Lexicalist Approach*. Kluwer Academic Publisher.

Piyapath Spencer. 2024. Documenting endangered languages with LangDoc: A wordlist-based system and a case study on Moklen. In *Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024)*, pages 28–36, Bangkok, Thailand. Association for Computational Linguistics.

Piyapath T Spencer. 2022. Grammatical sketch of moklen verb phrase. *Unpublished Manuscipt*.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations*.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

W3Techs. 2024. Usage statistics of content languages for websites.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Olga Zamaraeva, Chris Curtis, Guy Emerson, Antske Fokkens, Michael Goodman, Kristen Howell, T.J. Trimble, and Emily M. Bender. 2022. 20 years of the grammar matrix: cross-linguistic hypothesis testing of increasingly complex interactions. *Journal of Language Modelling*, 10(1):49–137.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15654–15669, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Prompts

A.1 System Prompt

> You are a research assistant in linguistics and you must complete every assigned task.

A.2 No context (-)

> You are given some examples of Moklen sentences `{moklen_only}`. Moklen is a language spoken around the Southern Thailand.
>
> **data structure:**
>
> $$[\ \text{sentence}_1,\ \text{sentence}_2,\ \vdots\ \text{sentence}_n\ ]$$

A.3 Bitext Context (B)

Here are some examples of Moklen sentences and their corresponding English translations `{Bitext}`

data structure:

$$\begin{array}{l} [\ [\ \text{Moklen}_1, \text{English}_1\ ] \\ \quad [\ \text{Moklen}_2, \text{English}_2\ ] \\ \qquad \vdots \\ \quad [\ \text{Moklen}_n, \text{English}_n\ ]\ ] \end{array}$$

A.4 Tokenised Context (T)

(1) T^0

Here are examples of Moklen sentences that are tokenised `{moklen_only_tokenised}`.

data structure:

$$\begin{array}{l} [\ [\ \text{word}_1^1, \text{word}_2^1, \ldots, \text{word}_n^1\ ]_1 \\ \quad [\ \text{word}_1^2, \text{word}_2^2, \ldots, \text{word}_n^2\ ]_2 \\ \qquad \vdots \\ \quad [\ \text{word}_1^n, \text{word}_2^n, \ldots, \text{word}_n^n\ ]_n\ ] \end{array}$$

(2) T^D

Here are some examples of Moklen sentences that are tokenised and each token is mapped with its corresponding POS and translations `{bitext_mapped}`.

data structure:

$$\begin{array}{l} [\ \{\ \text{word}_1^1 : (\text{meaning}_1^1, \text{pos}_1^1), \\ \qquad \text{word}_2^1 : (\text{meaning}_2^1, \text{pos}_2^1), \\ \qquad \vdots \\ \qquad \text{word}_n^1 : (\text{meaning}_n^1, \text{pos}_n^1)\ \}_1, \\ \quad \{\ \text{word}_1^2 : (\text{meaning}_1^2, \text{pos}_1^2), \\ \qquad \text{word}_2^2 : (\text{meaning}_2^2, \text{pos}_2^2), \\ \qquad \vdots \\ \qquad \text{word}_n^2 : (\text{meaning}_n^2, \text{pos}_n^2)\ \}_2, \\ \qquad \vdots \\ \quad \{\ \text{word}_1^n : (\text{meaning}_1^n, \text{pos}_1^n), \\ \qquad \text{word}_2^n : (\text{meaning}_2^n, \text{pos}_2^n), \\ \qquad \vdots \\ \qquad \text{word}_n^n : (\text{meaning}_n^n, \text{pos}_n^n)\ \}_n\ ] \end{array}$$

A.5 Example Context (E)

Here is the example of how English is documented using XLE:

```
DEMO   ENGLISH   RULES (1.0)

   S --> e: (^ TENSE);
      (NP: (^ XCOMP* {OBJ|OBJ2})=!
           (^ TOPIC)=!)
        NP: (^ SUBJ)=!
          (! CASE)=NOM;
        { VP |VPaux}.

   VP --> V (NP: (^ OBJ)=!
             (! CASE)=ACC)
  PP*:! $   (^ ADJUNCT).

   VPaux --> AUX VP.

----

DEMO ENGLISH LEXICON (1.0)

is     AUX * (^ TENSE)=PRES
    (^ ASPECT)=PROG.

in     P *   @(PREP IN).

with   P * @(PREP WITH).

the     D * @(DET DEF).

a    D * @(DET INDEF)
      @(NUMBER SG).

girl    N * @(N-SG GIRL).

boys    N * @(N-PL BOY).

----
```

A.6 Self-Explanation Context (S)

Here is the description of Moklen language `{grammar_description}` you analysed earlier.

```
Moklen is a language that exhibits a unique grammatical structure distinct from
English. Below, I will break down the grammar of Moklen, focusing on morphology,
syntax, and parts of speech, while also comparing it to English where relevant.

Morphology
1. **Word Formation**: - Moklen words are generally monomorphemic, meaning they
consist of a single morpheme without inflectional endings. For example, the
word “tj” (I, me) is a standalone pronoun without any grammatical inflection.
- Moklen does not use prefixes or suffixes to indicate tense, number, or case,
unlike English, which uses inflections (e.g., “walk” vs. “walked” for past
tense).
2. **Compounding**: - Moklen utilizes compound words to convey complex meanings.
For instance, “kan.k.poŋ.pù tíak” (asian sea bass) is a compound noun made up
of “kan” (fish) and “k.poŋ.pù.tíak” (specific type of fish). - This contrasts
with English, which often uses separate words or phrases to describe similar
concepts.
3. **Part of Speech**: - Each word in Moklen typically belongs to a specific
part of speech without changing form. For instance, “didun” (to lie, sleep) is
consistently a verb regardless of context.
...
```

B Ablation

Conditions	BLEU	ROUGE-L	METEOR	chrF	BERT (P)	BERT (R)	BERT (F1)
-	0.0256	0.1100	0.0882	11.3078	0.1270	0.1214	0.1250
- (with dictionary)	0.0875	0.4803	0.3158	11.6155	0.4242	0.3239	0.3740
T^{o}	0.1086	0.4793	0.3293	18.4021	0.4813	0.3931	0.4371
T^{D}	0.2683	0.5636	0.4840	17.3389	0.4978	0.5001	0.4989
B	0.1747	0.5198	0.5300	20.8929	0.3983	0.4615	0.4297
C	0.0673	0.5341	0.3574	23.2696	0.1415	0.2666	0.2026
E	0.0958	0.5062	0.3105	17.1944	0.4515	0.3320	0.3914
S	0.1496	0.5566	0.4464	18.6564	0.4605	0.4341	0.4474
T^{o}+B	0.1685	0.5144	0.5026	19.8641	0.3664	0.4227	0.3944
T^{o}+C	0.0673	0.5341	0.3574	23.2696	0.1415	0.2666	0.2026
T^{o}+E	0.0997	0.4836	0.3402	16.9057	0.4499	0.3590	0.4043
T^{o}+S	0.1518	0.5341	0.4376	19.8641	0.3886	0.3956	0.3922
T^{D}+B	0.1920	0.5433	0.5251	19.8641	0.4242	0.4653	0.4447
T^{D}+C	0.0673	0.5341	0.3574	23.2696	0.1415	0.2666	0.2026
T^{D}+E	0.2434	0.6184	0.5207	18.4021	0.4707	0.4920	0.4811
T^{D}+S	0.2260	0.6025	0.5231	13.3516	0.4992	0.4629	0.4810
B+C	0.1693	0.5218	0.5214	19.8641	0.3319	0.4395	0.3853
B+E	0.1737	0.5714	0.5338	19.8641	0.4279	0.4734	0.4504
B+S	0.1932	0.5751	0.5518	19.8641	0.4163	0.4609	0.4384
C+E	0.0673	0.5341	0.3574	23.2696	0.1415	0.2666	0.2026
C+S	0.0673	0.5341	0.3574	23.2696	0.1415	0.2666	0.2026
E+S	0.1336	0.5857	0.3638	20.1184	0.5168	0.4179	0.4672
T^{o}+B+C	0.1857	0.5612	0.5395	19.8641	0.4004	0.4589	0.4295
T^{o}+B+E	0.1754	0.5495	0.5311	19.8641	0.3946	0.4475	0.4209
T^{o}+B+S	0.1983	0.5719	0.5432	19.8641	0.4246	0.4592	0.4418
T^{o}+C+E	0.0673	0.5341	0.3574	23.2696	0.1415	0.2666	0.2026
T^{o}+C+S	0.0673	0.5341	0.3574	23.2696	0.1415	0.2666	0.2026
T^{o}+E+S	0.2151	0.6062	0.5183	19.8641	0.4253	0.4472	0.4362
T^{D}+B+C	0.1894	0.5300	0.5120	19.8641	0.3666	0.4491	0.4076
T^{D}+B+E	0.1861	0.5555	0.5501	19.8641	0.3913	0.4618	0.4263
T^{D}+B+S	0.1920	0.5587	0.5428	19.8641	0.4286	0.4876	0.4578
T^{D}+C+E	0.0673	0.5341	0.3574	23.2696	0.1415	0.2666	0.2026
T^{D}+C+S	0.0870	0.5639	0.3613	23.2696	0.2040	0.2809	0.2408
T^{D}+E+S	0.2490	0.5918	0.4984	17.3389	0.4926	0.4887	0.4904
B+C+E	0.1708	0.5329	0.5378	19.8641	0.3381	0.4314	0.3843
B+C+S	0.1890	0.5379	0.5311	19.8641	0.3820	0.4574	0.4195
B+E+S	0.1851	0.5252	0.5093	19.8641	0.3620	0.4124	0.3871
C+E+S	0.0673	0.5341	0.3574	23.2696	0.1415	0.2666	0.2026
T^{o}+B+C+E	0.1882	0.5519	0.5400	18.9681	0.4005	0.4553	0.4279
T^{o}+B+C+S	0.1901	0.5552	0.5329	19.8641	0.3882	0.4467	0.4173
T^{o}+B+E+S	0.1939	0.5677	0.5440	19.8641	0.4281	0.4789	0.4534
T^{o}+C+E+S	0.0673	0.5341	0.3574	23.2696	0.1415	0.2666	0.2026
T^{D}+B+C+E	0.2082	0.5982	0.5796	19.8641	0.4080	0.4793	0.4435
T^{D}+B+C+S	0.2017	0.5919	0.5934	19.8641	0.4309	0.4928	0.4617
T^{D}+B+E+S	0.1899	0.5722	0.5341	19.8641	0.3916	0.4627	0.4270
T^{D}+C+E+S	0.0870	0.5639	0.3613	23.2696	0.2040	0.2809	0.2408
B+C+E+S	0.1946	0.5763	0.5568	19.8641	0.4165	0.4738	0.4448
T^{o}+B+C+E+S	0.1932	0.5658	0.5392	19.8641	0.3735	0.4432	0.4082
T^{D}+B+C+E+S	0.1908	0.5969	0.5615	19.8641	0.4247	0.4964	0.4602

Figure 6: Moklen-to-English translation before including grammar

Conditions	BLEU	ROUGE-L	METEOR	chrF	BERT (P)	BERT (R)	BERT (F1)
-	0.0336	0.1293	0.0786	9.1679	0.2628	0.2185	0.2416
- (with dictionary)	0.1802	0.5387	0.5149	19.9009	0.5536	0.6102	0.5818
T^{o}	0.1936	0.6116	0.5879	19.8641	0.5774	0.6308	0.6038
T^{D}	0.2065	0.6282	0.6275	20.8929	0.5744	0.6559	0.6143
B	0.2256	0.5850	0.5850	19.9969	0.5335	0.6104	0.5717
C	0.1519	0.5507	0.5146	18.8720	0.5900	0.6097	0.5997
E	0.1848	0.5678	0.5614	18.8720	0.5346	0.6003	0.5670
S	0.2077	0.6146	0.5891	19.8641	0.6067	0.6512	0.6288
T^{o}+B	0.1771	0.5883	0.5768	19.9969	0.5042	0.5873	0.5455
T^{o}+C	0.1993	0.5893	0.5739	18.9681	0.5509	0.6032	0.5769
T^{o}+E	0.1824	0.5832	0.5824	20.8929	0.5893	0.6394	0.6140
T^{o}+S	0.1851	0.5408	0.5058	19.9969	0.5142	0.5855	0.5498
T^{D}+B	0.2225	0.6682	0.6394	20.8929	0.6489	0.6800	0.6645
T^{D}+C	0.1964	0.6172	0.5732	19.8641	0.5832	0.6142	0.5985
T^{D}+E	0.1923	0.5828	0.5529	20.8929	0.5324	0.5995	0.5657
T^{D}+S	0.2281	0.6391	0.6235	20.8929	0.5944	0.6491	0.6215
B+C	0.1949	0.5940	0.5365	16.3468	0.5699	0.5798	0.5744
B+E	0.1919	0.5569	0.5007	16.3468	0.5636	0.5740	0.5684
B+S	0.1965	0.5838	0.5587	19.8641	0.5581	0.6216	0.5895
C+E	0.1865	0.6042	0.5250	19.2275	0.5404	0.5344	0.5365
C+S	0.2143	0.6241	0.6071	19.8641	0.6161	0.6495	0.6325
E+S	0.2115	0.6163	0.5975	19.8641	0.5995	0.6412	0.6201
T^{o}+B+C	0.1993	0.5861	0.5873	19.9969	0.5547	0.6112	0.5828
T^{o}+B+E	0.1918	0.5781	0.5752	20.8929	0.5521	0.6012	0.5765
T^{o}+B+S	0.2048	0.5971	0.5802	18.8720	0.6047	0.6395	0.6218
T^{o}+C+E	0.2169	0.6119	0.6013	19.9969	0.5678	0.6324	0.5999
T^{o}+C+S	0.2015	0.6059	0.5973	19.8641	0.5573	0.6058	0.5811
T^{o}+E+S	0.1902	0.5679	0.5539	19.8641	0.5252	0.5761	0.5506
T^{D}+B+C	0.2649	0.6916	0.6663	22.1555	0.6822	0.7137	0.6979
T^{D}+B+E	0.2353	0.6479	0.6342	20.8929	0.6537	0.6893	0.6712
T^{D}+B+S	0.2011	0.6310	0.5739	19.8641	0.6021	0.6234	0.6128
T^{D}+C+E	0.2605	0.6453	0.6352	22.1555	0.5786	0.6699	0.6239
T^{D}+C+S	0.2735	0.7002	0.6524	20.8929	0.7074	0.7147	0.7110
T^{D}+E+S	0.2352	0.6400	0.6340	20.8929	0.6476	0.6897	0.6684
B+C+E	0.1754	0.5720	0.5657	20.8929	0.5379	0.6061	0.5717
B+C+S	0.2471	0.6293	0.6147	19.8641	0.6021	0.6570	0.6293
B+E+S	0.2069	0.5716	0.5529	19.9969	0.5401	0.6016	0.5707
C+E+S	0.1997	0.5793	0.5641	18.9681	0.5527	0.6030	0.5777
T^{o}+B+C+E	0.2022	0.5988	0.5867	19.9969	0.5714	0.6281	0.5996
T^{o}+B+C+S	0.2508	0.6323	0.6249	19.8641	0.6062	0.6586	0.6321
T^{o}+B+E+S	0.1985	0.5979	0.5819	20.8929	0.5509	0.6168	0.5837
T^{o}+C+E+S	0.1990	0.5891	0.5694	18.9681	0.5599	0.6046	0.5821
T^{D}+B+C+E	0.2357	0.7060	0.6601	20.8929	0.6708	0.7007	0.6857
T^{D}+B+C+S	0.1753	0.5844	0.5569	19.8641	0.5023	0.5897	0.5455
T^{D}+B+E+S	0.2093	0.6420	0.6097	22.1555	0.5365	0.6325	0.5836
T^{D}+C+E+S	0.2476	0.6677	0.6620	20.8929	0.6955	0.7181	0.7069
B+C+E+S	0.1645	0.5946	0.5552	19.8641	0.5565	0.5951	0.5757
T^{o}+B+C+E+S	0.2150	0.6337	0.6163	19.8641	0.5886	0.6319	0.6100
T^{D}+B+C+E+S	0.2278	0.6617	0.6393	19.8641	0.6372	0.6888	0.6627

Figure 7: Moklen-to-English translation after including grammar