

# An Interactive System for Generating Revisable Grammar Lessons for Extremely Low-Resource Languages Without Expert Annotation

Sebastien Christian

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

## Abstract

Endangered-language teaching often faces two practical bottlenecks: the scarcity of experts able to produce pedagogical grammars, and the dependence of most approaches on expert linguistic annotation. We present DIG4EL, a human-in-the-loop system for extremely low-resource languages that addresses both constraints by combining lightweight concept-based annotation, typological inference, structured sentence-pair augmentation, document retrieval, and constrained language model generation. Rather than aiming to produce definitive grammatical descriptions, the system generates revisable grammar lesson drafts grounded in heterogeneous evidence, including elicited sentence pairs, free translation pairs, and descriptive documents. The interface is designed so that speakers, teachers, and other language practitioners without formal linguistic training can contribute usable data, inspect intermediate inferences, and control source selection during generation. We describe the architecture, user workflows, and initial deployment experience in real-world revitalization settings. The contribution of the paper is an implemented workflow for early pedagogical draft generation under extreme data scarcity, not a controlled evaluation of pedagogical effectiveness.

## 1 Introduction

Two long-standing bottlenecks make teaching the grammar of endangered languages difficult. First, most formal documentation workflows that support grammatical description rely on expert linguistic annotation. Second, the number of linguists available to produce pedagogical grammars is far below the scale of the need. For many communities, waiting for expert annotation or expert-authored teaching materials means waiting indefinitely.

This paper starts from a different operational assumption. When the goal is not a definitive reference grammar but a revisable pedagogical draft,

extremely small amounts of data can already be useful. A few hundred sentence pairs, lightly enriched through annotations that language speakers can produce themselves, may be sufficient to generate lesson drafts for foundational grammatical topics. In this setting, expert linguistic annotation remains highly valuable when available, but it is not treated as a hard prerequisite for producing useful teaching materials.

We focus on endangered languages with extremely limited data, typically ranging from a few hundred to a few thousand transcribed sentence pairs, and lacking standardized pedagogical grammars. We refer to this setting as *extremely low-resource languages* (ELRLs)<sup>1</sup>.

Teaching grammar is an important component of language revitalization, as many community members learn their ancestral language as an additional language (Kachinske and DeKeyser, 2019; Nabizadeh et al., 2016). Yet the production of pedagogical materials remains a major bottleneck. Teachers are often expected to create structured lessons despite limited time, limited formal preparation, and the absence of classroom-ready resources. Comprehensive grammars typically take years to produce (Woodbury, 2011), and even when they exist, they usually require substantial transformation before they can be used in teaching (Sapién and Hirata-Edds, 2019). At the same time, many teachers report feeling unprepared to create grammar-focused teaching materials (Chaudhary et al., 2025).

Existing tools and methods address parts of this problem but do not remove these bottlenecks. Linguistic software such as FieldWorks Language Ex-

<sup>1</sup>We use the term *extremely* because, for instance, NLLB Team et al. (2024) consider languages with fewer than one million sentence pairs *low-resource*, and languages with fewer than 100,000 sentence pairs *very low-resource*. These thresholds remain orders of magnitude above the data conditions encountered in most community-led documentation and revitalization efforts.

plorer (SIL, 2025) supports corpus management and lexicon building, but does not generate pedagogical lessons and still assumes substantial expertise. Earlier systems such as PAWS (Black and Black, 2009) assist in drafting grammatical descriptions, but remain oriented toward linguistic analysis and require expert intervention. More recent approaches such as BASIL (Howell and Bender, 2022) and Autogramm (Corro and Kahane, 2024) focus on extracting formal grammatical structure from annotated data, but are not designed for pedagogical output or for operation under extreme data scarcity. Retrieval-augmented natural language processing (NLP) methods help retrieve relevant sources from a corpus, but they do not, by themselves, solve the problem of turning sparse, heterogeneous evidence into usable pedagogical drafts for non-specialist users.

We present an interactive system designed to reduce both bottlenecks. It enables speakers, teachers, and revitalization supporters without formal linguistic training to contribute usable data and generate editable grammar lesson drafts.

The system combines existing typological information, probabilistic inference, structured sentence augmentation, retrieval, and constrained generation by language models, used here for their ability to reason over linguistic data (Zheng et al., 2025) and format outputs. It is designed around usability, inspectability, and the progressive enrichment of available data. Its outputs are explicitly provisional: users are expected to inspect, revise, and adapt them before use. The system trades quality for existence.

The broader governance approach of the system is intended to align with CARE (Carroll et al., 2020) principles and, where appropriate, with FAIR (Wilkinson et al., 2016) principles in ways compatible with community governance requirements.

## Contributions

- We formulate grammar-lesson generation for ELRLs as a distinct problem characterized by two practical constraints: scarce expert annotation and scarce linguist-authored pedagogical grammars.
- We present a human-in-the-loop architecture that combines lightweight concept-based annotation, typological inference, structured sentence-pair augmentation, retrieval, and constrained generation to produce revisable lesson drafts.
- We describe interface and workflow choices that enable non-specialists to contribute structured evidence, inspect intermediate inferences, and control source use during generation.
- We report initial deployment experience in real-world revitalization settings, suggesting that the workflow can be operated by non-specialist users under deployment conditions.
- We release the source code as open-source software and provide access to a live interface.

**Scope: focused lessons rather than full grammars.** We deliberately target focused grammar lessons rather than full descriptive grammars. Under extreme data scarcity, bounded outputs on a user-selected topic are faster to generate and easier to revise.

**Software availability.** The version of the system described in this paper is archived on Zenodo (Christian, 2026). Ongoing development takes place in the public GitHub repository at <https://github.com/alterfero/dig4el>. The public interface is available at <https://dig4el.org>.

## 2 System Overview

The system can operate with any one of three input sources, although it benefits from combining them:

- **Annotated Elicited Pairs (AEPs):** a predefined set of English sentences and sentence segments to be paired with their translations into the ELRL, connections between expected concepts in the sentence and the word or words that express them in the ELRL (referred to as *concept-word(s) links*), optional back-translations from the ELRL to English, and comments;
- **Free Pairs (FPs):** additional sentence pairs from any trusted source, optionally enriched by users with concept-word(s) links, back-translations, and comments;
- **Documents:** descriptive resources such as academic papers, theses, sketches, or existing teaching materials.

This design reflects a central premise of the system: in ELRL settings, useful structure must often be extracted from whatever evidence is available when the target output is understood as provisional and revisable.

AEPs play a particularly important role because they are designed to collect structural information while requiring substantially less linguistic expertise than traditional annotation workflows. Each completed AEP contains (i) a source sentence in English, (ii) an optional translation of the English sentence into a lingua franca used in the field, (iii) a translation from English or the lingua franca into the ELRL, (iv) concept-word(s) links, (v) optional back-translations, and (vi) comments. In (iv), instead of requiring full morphological segmentation or glossing, the system asks users to indicate which word or words in the ELRL contribute to a given lexical or grammatical concept assumed to be present in the sentence. This is a deliberate trade-off: detailed annotation is costly and often unrealistic, whereas speakers and teachers can usually indicate which word or words in the sentence contribute to lexical or grammatical concepts. In our design, these many-to-many mappings between concepts and forms provide enough structure to support targeted inference and grounded draft generation while remaining far less demanding than expert glossing.

For example, the AEP based on the sentence *No, I don't cough.* asks the language documenter to identify the word or words in the translation, if any, that *contribute* to the concepts of (i) denial; (ii) reference to the speaker; (iii) negation; and (iv) coughing. These concepts are predefined for each AEP. If the ELRL translation is idiomatic and does not use these concepts directly, the language documenter is invited to enter a *back-translation*: a literal translation of the ELRL sentence into English.

We use the term *language documenter* broadly to refer to any person contributing to language description or teaching materials, including community members, teachers, external revitalization supporters, and linguists.

In the current implementation, the AEP inventory contains 215 prompts divided into five dialogs adapted from conversational questionnaires (François, 2019). These prompts can typically be completed in roughly ten hours within our workflow.

The output is a structured lesson on a user-

selected topic, including explanations, examples, and exercises. The lesson is intended to be inspected, corrected, and adapted by teachers or speakers rather than treated as a finished grammar. Figure 1 provides an overview of the pipeline.

The pipeline consists of seven stages:

1. **Data collection.** Users create or import AEPs, FPs, and documents through the interface.
2. **Grammatical inference.** Typological priors and automated observations extracted from AEPs are combined to infer values of grammatical parameters supported by the available evidence. These inferred values are exposed to the user for inspection and correction.
3. **Pseudo-glossing of AEPs.** AEPs are transformed into a linear representation combining the source sentence, the ELRL sentence, concept-word(s) links, and associated structural cues. This representation acts as a lightweight substitute for expert glossing.
4. **Augmentation and indexing of FPs.** FPs are automatically enriched with structured grammatical descriptions and cross-linguistically reusable concepts using schema-constrained language-model calls, and can then be further augmented by users through concept-word(s) links. These representations are vectorized and indexed for retrieval.
5. **Processing of descriptive documents.** Documents are segmented, embedded, and indexed so that relevant passages can be retrieved at generation time.
6. **Query and preferences.** The user specifies a lesson topic, either from a predefined list or through a free-form query, together with the desired output language and expected complexity level.
7. **Aggregation and formatting.** Constrained language-model calls combine inferred grammatical parameters, AEP pseudo-glosses, retrieved augmented sentence pairs, and retrieved document content into a structured grammar lesson with source traceability.

**Implementation note** The current implementation is available through an online interactive interface and does not require task-specific model

training. Instead, it relies on general large language models. This design is essential in ELRL settings, where available data are too sparse to support supervised pipelines.

### 3 System Components

The system combines probabilistic inference, structured representations, and retrieval-based generation. Each component is designed to produce interpretable intermediate outputs that can be inspected and controlled by the user.

#### 3.1 Grammatical Parameter Inference

The system estimates the values of grammatical parameters listed in major typological databases such as the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) and Grambank (Skirgård et al., 2023). These databases define collections of grammatical parameters and their possible values, and record, for each language, the values of a subset of those parameters. They are used by DIG4EL in three ways: (i) to retrieve known values for the target language when available; (ii) to provide statistical priors over parameter values; and (iii) to derive dependencies between values of different parameters.

Inference is performed over this network of interdependent grammatical parameters, using typological priors together with automated observations extracted from AEPs, following the Bayesian framing for grammatical parameter estimation from sparse observations explored by Christian (2025). Rather than attempting inference over a predetermined set of grammatical parameters, the system identifies the subset of parameters that can be reasonably supported by the available evidence. For each parameter, the interface exposes candidate values, confidence estimates, and the evidence streams that contributed to the inference, allowing users to validate or override the result.

#### 3.2 Sentence-Pair Representation and Augmentation

FPs are processed into structured representations that capture grammatical and semantic information and make them more comparable to AEPs. Each pair is augmented with a set of grammatical descriptors (e.g., tense, aspect, polarity, predicate type, pronouns, and clause type) and a set of cross-linguistically reusable semantic and grammatical concepts. These representations are pro-

duced through constrained language-model outputs.

Once sentence pairs have been augmented, users can optionally connect concepts to the word or words that express them in the ELRL sentences, as with AEPs, and add literal back-translations and comments. This approach avoids the need for full morphological annotation while still providing structurally useful information.

#### 3.3 Retrieval Layer

The system uses vector-based retrieval to select relevant examples and document content. Sentence-pair representations are embedded on the basis of their generated grammatical descriptions rather than raw surface text alone, allowing queries to retrieve examples on the basis of grammatical similarity instead of direct semantic overlap. User queries are encoded using the same representation, enabling retrieval of examples that match the intended grammatical phenomenon.

Document sources are processed separately through segmentation and embedding, allowing relevant descriptive passages to be retrieved. This retrieval layer provides grounded evidence for generation and allows the system to adapt dynamically to different topics and datasets.

#### 3.4 Aggregation and Constrained Generation

Based on user input, which includes the grammatical topic, the choice of output language (limited to widely supported languages), and the desired complexity level, a series of language-model calls aggregates information from multiple sources, including inferred grammatical parameters, AEP pseudoglosses, structured representations of retrieved sentence pairs, and retrieved document excerpts.

Generation is performed under explicit constraints: language models are instructed to rely only on the provided inputs, to explicitly highlight conflicts between sources, to return nothing if there is no supporting data, and to follow a predefined lesson structure. This structure includes (i) an introduction presenting how the grammatical topic is expressed in the chosen output language and an overview of how it is expressed in the ELRL, with emphasis on contrasts with the output language; (ii) a sequence of sections, each focusing on one aspect of the topic and composed of an explanation plus examples; (iii) a conclusion summarizing the most important points; (iv) additional examples that can



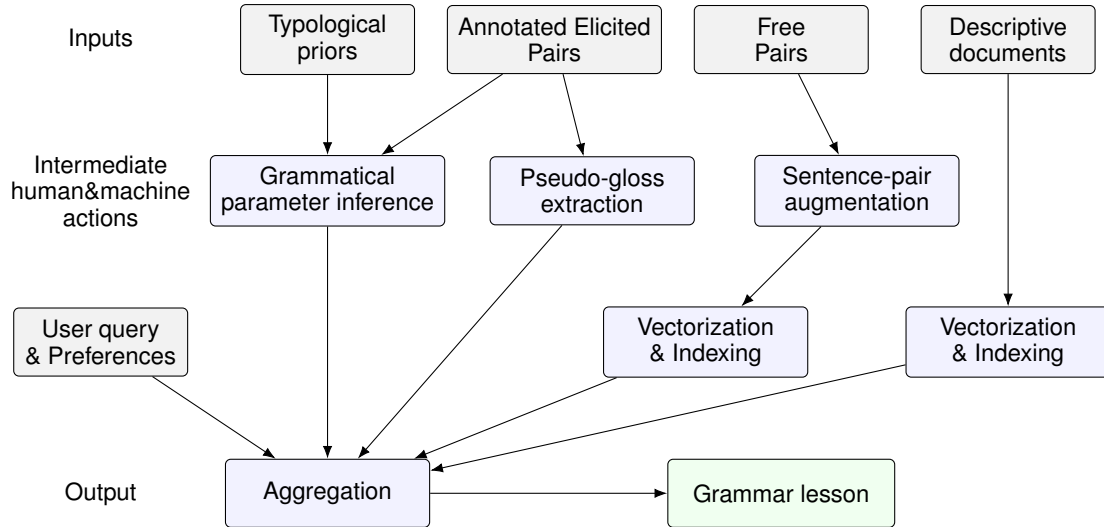


Figure 1: Pipeline of the system. Heterogeneous inputs are processed through inference, augmentation, and retrieval modules, then combined with the user query and generation settings to produce a structured grammar lesson. Intermediate representations remain inspectable and editable throughout the process.

be used as exercises; and (v) a list of the sources that contributed to the output.

#### 4 Interface Design Principles

We treat usability as a first-order requirement rather than an afterthought. The interface is designed to be usable by a broad range of language documenters and teachers.

User interactions can be divided into two broad categories: (i) providing data to the system and (ii) generating outputs. Data-entry workflows can tolerate moderate complexity, whereas generation workflows must remain simple and efficient. The interface is designed to reflect this asymmetry.

The system is organized around four design principles:

**Accessibility** The interface is designed to be usable by non-specialists with minimal training across a broad range of sensory and cognitive profiles. It minimizes technical and linguistic jargon, uses stable interaction patterns, separates more demanding data-entry tasks from simpler generation tasks, and offers on-demand details about processes.

**Human-in-the-loop control** Users retain substantial control over inference and generation. They can inspect and edit inferred grammatical properties, include or exclude sources, and revise generated outputs directly. Outputs are provided in forms that can be further edited and adapted for pedagogical use.

**Explainability through decomposition** The system maintains a separation between data sources and processing stages. Each source contributes independently to the final output, allowing users to trace generated content back to its origin.

**Progressive data enrichment** The system supports incremental improvement as new data become available. Outputs can be regenerated with updated inputs without retraining or reconfiguration, enabling iterative refinement of both data and pedagogical materials. A typical example is a teacher adding a new set of sentence pairs for a forthcoming lesson, which the system can incorporate into later reasoning and generation.

#### 5 User Workflows

The user workflows are designed to reduce the same two bottlenecks as the system itself: dependence on expert annotation at input time and dependence on linguist-authored grammars at output time. Accordingly, the interface separates workflows for making data available to the system from workflows for generating and revising lessons. Neither workflow requires formal training in linguistics.

Interaction is organized into three main workflows: (i) data entry and augmentation, (ii) grammatical parameter inference and validation, and (iii) lesson generation and revision.

## 5.1 Data Entry and Augmentation

The first workflow is designed for language documenters. Its purpose is to collect enough structured evidence for downstream inference and lesson generation.

Data entry is organized into three modules corresponding to the supported input types. All uploaded or created resources must include user-provided metadata, including provenance, licensing or reuse conditions, and confirmation that the data were collected with appropriate consent and may be processed by the system under the license governing its outputs.<sup>2</sup>

**Annotated Elicited Pairs (AEPs).** AEPs are the most structured input modality. Users translate a predefined set of elicitation prompts into the target language and add as many concept-word(s) links, back-translations in English or in a lingua franca, and comments as possible, as described in Section 2.

The interface supports both online entry and offline completion through downloadable spreadsheet templates that can later be re-imported. This hybrid design is important in practice, as many users work with intermittent connectivity or prefer offline data collection. Uploaded AEPs remain editable, allowing users to refine and enrich them over time. An example of the AEP interface is presented in Appendix 2.

**Free Pairs (FPs).** Users may also upload arbitrary collections of sentence pairs, using the provided spreadsheet template or exports from established corpus-collection software. These pairs are less constrained than AEPs and are intended to capture any available material.

Once uploaded, the user triggers their augmentation with structured grammatical descriptions and cross-linguistically reusable concepts by schema-constrained language models. This augmentation step is computationally heavier than standard data entry and is therefore handled as a long-running server-side process, with progress information displayed in the interface. After processing, users may enrich some pairs with concept-word(s) links, back-translations, and comments. In this way, FPs progressively move from raw examples toward structurally useful evidence.

<sup>2</sup>In the current implementation, resources retain their original license when one exists. Material created within the system is governed by the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 license.

**Documents.** Users can upload descriptive documents relevant to the ELRL, including academic papers, theses, sketches, or existing pedagogical materials. Once metadata have been entered, documents are segmented, embedded, and indexed automatically. These resources then become available for retrieval during lesson generation.

Taken together, these three input modalities allow the system to operate under a wide variety of data conditions.

## 5.2 Inference and Validation

The second workflow makes inferences about grammatical parameters visible and editable. This is central to the human-in-the-loop design: the system does not treat inferred structure as hidden internal state, but as provisional material that users can inspect and revise.

The system identifies a collection of grammatical parameters whose values can be robustly inferred. It infers those values and exposes them through the interface with optional supporting evidence. Users can review inferred grammatical properties and modify any inferred value accordingly. An example of this interface is provided in Appendix 3.

## 5.3 Lesson Generation and Revision

The third workflow is designed to be as straightforward as possible. The interface minimizes complexity and focuses on a small number of decisions directly relevant to lesson creation. Appendix 4 shows a screenshot of the generation interface.

Users specify a grammar topic, either by selecting a predefined topic or by entering a free-form query. They also choose the output language and the intended audience level. Based on these preferences, the system aggregates inferred grammatical parameters, pseudo-glossed elicited pairs, augmented sentence pairs, and retrieved document content into a structured lesson draft. Predefined grammar topics also benefit from response templates designed by linguists.

The generated output is intentionally framed as editable pedagogical material rather than as a definitive grammar. The interface also preserves traceability to the sources that contributed to the output, allowing users to inspect the basis of the generated content. When the system detects conflicts across sources or the absence of supporting evidence, it is designed to expose them explicitly in the output so that they can be resolved by the user.

This workflow operationalizes the core claim of the paper: teachers do not need to wait for a linguist-authored pedagogical grammar before they can obtain usable teaching material. With an appropriate workflow and interface, non-specialists can provide enough structure to support first-draft generation in at least some ELRL settings, after which teachers can adapt the output to their classroom needs.

## 6 Example

This section illustrates the workflow through the generation of a lesson on negation in Tahitian<sup>3</sup>. The available sources for this example include 215 Annotated Elicited Pairs (AEPs), collected in approximately ten hours; 3,000 Free Pairs (FPs) gathered from multiple sources; and one descriptive document: a French academic document covering selected topics in Tahitian grammar.

Tahitian is partially documented in the typological databases WALS and Grambank. When the grammatical-parameter inference module was run, 37 inferred values were presented to the user who was invited to either accepting them, changing them, or removing the corresponding parameter when uncertain.

The FPs were automatically augmented and indexed by the system. In the current configuration, this preprocessing step took approximately one day for the 3,000 sentence pairs.

The descriptive document was segmented, vectorized, and made available for retrieval within a few minutes after upload. These preprocessing steps are performed only once and are re-run only when the corresponding sources are modified.

On the *grammar generation* page, English was selected as the output language and *teenagers* as the intended audience level. Although the document is intended for teachers, the *teenagers* selection is often used as it removes most linguistic jargon from the output. The predefined topic *Negation* was chosen. The generation interface is shown in Appendix 4.

Because of the use of large language models, outputs are not identical across runs. However, repeated generations for the same query and source configuration consistently include the same core grammatical information. The generated lesson begins with an introduction that contrasts negation in the output language, here English, with negation in

Tahitian. One generated introduction is reproduced below.

Negation lets us say that something is not so, does not exist, will not happen, or must not be done. In English you usually add words like *not*, *no*, *don't*, *won't*, *isn't*, *never*, or *no longer*. English tends to keep the sentence shape and add *not* to the helper verb, or uses *no* for existence (*there is no water*), or uses *don't/won't* for commands and future. In Tahitian, the choice of the negative word depends on the kind of statement being made: actions that already happened or are viewed as facts, future events and abilities, equative “X is Y” sentences, existence/possession, and prohibitions each take a different dedicated negator. You will see four core forms throughout this lesson: *'aita* for factual events and for existence/possession, also as a free-standing “no”; *e'ital'eita* for future and ability; *e'ere* for “is not” classifications; and *'eiaha* for “don't.” Other useful pieces include *'ore* in certain subordinate clauses, the pattern with *i* before the verb after *'aita*, the use of *e* before the verb after *e'ital'eita* and *'eiaha*, and adverbs such as *a'e nei* “ever/never,” *fa'ahou* “any more/again,” and *ā* “still/not yet” under negation. Where English uses one general *not*, Tahitian selects the appropriate negative marker for the clause type and places it at the front of the clause.

The lesson then presents a sequence of sections covering different types of negation in Tahitian. In this generation, the first three sections after the introduction are: (i) *How Tahitian negation flips the sentence*; (ii) *'Aita: general “not,” past/experienced actions, and “there is no ...”*; and (iii) *E'ita: future/irrealis and general “won't / wouldn't / can't”*.

Each section contains examples retrieved from the corpus together with short explanations of their grammatical relevance. For example, the first section includes the following item:

### 1) *'aita, 'aita e 'ōro'a*

*No, there is no celebration.*

The pattern *'aita e NP* means “there is no NP.”

<sup>3</sup>Tahitian is used here because all resources are public.

Three complete generations on negation in Tahitian are available for download at <https://zenodo.org/records/20045485>.

## 7 Initial Deployment Experience

This section reports initial deployment experience and qualitative observations rather than a controlled evaluation of pedagogical effectiveness. DIG4EL is currently used with 24 typologically diverse languages, ranging from North Africa to Papua New Guinea. These include relatively well-documented languages, such as Mwotlap in Vanuatu, as well as minimally documented languages that test the limits of the system, such as Bwato in New Caledonia.

### 7.1 Workshop use by non-specialists

One illustrative deployment was a workshop held at Vanuatu National University. Vanuatu is home to more than one hundred languages. Participants came from diverse professional backgrounds and were all involved in supporting local languages. For many of them, this was their first experience documenting their own language without direct assistance from a linguist.

During the workshop, participants entered AEPs, created concept-word(s) links, and generated lesson content. Not every participant completed the full workflow in their own language during the session, largely because of time constraints, but the workshop showed that non-specialist users could learn and operate the core components after an initial introduction. The claim here is intentionally modest: the lightweight input workflow and the lesson-generation workflow were usable under workshop conditions. The workshop also suggested that participants valued documenting their own language directly and responded positively to generating and revising content. One outcome of this workshop was a formal cooperation agreement and deployment program that will pilot DIG4EL in a group of schools over the next two years.

### 7.2 Use within ongoing documentation work

Beyond workshop settings, the system has also been adopted by the CNRS Heliceo project, which aims to document a large number of Pacific languages and modernize parts of the documentation process.

In this context, the system is used not only as a generation tool, but also as a way to structure data

entry, lightweight annotation, and the progressive enrichment of language resources by a broad range of language documenters.

This second deployment matters because the system is being used as infrastructure for iterative data collection and reuse: AEPs, FPs, and uploaded documents can be expanded over time, reprocessed, and fed back into later lesson generation.

Taken together, these deployments support a bounded but important conclusion: the workflow is operational outside a laboratory setting and can support both first-draft lesson generation and incremental resource building.

## 8 Discussion

The system reflects a shift in how NLP can be applied to ELRLs. In such settings, the central challenge is often not maximizing benchmark performance, but enabling useful work to happen despite the absence of expert annotation, large corpora, and sufficient linguistic labor. The contribution of the system is therefore not to eliminate uncertainty, but to make progress possible in such documentation and teaching process.

More specifically, the system is designed to mitigate two long-standing bottlenecks. First, it reduces dependence on expert linguistic annotation by replacing glossing- or treebank-centered workflows with lightweight concept annotation that most speakers can provide. Second, it reduces dependence on linguist-authored pedagogical grammars by generating structured lesson drafts that teachers and speakers can revise.

Several broader observations follow from this design. First, grammar teaching material can be initiated by a broader range of language practitioners than is usually assumed in NLP pipelines, provided that the workflow and interface are designed accordingly. In ELRL contexts, shifting data collection and content creation toward speakers and community-based practitioners, rather than positioning them primarily as assistants to external experts, is a meaningful practical and ethical change. Second, combining heterogeneous sources of evidence improves robustness under extreme data scarcity. Third, transparency is not merely a usability feature, but a methodological requirement in contexts where uncertainty is unavoidable. Users must be able to inspect intermediate inferences, identify implausible outputs, and decide which sources to trust. Finally, human oversight



remains essential. The system is useful precisely because it supports revision, not because it removes the need for it.

**Risks and Limitations.** The same design choices that make the system usable under extreme scarcity also introduce risks and limitations.

- **Error propagation.** Generated outputs may be difficult for teachers to correct, especially when they appear authoritative or contain technical linguistic terms. This creates a risk that errors may be propagated rather than identified.
- **Reuse conditions.** The current implementation assumes that input materials can be processed under declared reuse conditions. This restricts applicability in contexts where communities, authors, or institutions impose cultural, ethical, or legal constraints on access and reuse.
- **Institutional barriers.** The system lowers technical and linguistic barriers, but it does not resolve social and institutional ones. Some teachers may still feel that they are not legitimate producers of explicit grammar lessons even when appropriate tools are available.
- **Usable evidence and willingness to revise.** The system’s usefulness depends on the availability of at least some usable evidence related to the grammar topic expressed in the query, on the relevance of retrieved materials, and on the willingness and ability of users to revise outputs.

## 9 Ethical Considerations

**Consent, rights, and governance** Language data and descriptive materials may be culturally sensitive, collectively governed, or subject to local restrictions that go beyond standard copyright. The system is therefore designed to require metadata about provenance, consent, and reuse conditions at upload time. The intent is to observe CARE first, then FAIR principles, which may diverge in practice depending on the situation, but no interface can guarantee that user-provided declarations fully capture local governance norms or community expectations.

**Representation of variation** Endangered languages often display dialectal variation, intra-community disagreement, or incomplete standardization. Any system that aggregates across examples and documents risks overrepresenting one variety, suppressing variation, or presenting contingent analyses as settled facts. This risk cannot be fully removed and is managed in the system by allowing users to specify the relevant language variety, if any.

**Access and dependency** The current implementation depends partly on online services and on data that can be processed under declared reuse conditions. These assumptions may exclude communities with limited connectivity or stricter cultural and legal controls over language materials. Future work should therefore prioritize offline deployment and more flexible governance mechanisms.

## 10 Conclusion and Future Work

We presented a human-in-the-loop system for generating revisable grammar lessons for extremely low-resource languages from minimal and heterogeneous data. By combining lightweight concept-based annotation, typological inference, structured augmentation, retrieval, and constrained generation, the system makes it possible to produce first pedagogical drafts without requiring expert annotation as a prerequisite.

The contribution of the paper is an implemented workflow and a deployed system design for early pedagogical draft generation under severe data scarcity. Initial qualitative feedback from multilingual communities and language documenters has been positive, although this approach also introduces its own technical, ethical, and social risks and limitations.

Future work includes controlled evaluation of lesson quality and usability, broader offline deployment, and deeper integration into educational and community-led documentation settings.

## Acknowledgments

This work was supported by the French CNRS Heliceo project under grant ANR-24-RRII-001.

We are deeply grateful to the language documenters and community members with whom we have had the opportunity to work. Their patience, feedback, and support have been essential in making the system possible and in adapting it to

practical language documentation and teaching contexts.

We also thank the Government of Vanuatu and its partner institutions for their willingness to support the first school-based deployment of DIG4EL.

## References

- 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*.
- 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*. 19.
- Aditi Chaudhary, Arun Sampath, Ashwin Sheshadri, Antonios Anastasopoulos, and Graham Neubig. 2025. *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*. 12:100164.
- Sebastien Christian. 2026. *DIG4EL*.
- 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.
- Matthew S. Dryer and Martin Haspelmath. 2013. Wals online. <https://wals.info/>. Max Planck Institute for Evolutionary Anthropology, Leipzig. Accessed 25 February 2026.
- 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*. 8(1).
- Ilina Kachinske and Robert DeKeyser. 2019. *The interaction between timing of explicit grammar explanation and individual differences in second language acquisition*. 2(2):197–232.
- A. Nabizadeh, A. Taghinezhad, and Maral Azizi. 2016. The effect of implicit / explicit instruction on learning english grammar.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, 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*. 630(8018):841–846.
- Racquel-María Sapién and Tracy Hirata-Edds. 2019. *Using existing documentation for teaching and learning endangered languages*. 33(6):560–576.
- SIL. 2025. *Fieldworks language explorer (flex)*.
- Hedvig Skirgård, Hannah J. Haynie, Haejoong Jang, Simon J. Greenhill, Harald Hammarström, Robert Forkel, Sebastian Bank, Claire Bown, Russell D. Gray, and Nicholas Evans. 2023. Grambank 1.0. <https://grambank.clld.org/>. Cross-Linguistic Data Formats (CLDF) database, CLLD platform. Accessed 25 February 2026.
- 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*. 3(1):160018.
- Anthony C. Woodbury. 2011. Language documentation. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages*, pages 159–186. Cambridge University Press.
- Tianshi Zheng and 1 others. 2025. *Logidynamics: Unraveling the dynamics of inductive, abductive and deductive logical inferences in large language model reasoning*. *arXiv preprint arXiv:2502.11176*.

## A AEP Augmentation Interface

A says: "I became sick after that."

Equivalent in Mwootap

basō, et misin te e nok gom

Validate translation

In this sentence, would you know which word(s) would contribute to the expression the following concepts?

Asserting (affirming):

Choose options

Becoming:

Choose options

Sick:

gom

After:

basō

Reference to an unknown event:

Choose options

A reference to the speaker.:

nok

What would be the literal English translation of this Mwootap sentence? (Relevant if you judge that the sentence in Mwootap differs significantly from the English original)

Comments/Notes

Liter. <Then, not a long time, and I (fell) sick.>

Validate expression and connections

Local dialog context

A: They were sweet.

A: I ate many of them.

A: I became sick after that.

B: I see.

B: It must have been these fruits that made you sick.

B: If you hadn't eaten so much, you wouldn't have gotten sick.

A: Oh Doctor, you're right.

A: I shouldn't have.

A: What should I do now?

B: Don't worry.

Figure 2: AEP interface showing data in Mwotlap, a language from Vanuatu. The sentence in English is at the top, with no lingua franca used. The right column helps the user keep the context in mind by showing the portion of the dialog surrounding the current sentence. Below the translation, each concept is listed with a box allowing the user to select none, one, or multiple words from the translation for each expected elicited concept. Below this, a text box invites the user to enter the literal back-translation, if relevant, followed by comments.

Figure 4: Interface used to generate grammatical descriptions. Outputs stored online can be directly displayed or downloaded; Outputs are not stored by default, the user makes the decision. The generation menu includes a selection of the format, the output language, and the complexity modeled as age brackets. The user then selects a pre-defined grammar topic or enter a free query before launching the generation.

## B Grammatical Parameters Inference Feedback Interface

### Parameter discovery

10 parameters observed, 0 known from WALs, 0 known from Grambank.

5 Strong parameters, enabling a reach of 239 other parameters.

Running General Agent with 69 parameters.

### Beliefs

☐ Let me edit beliefs

Based on statistical information, existing knowledge, observations and inferences across parameters, the following beliefs formed a consensus.

Parameter	Origin	Winner	Confidence
33 Exponence of Tense-Aspect-Mood Inflection	Inferred	monoexponential TA	92
17 Gender Distinctions in Independent Personal Pronouns	Inferred	No gender distinction	88
4 What is the pragmatically unmarked order of S and V in intransitive clauses?	Inferred	SV	87
31 Polar Questions	Inferred	Question particle	87
26 Order of Negative Morpheme and Verb	Inferred	NegV	86
21 Adjectives without Nouns	Inferred	Without marking	85

Figure 3: Interface showing the result of the grammatical-parameter inference process. At the top are the parameter-discovery results. Then, for each parameter, the interface displays the value currently inferred for the language, together with the confidence derived from the entropy of the distribution. The user can flip the *Let me edit beliefs* switch, which allows any value to be corrected before further processing.

## C Grammar Lesson Generation Menu

### Access stored outputs from previous queries

These outputs are **raw outputs** from DIG4EL, provided for research purposes. **They have not been corrected** by an expert of the language and may contain errors and inaccuracies. **They should not be used as is for teaching or learning** the language described.

Select a query to access the files

lesson\_Negation\_(English) ▼

Download JSON file

Download DOCX file

► Click here to see the output

### Generate a new grammatical description in Tahitian

Format

Grammar lesson ▼

What is the language of readers?

English ▼

The grammar is generated for...

Teenagers ▼

Choose a typical lesson topic

Select a standard grammar lesson...

no selection ▼

Or enter your custom query

... or enter your own topic.