Creating Grammar Teaching Material for Endangered Languages with Hybrid Grammar Induction

Sebastien Christian
Research Center for Pacific Societies and Humanities
University of French Polynesia, CNRS
sebastien.christian@upf.pf

Abstract

Explicit grammar teaching plays a central role in endangered language revitalization, yet creating grammar lessons is labor-intensive and is typically assigned to already overwhelmed teachers. We introduce HYGRAM, a Hybrid Grammar Induction method for grammar induction from sparse data that integrates expert linguistic abstractions, typological priors, Bayesian inference, and constrained LLM reasoning. This method is deployed in a teacher-facing software platform designed for languages with minimal digital footprint. Given as little as a corpus elicited within approximately 10 hours of fieldwork and any available descriptive documents, HYGRAM produces topic-specific, structured grammar lessons designed for classroom use. We evaluate the system on six typologically diverse endangered languages with expert linguists, who rate outputs for similarity to expert-authored materials, linguistic quality, and pedagogical usefulness. Results show consistently high similarity, reliable quality once a modest data threshold is reached, and strong recommendations for teacher use (mean 7.7/9). Community feedback from Vanuatu-based language practitioners further indicates high perceived relevance for local revitalization efforts. Our findings demonstrate that controlled, hybrid grammar induction can support practical grammar teaching in extremely low-resource settings.

1 Introduction

For more than 1,700 languages worldwide classified as endangered or worse[1], revitalization efforts rely on explicit language instruction for both children and adult learners. Explicit grammar teaching plays an important role in efficient additional-language learning (Kachinske and DeKeyser, 2019; Nabizadeh et al., 2016).

For endangered languages, producing grammar lessons often falls to school teachers because standardized pedagogical materials are rarely available. Even when academic descriptions and corpora exist, they are generally unsuitable for classroom use; transforming academic literature into a didactic grammar is a substantial undertaking (Sapién and Hirata-Edds, 2019) and is therefore seldom attempted. Many teachers, especially those without formal training in grammar, report feeling underprepared to design explicit instructional materials while remaining open to computational support (Chaudhary et al., 2022).

Automating the generation of grammar teaching materials is therefore an important component of language revitalization. Yet existing methods do not meet the requirements imposed by the scarcity and heterogeneity of available data, the limited number of trained linguists, and teachers' need for practical, time-saving tools.

Our contributions are as follows:

1. We formalize the task of grammar lesson generation for the vast majority of endangered languages, including extremely low-resource languages, for which only minimal, rapidly collectable data are available.

2. We present a hybrid method that combines typological knowledge, statistical inference, and LLM-based reasoning to generate grammatical descriptions from a corpus collected in approximately 10 hours and from arbitrary descriptive documents.

3. Through expert evaluation across six endangered languages, we show that the resulting outputs are structurally consistent, pedagogically useful, and reliable once a minimal data threshold is reached.

DIG4EL, an online software platform implementing HYGRAM, has been developed in col-

[1]Endangered according to Campbell et al. (2017).

Findings of the Association for Computational Linguistics: ACL 2026, pages 26628–26643

laboration with language experts and communities across the Pacific region (Christian, 2026b). It follows the CARE (Carroll et al., 2020) and compatible FAIR (Wilkinson et al., 2016) principles for data governance and authorship.

2 Background and related work

2.1 A Note on Low-Resource Languages

There is a substantial mismatch between what some computational linguistics papers label as *low-resource* languages and the actual data conditions under which linguists work on endangered languages. In large-scale NLP projects such as Meta's No Language Left Behind (NLLB Team et al., 2024), a language with fewer than one million aligned sentence pairs is described as *low-resource*, and one with fewer than 100,000 aligned pairs as *very low-resource*. For most field linguists working on endangered languages, such thresholds are unimaginable: a corpus of 1,000 aligned sentence pairs is already considered substantial, and a corpus of 500 pairs with high-quality annotations is often sufficient to write a grammar.

The present work targets endangered languages that lack the digital footprint, audio resources, textual corpora, or structured annotations required to train even modest neural models. We refer to these here as *Extremely Low-Resource Languages* (ELRLs). ELRLs represent the majority of the world's linguistic diversity, as 95% of the world's languages are spoken by only 5% of the world's population (Nettle and Romaine, 2000).

2.2 Corpus Creation Bottlenecks

Producing grammatical descriptions presupposes the existence of a corpus of transcribed, partially annotated data. Two bottlenecks constrain this process.

First, the *transcription bottleneck*: one hour of audio may require from 5 to 60 hours of manual transcription (Foley et al., 2018; Adams et al., 2021), and current ASR tools remain either not efficient enough or difficult for non-technical fieldworkers to use (Coto-Solano et al., 2022).

Second, the *glossing bottleneck*: producing Interlinear Glossed Text (IGT) requires specialist linguistic expertise, and documentation projects regularly accumulate large backlogs of unannotated text (Seifart et al., 2018).

These constraints limit the data available for computational analysis and motivate methods able to operate with minimal, rapidly collected corpora.

2.3 The Automation of Grammatical Descriptions

Traditional field linguists' grammars require years of work because they aim to describe a language comprehensively. After publication, they are often inaccessible to teachers and communities because of their technical metalanguage, and they are rarely written in the local language of instruction.

Formal grammars derived from IGT or other sentence-level representations encode a significant portion of a language's structure and are well suited to computational analysis. They support advances in structural modeling and inference, as illustrated by BASIL (Howell and Bender, 2022) and the Autogramm project (Corro and Kahane, 2024). These approaches, however, are designed for linguistic research rather than for pedagogical or documentation use.

Industry-led efforts have taken a more pragmatic stance. Systems such as PAWS (Black and Black, 2009) aim to accelerate the creation of minimal grammars for endangered languages in order to facilitate language learning for outsiders. While useful for specific tasks, they do not generate didactic grammar lessons for community learners.

Community-driven revitalization programs, by contrast, place language teaching at their core (Shah and Brenzinger, 2018), with notable successes such as Hawaiian-medium immersion schools (Brenzinger and Heinrich, 2013). However, community-written full grammars, when they appear at all, are typically produced only after a dictionary has been completed, a process that itself may take many years.

Despite these efforts, a structural gap remains between (i) research-oriented formalisms designed for computational or theoretical analysis, (ii) industry-driven tools optimized for translation and rapid minimal learning by outsiders, and (iii) the pedagogical needs of communities engaged in revitalization. None of these approaches is designed to produce accessible, topic-structured grammar lessons that local teachers can adapt and use in classrooms with minimal training.

3 Task Definition

We define the task as the automatic generation of structured, pedagogical grammar lessons for endangered languages from available data and minimal

additional data collection.

HYGRAM must satisfy three practical requirements: (i) operate with data that a fieldworker or teacher can collect in under ten hours; (ii) produce outputs readable by non-specialist educators; (iii) restrict itself to observable evidence.

3.1 Inputs and Outputs

HYGRAM operates on transcribed and structured linguistic material rather than directly on speech: speech remains central to documentation, but in the target settings, transcription is itself the main bottleneck, and reliable automatic speech recognition is either not yet available or not yet usable in practice by language documenters.

The system may receive any combination of:

- A minimal elicited corpus of 215 pre-defined sentence pairs;
- Additional free sentence pairs;
- Publicly available PDF documents (e.g., descriptions, grammar sketches).

The output is a topic-specific grammar lesson organized into an introduction, focused explanatory sections, and examples drawn from the data, which can also be used as practice items. Topics may be selected from a predefined list (e.g., Negation, Possession, Demonstratives) or provided as free-form queries.

4 Method

4.1 Overview

HYGRAM separates the problem into two parts: inferring what can be known about a language from sparse evidence, and expressing this knowledge as a pedagogically usable lesson.

The method uses information from (i) the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) and Grambank (Skirgård et al., 2023), together with any subset of (ii) a core corpus of 215 pre-defined sentence pairs, (iii) any number of additional sentence pairs, (iv) any collection of descriptive documents, and (v) an optional topic-specific template. These inputs are enriched and combined before being passed to an LLM, which aggregates them into a structured document. All corpora used in this work were collected and made available with appropriate informed consent from speakers and communities, in accordance with the documentation practices of their respective sources.

4.2 Bayesian Parameter Inference with Frontier-Based Parameter Discovery

Our method extends the Bayesian framework introduced in Christian (2025). It leverages direct and probabilistic knowledge encoded in large-scale grammatical databases to infer the values of grammatical parameters for a target language, using, by default, a core corpus of 215 annotated sentence pairs derived from a core set of Conversational Questionnaires (CQs) proposed by François (2019). The number 215 is not a theoretically determined minimal threshold; rather, it corresponds to the number of segments derived from the five CQs presented by François (2019). Using the associated software tools, this corpus can be collected and annotated in under ten hours. Annotation does not require linguistic training: speakers simply indicate which word(s) in a sentence contribute to each expected semantic concept.

The original Bayesian formulation required preselecting which parameters to infer. Since inferable parameters vary across languages and are unknown a priori, we introduce a **frontier-based discovery** mechanism. The system performs Bayesian updates on all parameters for which priors can be derived from WALS and Grambank, and incorporates automated observations extracted from the CQ corpus when available. Once beliefs have been updated, it selects an initial seed of parameters whose posterior probabilities exceed a confidence threshold. From this seed, a frontier of high-confidence inferences expands through the conditional-probability network derived from WALS and Grambank, yielding a data-driven and robust set of parameters for the target language (computation details are provided in Appendix A).

4.3 Additional Sentence Pair Augmentation and Indexing

Augmentation. A set of predefined grammatical descriptors (e.g., tense, aspect, polarity, predicate type, pronouns) is assigned to each sentence pair using an LLM[2] with constrained output (prompt in Appendix C). The model returns a structured representation containing:

- descriptor values, organized both as a vector and as a plain-language description of the sentence;

[2] o4-mini-2025-04-16, from the April 2025 "O4" family, accessed via the OpenAI API.

- key translation concepts enabling cross-linguistic alignment.

Once this initial augmentation has been performed, the interface allows users to link the key translation concepts in each sentence to word(s) in the endangered language, yielding a pseudo-gloss of the sentence. This step is optional and necessarily partial, but it provides valuable additional information.

The plain-language descriptions of the augmented sentences are encoded using the `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019) and indexed with FAISS (Johnson et al., 2019).

This retrieval layer ensures that vectorization reflects the sentence's grammatical description, since user queries target grammatical phenomena.

4.4 Documents Vectorization and Indexing

Users may also upload publicly available PDF documents, such as theses, grammar sketches, or community manuals. HYGRAM delegates document embedding, indexing, and semantic retrieval to OpenAI's vector-store API.[3]

4.5 Aggregation

User inputs. Once a language has been selected, users are invited to specify a grammar topic, either by choosing from a list of predefined topics or by entering a free-form query. Each predefined topic is associated with a template designed by linguists familiar with typologically diverse languages; these templates propose both a discovery process and an output structure to guide generation. Users also choose the output language, which must be a mainstream language, and the intended audience, ranging from teenagers to linguists.

Aggregation Process A user query triggers the following processes:

- **Template**: If a generation template exists for the query, it is retrieved and appended to the user input (see an example in Appendix C.6).
- **Parameters**: From the full set of inferred grammatical parameters, the subset most relevant to the query is selected by an LLM[4] (see the prompt in Appendix C.3).

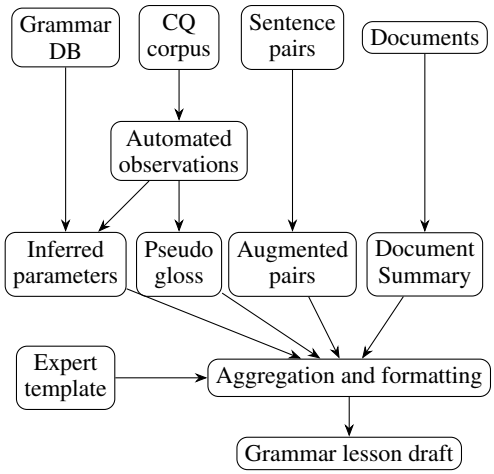

Figure 1: HYGRAM pipeline overview.

- **Sentence Pairs**: The user query is encoded using the same model as the sentence-pair representations, and the top N nearest neighbors (up to 50 by default) are retrieved.
- **Documents**: Given a query, the system extracts relevant information from the documents using a constrained LLM prompt via OpenAI's vector-store API[5] (see the prompt in Appendix C.4).

Finally, the selected parameters and their beliefs, CQ pseudo-glosses, the retrieved template, augmented sentence pairs, and document summaries are passed to a final *aggregation* LLM[6], which synthesizes the grammatical description as a structured output while being constrained to preserve source grounding and to surface contradictions between sources explicitly rather than silently resolving them (see the prompt in Appendix C.5).

A synthetic diagram is presented in Figure 1, and a detailed flowchart is presented in Appendix B.

5 Evaluation

Evaluating grammar lesson generation for endangered languages is challenging: languages differ widely in documentation, linguistic complexity, and available corpora, and no established benchmark exists for this task.

We evaluate the system on real endangered languages rather than on synthetic or mainstream languages. Synthetic languages fail to capture the uneven and heterogeneous distribution of evidence observed in real-world documentation settings, while

[3] https://platform.openai.com/docs/overview

[4] o4-mini-2025-04-16, from the April 2025 "O4" family, accessed via the OpenAI API.

[5] https://platform.openai.com/docs/overview

[6] gpt-5-2025-08-07, from the "O5" family, accessed via the OpenAI API.

evaluation on mainstream languages may introduce uncontrolled knowledge leakage from LLM pretraining.

We evaluate HYGRAM along two complementary dimensions: (i) expert judgments of lesson quality and pedagogical adequacy across six endangered languages, and (ii) feedback from speaker communities involved in revitalization activities.

Given the small number of languages, the results are interpreted descriptively rather than as statistically significant trends.

As a qualitative point of comparison, we also tested direct prompting of a general-purpose LLM without HYGRAM's structured aggregation pipeline. These outputs were poor, with hallucinated information and cross-language contamination, and were therefore not analyzed further.

5.1 Expert Evaluation Setup

We collaborated with six expert linguists to evaluate outputs for six typologically diverse languages: Anne-Laure Dotte for Iaai (New Caledonia, Dotte 2017), Alexandre François for Mwotlap (Vanuatu, François 2023), Nick Thieberger for Nafsan (Vanuatu, Thieberger 2026), Tatiana Korol for Ngen (Ivory Coast, Korol 2022), Eline Visser for Uruangnirin (Papua New Guinea, Visser 2026), and Jacques Vernaudon for Tahitian (French Polynesia, Vernaudon 2018).

For comparability, we generated *three* lessons per language under identical conditions: for an English-speaking teenage audience and on the topic of *Negation*. Negation was chosen because it is typologically rich yet sufficiently circumscribed to support a standalone lesson. Three generations were selected as a compromise between capturing controlled generation variability and limiting evaluator burden.

All data, system inputs, and generated outputs used in this evaluation are publicly available (Christian, 2026a). Appendix C.7 shows an excerpt from a system output, including the introduction and the first part of a focused section.

Experts were shown the unedited lessons and asked to rate nine statements on a 1–9 scale[7]: (1) The content of the three lessons is similar; (2) The lessons adequately cover the topic for use in teaching the language to beginners; (3) There are errors in the lessons; (4) The lessons can be understood by a school teacher; (5) These lessons could be passed directly to a teacher who speaks the language; (6) These lessons could be passed directly to a teacher who is not a fluent speaker of the language; (7) You would recommend using HYGRAM to support teachers who speak the language in creating grammar lessons; (8) This tool helps create beginner's grammar lessons faster than the way they are created today; and (9) You would recommend using HYGRAM to assist linguists in creating grammar sketches.

This evaluation setup reflects the constraints of the target domain. For ELRLs, no gold-standard benchmark of pedagogical grammar lessons exists, and correctness cannot be reliably assessed using automatic comparison metrics. Evaluation therefore relies on expert judgment and is interpreted descriptively.

Table 2 reports feedback from the six linguists. To ensure that higher scores consistently correspond to more positive evaluations, we transformed the *error* score g into $(9 - g)$ and relabeled the category as "No errors."

We grouped the expert evaluation criteria into three dimensions: **Similarity**, capturing how similar the three generated lessons for a language are to one another; **Quality**, encompassing coverage and the absence of errors; and **Recommendation**, covering all items related to perceived usefulness for teachers, learners, and linguists.

We define Quality as the mean of Coverage and No Errors scores.

5.2 Evaluation of Similarity.

Similarity scores cluster tightly around 7 across all languages. This suggests that the controlled-generation constraints applied to the LLM yield lessons that consistently follow the expected format and style while allowing acceptable variations.

5.3 Evaluation of Quality.

To visualize the correlation between available data and Quality, we applied min–max normalization to the three input types (CQs, sentence pairs, and document volume) across languages. We summed them to produce a *combined data index* ranging from 0 to 3. Figure 2 plots Quality as a function of this index. Despite the heterogeneity of the datasets, a clear trend emerges: the 3 most documented languages plateau between 6.5 and 7.0,

[7]The exact instruction to experts was: You have been provided with 3 lessons on Negation generated by DIG4EL for English-speaking teenagers learning the language, with an identical prompt. As part of the software evaluation, grade the following parameters.

Language	CQs	Sentence Pairs	Documents (MB)
Iaai	0	238	18
Mwotlap	3	4800	53
Nafsan	1	0	52
Ngen	0	378	4
Tahitian	5	3000	3
Uruangnirin	1	248	1

Table 1: Input data provided to HYGRAM for each evaluated language.

Table 2: Evaluation of outputs by language experts.

Item	Iaai	Mwotlap	Nafsan	Ngen	Tahitian	Uruangnirin
Similarity	6	7	7	8	6	7
Coverage	3	7	7	2	7	6
No Errors	2	6	7	4	6	2
Clarity	7	7	7	6	7	9
Direct to fluent	2	8	6	3	6	7
Direct to non-fluent	3	7	6	5	3	3
Recommended to teachers	7	8	8	6	8	9
Improves speed	7	7	8	6	8	5
Recommended to linguists	7	8	8	7	9	5

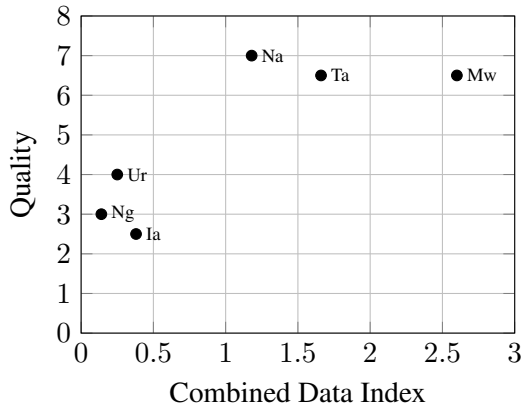

Figure 2: Quality as a function of the combined normalized data index.

whereas the three least documented languages cluster in the 2.5 to 4.0 range.

5.3.1 Data Type Influence Analysis

To isolate the contribution of different data types, we analyzed the trend between Quality and the volume of each data type, and performed an ablation test on Tahitian.

Influence of the data type on Quality Figures 3, 4, and 5 plot Quality as a function of each data type considered independently. All three sources show a broadly positive association with Quality: languages with more CQs, more sentence pairs, or more document volume tend to achieve higher scores.

For CQs (Figure 3), even a small number of questionnaires can support high Quality when complemented by rich sentence-pair or document evidence (e.g., Nafsan).

Sentence pairs (Figure 4) show the clearest monotonic trend: Mwotlap and Tahitian, which provide thousands of pairs, achieve stable Quality around 6.5–7.0. Yet Nafsan attains the highest Quality despite having *no* sentence pairs, indicating that sentence pairs alone are not the determining factor.

Documents (Figure 5) exert a strong but nonlinear influence. Large document collections (Mwotlap, Nafsan) correlate with high Quality. Still, document volume is neither necessary nor sufficient: Tahitian achieves high Quality with only 3 MB of documentation, whereas Iaai exhibits low Quality despite 18 MB of material. An investigation in documents shows that the Tahitian documents include a discussion of negation, whereas the Iaai documents do not.

Overall, these patterns indicate that **Quality depends on a balanced combination of heterogeneous evidence rather than any single data type**. CQs offer high-precision cues for key grammatical distinctions; sentence pairs provide diverse exemplars for RAG; and documents contribute global

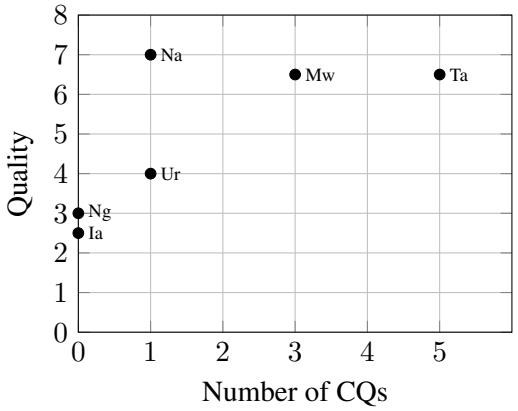

Figure 3: Quality as a function of the number of CQ.

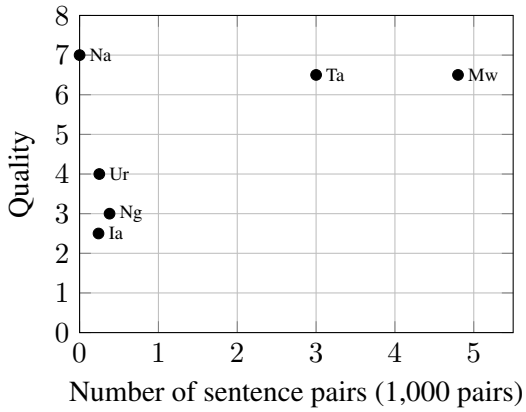

Figure 4: Quality as a function of the number of sentence pairs.

descriptive knowledge.

HYGRAM appears to perform reliably once sufficient explanatory evidence is available, whether from a single source or across multiple modalities. In contrast, reliance on a weak or poorly aligned source leads to inconsistent outcomes.

Ablation study on Tahitian (i) CQs only, (ii) CQs + sentence pairs, and (iii) all data (CQs, sentence pairs, and documents). As expected, Quality decreased as data were progressively removed. Still, all outputs remained coherent and structurally valid. Outputs of the ablation study are publicly available (Christian, 2026a). Specifically, in Tahitian, Quality decreased from 6.5 with all data to 6 when documents were removed, and remained at 6 when sentence pairs were also removed. Although explanations became less well-documented and fewer examples were available, the perceived Quality did not degrade sharply. This indicates that, for the negation topic, CQ data alone can provide sufficient explanatory evidence to support useful

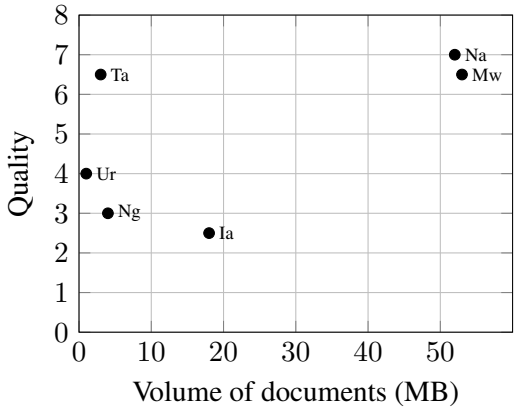

Figure 5: Quality as a function of the volume of documents.

lesson generation, particularly when key grammatical parameters are also attested in databases such as WALS.

5.4 Recommendation Scores.

Recommendation scores, aggregating expert judgments on usefulness for fluent and non-fluent learners, suitability for teachers and linguists, and perceived reductions in preparation time, follow the same general pattern as Quality. Languages with more available data fall within a narrow upper band (6.8–7.6), while the lowest-resource languages score lower (5.4–5.8).

Crucially, **the experts' ratings for "recommendation to teachers" remain high across all languages**, with a mean of 7.66. This suggests that, even when data are sparse, the generated lessons are viewed as practical and usable by educators. A similar trend holds for "recommendation to linguists" (mean 7.33), indicating strong perceived usefulness in a community-led revitalization context. Overall, these results reflect a genuine appetite among experts for tools that can reduce lesson-preparation time while maintaining acceptable pedagogical quality.

5.5 Error Analysis

A qualitative review of the generated lessons reveals four main categories of errors.

Factual errors. Most factual errors arise from fragile inferences that are presented as definitive statements in the lesson. As expected, these errors occur most frequently in the lowest-resource languages (Ngen and Uruangnirin), but they also appear in Iaai, despite its relatively large volume of available documentation. In this case, the issue

stems from a lack of relevant information about negation in the documents themselves, illustrating that document volume alone is not a reliable predictor of factual accuracy at this scale.

Omissions. Even when multiple negation strategies are attested in the corpus or the documents, some may be omitted from the generated lesson. These omissions typically affect less frequent or more specialized constructions and reflect the system's tendency to prioritize the most salient patterns when evidence is sparse.

Variability. Although the three lessons generated for a given language are generally similar, they are not identical. A negation pattern correctly identified in one lesson may be absent from another, reflecting residual variation in controlled generation. While this variability is limited, it can result in uneven coverage across lessons.

Metalanguage. Lessons intended for non-linguist audiences require careful adaptation of metalanguage to the target readership and the chosen language of instruction. HYGRAM generally produces appropriate pedagogical explanations, but in some cases specialized terminology from source documents is carried over without sufficient simplification or rephrasing, which reduces accessibility for non-specialist users. In other cases, the chosen metalanguage may be misleading.

5.6 Community Evaluation

HYGRAM has received positive feedback from communities across the Pacific region. In October 2025, we conducted an evaluation in Vanuatu, where 135 languages are spoken and where the national education strategy aims to make at least 60 of these languages viable options for early primary schooling. Following our presentations, a cooperation agreement was signed between our university and the National University of Vanuatu, with formal support from the Ministry of Education. A presentation and hands-on training workshop was conducted with local staff. This workshop represented the first opportunity for most participants to work on documenting their own language without requiring a linguist or extensive prior training.

Of the 18 participants, 8 completed a feedback survey. On the item "How relevant and helpful is DIG4EL for supporting the local languages of Vanuatu?" respondents assigned an average score of 7.6, indicating **strong perceived usefulness**. A multi-year project is now being set up to support the use of the software in schools. The system is also already being experimented with for educational purposes in four languages in French Polynesia.

6 Limitations

This work has multiple limitations in its pursuit of practical solutions, including the following.

First, evaluation is conducted on a limited number of languages (six for expert evaluation; 27 currently supported as of April 2026), which does not allow statistical generalization despite deliberate typological diversity. Broader evaluation across additional language families, documentation profiles, and elicitation conditions is required to assess robustness.

Second, expert evaluation is constrained by the reality of ELRL work: for many languages, only one academic specialist is available. As a result, each language in our study was evaluated by a single expert rather than by multiple independent raters, which limits calibration and inter-rater comparison. Accumulating additional evaluations over time will be necessary to strengthen reliability.

Third, evaluation relies on expert judgments of lesson quality and pedagogical usability because no gold-standard benchmark of beginner-oriented grammar lessons exists for most ELRLs.

Fourth, the present evaluation does not directly measure classroom effectiveness, teacher workload reduction, or learner outcomes, which are central to real-world adoption. The next phase of this research, including ongoing deployments in the Pacific region, is intended to address these dimensions.

Finally, the current implementation relies on online services for embedding and LLM inference, which constrains deployment in low-connectivity environments; fully offline or locally deployable versions remain future work.

7 Risks

Content inaccuracies: Despite constrained generation and evidence-based aggregation, LLM-assisted components may still produce unsupported statements, omissions, or overly confident formulations. If such issues are not identified during human review, they may propagate into teaching materials.

Ignoring minority dialects: The generation of structured pedagogical material may implicitly

privilege specific variants or analyses, potentially underrepresenting natural linguistic variation and ignoring dialectal diversity.

To mitigate these risks, HYGRAM is explicitly designed as a human-in-the-loop system: all outputs are intended as provisional resources, accompanied by uncertainty cues where available, and require expert or teacher validation, correction, and contextual adaptation before classroom use.

Conclusion

We introduced Hybrid Grammar Induction (HYGRAM), a method for generating pedagogically structured grammar lessons for endangered languages from extremely sparse data. HYGRAM combines typological priors, Bayesian inference, and controlled LLM-based reasoning to operate in settings where conventional NLP approaches are infeasible.

Evaluation across six endangered languages shows that HYGRAM produces consistent and usable lessons once a modest amount of data is available, with strong expert recommendations for teacher use even under low-resource conditions. Community feedback further indicates its practical relevance for revitalization contexts.

This work shows that HYGRAM can meaningfully support grammar teaching and documentation under severe data constraints. Future work will focus on larger-scale evaluation, classroom impact studies, improved handling of low-frequency constructions, and offline deployment.

Acknowledgments

We thank the speakers of the six languages considered in this study for sharing their knowledge and linguistic expertise, and acknowledge that they should remain the primary decision-makers in any revitalization effort. We are also grateful to the six linguists who evaluated the generated outputs. An AI assistant was used for grammar, wording, and clarity.

References

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, Christopher Cox, Katya Aplonova, Guillaume Jacques, and Nathan Hill. 2021. User-friendly automatic transcription of low-resource languages: Plugging ESPnet into elpis. Version Number: 2.

Cheryl Black and Andrew Black. 2009. PAWS: Parser and writer for syntax, drafting syntactic grammars in the third wave. In *SIL Forum for Language Fieldwork*.

Matthias Brenzinger and Patrick Heinrich. 2013. The return of hawaiian: language networks of the revival movement. *Current Issues in Language Planning*, 14(2):300–316.

Lyle Campbell, Gary Holton, Anna Belew, Lyle Campbell, Gary Holton, and Anna Belew. 2017. The catalogue of endangered languages as CLDF dataset.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. *Data Science Journal*, 19.

Aditi Chaudhary, Arun Sampath, Ashwin Sheshadri, Antonios Anastasopoulos, and Graham Neubig. 2022. Teacher perception of automatically extracted grammar concepts for l2 language learning. Version Number: 1.

Sebastien Christian. 2025. Enhancing grammatical documentation for endangered languages with graph-based meaning representation and loopy belief propagation. *Natural Language Processing Journal*, 12:100164.

Sebastien Christian. 2026a. Dig4el inputs and outputs on negation. Dataset. Zenodo. DOI: 10.5281/zenodo.19480176.

Sebastien Christian. 2026b. Dig4el software. Software. Zenodo. DOI: 10.5281/zenodo.19465517.

Caio Corro and Sylvain Kahane. 2024. Sparse logistic regression with high-order features for automatic grammar rule extraction from treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15114–15125. ELRA and ICCL.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022. Development of automatic speech recognition for the documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882, Marseille, France. European Language Resources Association.

Anne-Laure Dotte. 2017. Dynamism and change in the possessive classifier system of iaai. *Oceanic Linguistics*, 56(2):339–363. HAL Id: hal-02877380.

Matthew Dryer and Martin Haspelmath. 2013. The world atlas of language structures online.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan Van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL endangered language pipeline and inference system (ELPIS). In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 205–209. ISCA.

Alexandre François. 2023. *A Mwotlap–French–English Cultural Dictionary*. HAL open archive, hal-04838987.

Alexandre François. 2019. A proposal for conversational questionnaires.

Kristen Howell and Emily M. Bender. 2022. Building analyses from syntactic inference in local languages: An HPSG grammar inference system. *Northern European Journal of Language Technology*, 8(1).

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with FAISS. *arXiv preprint arXiv:1702.08734*.

Ilina Kachinske and Robert DeKeyser. 2019. The interaction between timing of explicit grammar explanation and individual differences in second language acquisition. *Journal of Second Language Studies*, 2(2):197–232.

Tatiana Korol. 2022. Preliminary description of ngen pronominal elements. *Mandenkan*, 68:43–58.

A. Nabizadeh, A. Taghinezhad, and Maral Azizi. 2016. The effect of implicit / explicit instruction on learning english grammar. *Modern Journal of Language Teaching Methods*.

Daniel Nettle and Suzanne Romaine. 2000. *Vanishing Voices: The Extinction of the World's Languages*. Oxford University Press, Oxford.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Racquel-María Sapién and Tracy Hirata-Edds. 2019. Using existing documentation for teaching and learning endangered languages. *Language and Education*, 33(6):560–576.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Sheena Shah and Matthias Brenzinger. 2018. The role of teaching in language revival and revitalization movements. *Annual Review of Applied Linguistics*, 38:201–208.

Hedvig Skirgård, Hannah J. Haynie, Harald Hammarström, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Noor Karolin Abbas, Sunny Ananth, Daniel Auer, and 100 others. 2023. Grambank v1.0.

Nick Thieberger. 2026. Nafsan, a language from central vanuatu: time depth and variability in records spanning 160 years. *Australian Journal of Linguistics*.

Jacques Vernaudon. 2018. Les métalangues du tahitien à l'école. *Contextes et Didactiques*, (12). En ligne, mis en ligne le 15 décembre 2018.

Eline Visser. 2026. A grammar sketch of uruangnirin. *Journal of the Southeast Asian Linguistics Society*, 19(1):clvi–cxciv.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, and 34 others. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Appendix

A Frontier-Based Parameter Discovery

Details of the selection of grammatical parameters

Defining the seed We define a *seed*, set of strongly supported values, as

$$S = \{ v \in \mathcal{V} : b(v) \geq \theta_{\text{belief}} \},$$

with $\theta_{\text{belief}} = 0.9$. The seed acts as the origin for the frontier expansion.

Given the posterior belief distribution $b(v)$ over all parameters $v \in \mathcal{V}$, the parameter discovery stage identifies parameters whose values are (i) highly probable and (ii) typologically connected to the seed in the WALS/Grambank graph G.

Graph pruning and bounded expansion. Then, using the conditional probabilities between values, the process is to follow paths bringing new values with still strong beliefs:

Let W_{uv} denote the edge weight between value nodes u and v. Edges with insufficient conditional support are removed:

$$W_{uv} \leftarrow \begin{cases} W_{uv} & \text{if } W_{uv} \geq \theta_{\text{CP}}, \\ 0 & \text{otherwise.} \end{cases}$$

A breadth-limited expansion to depth d defines path strengths via

$$(W^k)_{sv} \;=\; \sum_{s=u_0,u_1,\ldots,u_k=v} \prod_{t=0}^{k-1} W_{u_t u_{t+1}}$$

$$1 \leq k \leq d.$$

Influence of strong seeds. For any value v, its influence score from the seed is is

$$I(v) \;=\; \max_{s \in S} \Big[b(s) \max_{1 \leq k \leq d} (W^k)_{sv} \Big].$$

This quantity captures how strongly v is linked to highly certain values through reliable, short paths in G.

Parameter-level scoring. Each parameter P corresponds to a set of values $\mathcal{V}_P$. Its posterior mass is

$$B(P) \;=\; \sum_{v \in \mathcal{V}_P} b(v),$$

and its maximal connectivity to strong seeds is

$$C(P) \;=\; \max_{v \in \mathcal{V}_P} I(v).$$

A combined score ranks parameters for elicitation:

$$\text{Score}(P) \;=\; B(P)\, C(P).$$

Parameters with $\text{Score}(P) \geq \theta_{\text{score}}$ are selected as part of the Loopy Belief Propagation process that concludes the inferential process on grammatical parameters.

B HYGRAM Flowchart

Figure B.1 presents the full HYGRAM workflow, including intermediate inference and retrieval steps.

C Sentence Augmentation

C.1 Sentence descriptors

Listing 1: Sentence Descriptors

```
    intent: List[Intent]
    mood: List[Mood]
    act_of_speech: List[ActOfSpeech]
    type_of_predicate: Optional[List[
        TypeOfPredicate]] = None
    modality: Optional[List[Modality]] =
         None
    evidentiality: Optional[List[
        Evidentiality]] = None
    voice: Optional[List[Voice]] = None
    polarity: Optional[List[Polarity]] =
         None
    tense: Optional[List[Tense]] = None
    aspect: Optional[List[Aspect]] =
        None
    spatial_location: Optional[List[
        SpatialLocation]] = None
    directionality: Optional[List[
        Directionality]] = None
    sentence_complexity: list[
        SentenceComplexity]
    possession: Optional[List[Possession
        ]] = None
    personal_pronouns: Optional[List[
        PersonalPronoun]] = None
    other_pronouns: Optional[List[
        OtherPronoun]] = None
    classifiers: Optional[List[
        ClassifierType]] = None
    numbers: Optional[List[
        NumberQuantifier]] = None
    key_translation_concepts: List[str]
```

C.2 Sentence augmentation prompt

Listing 2: Sentence Descriptors

```
You analyze one English sentence and
    return only JSON that instantiates
    the Sentence schema below.
Be exhaustive but conservative: include
    what is textually expressed or
    licensed by clear grammatical cues.
    Do not infer world knowledge.
When unsure, leave fields empty lists (
    or "" for comment). Use "other" only
     if the sentence clearly encodes a
    category that is not listed.

OUTPUT RULES:
- Return JSON only (no prose, no
    markdown, no code fences).
- Echo the input in "original_sentence".
- Fields typed as lists must be lists (
    even with one value). If nothing is
    expressed, use [].
- Do not invent labels. Use exactly the
    strings in the allowed sets below.
- Prefer precision over recall: omit if
    uncertain.
```

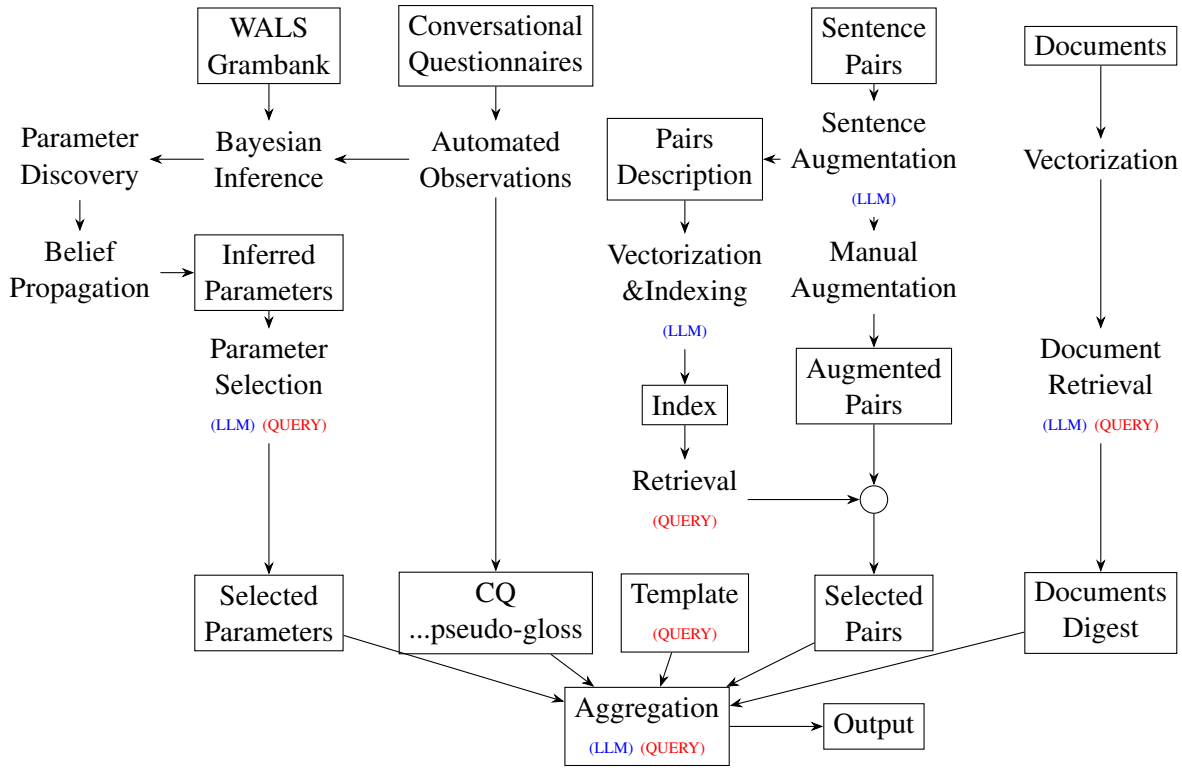

Figure B.1: Full HYGRAM workflow. The diagram shows intermediate inference, augmentation, indexing, and retrieval steps omitted from the main text for clarity. Blue (LLM) labels indicate LLM-based components; red (QUERY) labels indicate query-dependent operations.

```
- If multiple values are present for a
    category, list them all in salience
    order.

MAPPING RULES:

- Intent:
  declaratives -> "assert";
  yes/no or wh-forms / tag questions ->
      "ask";
  imperatives/requests -> "order";
  "I think/maybe/perhaps" -> add "doubt"
      ;
  "wish/hope/if only" -> add "wish";
  greetings -> "greet";
  thanks -> "thank";
  "okay/understood" -> "acknowledge".
  Multiple intents allowed.

- Mood
  Imperative forms ("Do X", "Please ..."
      , "Let's ...") -> "imperative"
  (and often "hortative" for "let's", "
      injunctive" for negative commands
      "don't").
  Questions (final "?" or inversion/wh-
      fronting) -> add "interrogative".
  Counterfactuals/wishes ("if I were..."
      , "I wish...", "would that...") ->
      "subjunctive" or "optative".
  Obligations ("must/should/ought to")
      -> add "necessitative".
  Permissions ("may/can" in permission
      sense) -> add "permissive".
  Mild 3rd-person exhortations ("Let him
      ...", "May she...") -> "jussive".
  Desire/volition ("want/plan/try/let's"
      ) -> "desiderative" and/or "
      hortative".
  Otherwise default includes "indicative
      " for asserted finite clauses.

- Act Of Speech
  Statements -> "assertive";
  commands/requests/suggestions -> "
      directive";
  promises/offers ("I promise/shall...",
       "I'll...") -> "commissive";
  thanks/apologies/congratulations -> "
      expressive";
  institutional acts ("I resign", "I
      pronounce...") -> "declarative";
  greetings/small talk -> "phatic";
  explicit performatives ("I apologize",
       "I promise") -> "performative";
  questions -> also include "
      interrogative".

- Type Of Predicate
  Sentence about existence -> "
      existential";
  about location (time or space) -> "
      locative";
  about qualities -> "attributive";
  about possession -> "possessive";
  about category/nature -> "inclusive";
  about something happening -> "
      processive";
  otherwise -> "other".
```

```
- Polarity
  Clausal negators ("not/never/no/nobody
      /nothing/without", contracted
      forms) ->
       "negative".
  Mixed scopes (one clause negated,
      another positive; or both present)
       -> "mixed".
  Otherwise -> "positive".

- Tense
  Past morphology ("VBD", "was/were", "
      did") -> "past".
  Present ("VBP/VBZ", present copula) ->
       "present".
  Future ("will/shall/going to"; or
      explicit "tomorrow/next year") ->
       "future", or "near_future" for "
          about to/soon", "remote_future
          " for far-future.
  Past perfect ("had + V-en") -> include
       "perfect" in aspect and "past" in
       tense.
  "used to / was V-ing" as habitual/
      ongoing background -> "
      imperfect_past".
  Gnomic: encode under aspect, not tense
       ("gnomic").

- Aspect
  Progressive: be + V-ing.
  Perfect: have + V-en.
  Habitual: adverbs ("usually/often/
      always") or generic presents ("
      Cats purr").
  Prospective: be "going to" / "about to
      ".
  Inceptive/terminative: "start/begin to
      ", "finish/stop".
  Iterative/frequentative/semelfactive/
      durative/resultative/experiential/
       continuative when adverbially or
          lexically encoded.
  Perfective for bounded/completed
      events; imperfective for ongoing/
      unbounded.
  Gnomic for timeless truths or general
      facts.

- Voice
  Passive: be/get + past participle,
      logical subject demoted; optional
      "by"-phrase.
  Causative: "make/let/have/get + object
       + V".
  Middle/antipassive/applicative:
      include only when explicitly
      marked.
  Default: "active".

- Modality
  Dynamic-ability: can/could (ability).
  Dynamic-volition: want/try/plan/let's.
  Desiderative: wish/hope/want (desire
      sense).
  Deontic-obligation: must/should/ought/
      need to/have to.
  Deontic-permission: may/can (
      permission).
  Deontic-prohibition: mustn't/shouldn't
      /don't (imperative negation).
  Epistemic-*: may/might/could/probably/
      definitely/seem/apparently/etc.

- Evidentiality
  Include only if overtly marked:
    "I heard/they say" -> hearsay/
        reported;
    "apparently/seems/looks/sounded" ->
        inferred/visual/nonvisual.
  Omit when not expressed.

- Sentence complexity
  One clause -> simple.
  Coordination with clausal "and/or/but"
       -> coordination.
  Subordinator/relative/embedded clause
      -> subordination.
  Both -> compound_complex.
  Clear ellipsis/headline -> elliptical.
  Else -> unknown.

- Pronouns
  List all persons/numbers present.
  For "we": if "let's", treat as "first
      person plural inclusive";
       otherwise include both inclusive
          and exclusive.

- Other pronouns
  demonstratives: this/that/these/those
  interrogatives: who/what/which/where/
      when/why/how
  indefinites: some/any/someone/none
  relatives: who/that/which
  reflexives: myself/yourself/etc.
  reciprocals: each other

- Numbers / classifiers
  numbers/quantifiers: one, two, some,
      many, all, none, a few.
  classifiers=["gender"] only if
      morphologically relevant; omit in
      English.

- Spatial location / directionality
  Include only when overt:
    here/there/this/that -> proximal/
        distal;
    come/go/back/toward/away/in/out/
        across/along/up/down ->
        directionality.
  Use uphill/downhill/seaward/landward
      only if explicitly expressed.

- Possession
  "X's Y / of-genitive / my/your/his/her
      " -> attributive (and
      genitive_case if marked).
  "X has Y" / have/has -> have_verb and
      predicative.
  Add "alienable"/"inalienable" as
      appropriate (body parts =
      inalienable).

- key_translation_concepts:
  A list of short anchors copied from
      the English sentence.
```

```
  Choose items whose translation has
      highest cross-linguistic
      explanatory power.

  Examples:
    "Please don't open the door." ->
        ["do not (neg)", "please (
            politeness)", "open", "door"
            ]
    "Let's head back inland tomorrow."
        ->
        ["let us (hortative)", "back (
            dir)", "inland (landward)",
            "tomorrow (future time)"]
    "If I had known, I would have called
         you earlier." ->
        ["if (conditional)", "had (
            perfect)", "would have (
            conditional+perfect)", "
            calling", "earlier"]
    "I have never seen those mountains."
         ->
        ["have ... seen (perfect)", "
            never (neg)", "those (
            demonstrative distal plural)
            ", "mountains (plural)"]
    "The report was completed because
        the data couldn't be accessed."
        ->
        ["was completed (passive)", "
            because (causal)", "could
            not (modal+neg)", "be
            accessed (passive)", "data"]

- comment:
  <=120 characters; include ambiguity
      notes or parsing rationale when
      useful.
```

C.3 Grammatical Parameters Selection Prompt

Listing 3: Grammatical Parameter Selection Prompt

```
You are a grammar expert selecting, from
     a provided list, the grammatical
    parameters
most relevant to answering a user's
    query.

INPUT:
- A user query.
- A list of grammatical parameters.

OUTPUT:
Return the subset of input parameters
    that are relevant to answering the
    user's query.

EXAMPLE:
User query: "How does Arawak express
    aspect?"
Parameters list:
[
  "Is a morphological distinction
      between perfective and
      imperfective aspect available on
      verbs?",
  "voice",
  "The Perfect",
  "The Past Tense",
  "Perfective/Imperfective Aspect"
]

Expected output:
[
  "Is a morphological distinction
      between perfective and
      imperfective aspect available on
      verbs?",
  "Perfective/Imperfective Aspect"
]
```

C.4 Document Content Retrieval Prompt

Listing 4: Document Content Retrieval Prompt

```
You are an agent specialized in
    retrieving grammatical information
    about {indi_language} in the
    provided documents.
    to answer a user's query. Retrieve
        all relevant information from
        the documents and compile them
        into a detailed
    answer to the user's query, with
        examples taken from the
        documents.
    - Use only information from the
        documents. Do not invent any
        additional information or
        examples. If there are no
        relevant
    information in the documents, just
        output "no relevant information
        about the query in the documents
        ".
    - If the query comes with
        instructions about how to
        formulate an answer, follow
        these instructions.
    USER QUERY: {query}
```

C.5 Aggregation prompt

Listing 5: Aggregation Prompt

```
You are an agent specialized in creating
     grammar teaching material for an
    endangered language.
Students speak the source language.
You receive a user query (with optional
    instructions) and multiple sources
    of information.
Your task is to compile these sources
    and add your own input to create a
    grammar lesson
about the query. You must follow any
    instructions included in the query
    about how to think
and how to formulate the answer. Your
    output is a grammar lesson following
     the schema below.
```

```
COMPLY WITH THE FOLLOWING:
- Do not add information from any other
    source or from your own knowledge.
- Add ALL relevant information from ALL
    provided sources.
- Follow any instructions included in
    the user query.
- Adapt the output to the readers: write
     in the readers' language, match
    their level,
  and highlight contrasts between the
      endangered language and the
      readers' language.
- All lesson content must come from the
    provided material only. Do not
    invent content.
- In all examples, translate the English
     source sentence into the readers'
    language.
- Avoid metalinguistic jargon. Use
    everyday words and periphrases,
    especially if the
  audience is young.

INPUT:
- Name of the endangered language.
- Query and optional instructions about
    how to create the lesson.
- Type of readers and their language.
- List of grammatical parameters and
    their values in the endangered
    language
  (some with examples).
- Description from sentence analysis.
- Description from a compilation of
    documents.
- List of examples, each with its
    description.
- List of sentence pairs, some with
    explicit mappings between the
    concepts in the
  English sentence and the words in the
      endangered language.

NOTE ON INPUTS:
If there are contradictions between
    sources, be explicit about them and
    cite each source.

OUTPUT:
A grammar lesson in the readers'
    language, structured as follows:

1. Title
   - Derived from the user query.

2. Introduction
   - A simple, language-independent
       explanation of the grammar topic.
   - A short description of how this
       topic works in the readers' own
       language, with examples.
   - A hint about how the endangered
       language expresses this topic.
   - Purpose: help readers understand
       the topic without assuming prior
       grammar knowledge.

3. Information chunks
   - A list of focused paragraphs, each
       covering one aspect of the topic.
   - Each chunk contains:
       * A title.
       * An explanation.
       * Several examples from the
           corpus (preferably 5 when
           possible).
       * For each example: the
           endangered-language sentence,
            its translation into the
          readers' language, and a short
              explanation using the
              grammar topic lens.

4. Conclusion
   - What students must absolutely
       remember.

5. Drills
   - Sentence pairs illustrating the
       topic that can be used for
       exercises.
   - Translate the source-language
       sentences into the readers'
       language when needed.
   - Include ALL relevant examples
       provided in the input.

NOTE ON OUTPUTS:
When referring to a target word or
    sequence of words inside an
    explanatory sentence,
surround it with "**". Example: "**uru
    **", "**ia ora na**".
```

C.6 Example of Template

Listing 6: Example of template associated to the "Negation" prompt, created by Jacques Vernaudon, University of French Polynesia.

```
"guidance": "Each language uses a 'most
    frequent negation pattern' to
    express negation.
It typically prefers one of the
    following three patterns:

1. Specialised negative particles
   (Example: 'not' in 'This is not a dog
       ') (~50% of languages).

2. Affixes
   (Example: 'im-' in 'This is
       impossible' = 'This is not
       possible') (~33% of languages).

3. Negative verbs
   (Example: 'ita / 'ere / 'ore in Proto
       -Tahitian) (~12% of languages).

You must determine which pattern is the
    most frequent negation strategy in
    the target language.
However, you must also take into account
     any other patterns that appear in
    the examples.
```

```
You must also consider the 'negation
    scope'.
- Sentence negation applies to the
    predicate and its arguments
  (Example: 'The dog is not eating the
      fish').
- Constituent negation applies only to
    part of the clause, often not the
    predicate
  (Example: 'He participated in a non-
      violent protest').

When analysing sentence negation,
    examine these specific types:

- Negation of existence
  (Example: 'There is no dog').

- Negation of location
  (Example: 'The dog is not there / in
      the house').

- Negation of category membership
  (Example: 'This is not a dog').

- Negation of identification with a
    specific entity
  (Example: 'This is not my dog').

- Negation of a quality
  (Example: 'This is not small').

- Negation of a process
  (Example: 'The dog is not running').

If the negative words or structures
    differ across these types, you must
    explicitly present
those formal differences.

In all cases, consider how aspectual
    nuances are expressed.
Examples:
- 'There are no birds'
- 'There are no more birds'
- 'There are no birds yet'.

When a sentence involves possession,
    determine whether the negation scope
     targets existence
or the possessive relation:

- 'I have no dog' = 'There is no dog in
    relation to me'
  (negation scope on existence).

- 'This dog is not mine' = 'This dog is
    not in relation to me'
  (negation scope on the possessive
      relation)."
```

C.7 Output's excerpt

Introduction to Negation in Tahitian Every language has ways to say things like "not," "no," "never," or "don't." In English, we use helpers like *don't/doesn't/didn't*, *can't*, *won't*, *isn't/aren't*, and words like *never* and *no*: "I don't know," "It isn't a dog," "Don't run," "There isn't any water," "I don't have a pencil," "I've never been there."

In Tahitian, there isn't just one single "not" word. Tahitian selects a different negator depending on the kind of sentence: facts and existence use *'aita*; future, ability, or "wouldn't" use *e'ita*; saying "X is not Y/Adj" uses *e'ere*; and telling someone not to do something uses *'eiaha*. You will also see helpers like *i* and *e* immediately after the negator. This lesson shows how each one works and how they differ from English.

Where the "no-word" goes in Tahitian In Tahitian, the negator comes first. After it, a small helper marks the verb phrase: *i* after *áita* (facts) and *e* after *e'ita* (future or ability). With *e'ere* ("is not"), you will often see *i* before the noun or quality. With *'eiaha* (prohibitions), the verb follows with *e*, or sometimes with *'ia* before a passive-form verb.

English places *not* or *do not* after the subject; Tahitian places the negator first.

Example

1. *'aita vau i 'ite*
   "I don't know."

Here, *'aita* stands first, followed by the subject (*vau* "I"), then *i*, and finally the verb. English says "I don't know," but Tahitian leads with the negator.