

Worldview Annotation for Low-Resource Languages

Tetiana Ilman

Independent Researcher

worldviewannotationschema@gmail.com

Abstract

Low-resource language technology has advanced significantly in preserving linguistic form, yet a critical gap remains: the conceptual systems that languages encode — how communities structure personhood, agency, time, evidence, relationality, and moral responsibility — are systematically lost in current documentation pipelines. We argue that the problem is not only the annotation scale but annotation level. We introduce the Worldview Annotation Schema (WAS), presented here as a three-layer framework that begins with community documentators in the field and ends with machine-readable structured data, catching culturally grounded meaning before downstream AI analysis can homogenize it. We track where worldview meaning dissolves in existing pipelines, propose the WAS architecture with a worked example from Maa (Eastern Africa), and outline a governance model that grounds curator interpretation in community-provided evidence while producing graph-compatible structured output. In the age of AI, language preservation without worldview preservation risks producing languages that survive archivally while disappearing cognitively.

1 Introduction

More than 7,000 languages are spoken worldwide (Hammarström et al., 2020), yet global NLP resources remain concentrated in a small number of high-resource languages (Joshi et al., 2020), while many languages face ongoing endangerment—concerns reflected in international initiatives such as the International Decade of Indigenous Languages (UNESCO, 2022). Language loss is not merely the loss of another dictionary: languages carry oral traditions, ecological knowledge, social memory, and specific ontologies. When a language disappears, humanity loses a knowledge operating system — an epistemic and ontological loss that may also contribute to future AI bias.

This paper introduces a conceptual annotation framework designed to preserve worldview information in low-resource language documentation. Our goal is to define the architecture, annotation workflow, and governance principles that can support future empirical evaluation.

Traditional language preservation focused on documentation, literacy, and heritage transmission. Although essential, these approaches are no longer sufficient in the age of artificial intelligence. A language archived but not computationally integrated risks symbolic survival alongside functional disappearance. Preservation must now include the structured representation of the worldview systems encoded in language.

Advances in machine translation, speech recognition, corpus construction, and automated annotation have created new possibilities for low-resource language inclusion. Nevertheless, global NLP remains heavily concentrated in a small number of high-resource languages (Joshi et al., 2020). However, there is a foundational asymmetry. AI systems learn from what becomes machine-readable. Deeper conceptual distinctions encoded in the language often are not.

This study was motivated by an exploratory experiment that examined this asymmetry directly. In Ukrainian affective language contexts, human speakers consistently used diminutive morphology to express relational closeness and care, whereas the outputs of several AI systems omitted these forms almost entirely under neutral prompting. When explicitly prompted to use diminutives, outputs shifted to an exaggerated, saturated form that read as caricatured or performative rather than conveying genuine affective warmth. In our exploratory Ukrainian data, relational warmth was frequently expressed through layered diminutive morphology. This use is consistent with cross-linguistically documented affective functions of diminutives (Jurafsky, 1996). The exploratory com-

parison therefore suggests that one of Ukrainian’s most productive worldview-bearing features is consistently misrepresented during current AI language generation. If AI systems fail to preserve this distinction even in a well-documented language, the same compression is likely to have greater consequences in languages where the documentation infrastructure itself is limited. The complete experimental methodology and quantitative analysis are beyond the scope of this paper and will be reported separately. Here, the observation serves only to motivate the annotation framework introduced below.

Current annotation systems scale the analysis of observable content: topics, sentiment, named entities, discourse functions. However, many culturally meaningful distinctions exist at another level — how speakers encode evidential source, distribute agency, conceptualize time, or assign moral responsibility. These distinctions are not reliably recoverable through surface semantic labeling alone. The problem is not only the annotation scale but annotation level.

A language that survives in archives but loses its conceptual architecture has not fully survived. Preserving a language’s forms while losing its ontological structure is just a slower kind of disappearance. We frame WAS around low-resource languages specifically because these are the contexts where a lost worldview distinction has no fallback: no large parallel corpus, no established reference grammar, no standing community of professional linguists to reconstruct it later. Better-documented languages such as German or French retain enough alternative documentation to recover such distinctions after the fact even when an annotation pipeline flattens them once; thinly documented languages typically do not, so losses at the annotation stage tend to be final.

Contributions. This paper makes five contributions: (1) it repositions annotation as a site of epistemic decision-making where worldview distinctions are preserved or erased, introducing the Worldview Annotation Schema (WAS) to operationalize worldview through dimensions such as time, agency, and emotion; (2) it defines the current three-layer annotation architecture and a two-tier community/curator workflow, designed for non-technical first-tier language specialists (translators, educators, field linguists) and separating linguistic interpretation from technical curation; (3) it

treats non-alignment and untranslatability as valid outcomes, addressing the upstream data formation stage that benchmarks implicitly assume already been handled; (4) it illustrates the framework through a fully worked Maa proverb annotation, including two distinct documented usage events and a complete literal gloss; and (5) it reports a pilot field application with a native Maa-speaking community documentator, illustrating that the workflow can be carried out by a non-technical community member in an authentic setting.

2 The Dissolution of Worldview in Current Pipelines

The dissolution of worldview-level meaning happens cumulatively, across three transitions, each introducing a compression that the next stage cannot recover.

From interaction to transcription: spoken language carries context, prosody, timing, gesture, silence, and relational positioning. When interaction becomes text, these channels are largely discarded. Culturally meaningful pause lengths, intonation contours, evidentiality, and bodily conduct either disappear or require specialist annotation that most low-resource pipelines cannot sustain.

From transcription to translation: the target language (most often English) lacks direct equivalents for many source-language distinctions. Evidential morphology may be rendered as a neutral assertion, echoing Jakobson’s (1959) observation that translation entails loss of grammatically obligatory source-language distinctions. A Mayan greeting encoding a path-based model of life (*Bix a beel*: how-your-path) may be glossed simply as “Hello.” Conceptual architecture is not preserved.

From translation to annotation: standard frameworks capture structural linguistic information or task labels, not worldview-level distinctions. An additional epistemic filter operates — for example, Maya field linguists have reported that greeting expressions encoding life as trajectory are often dismissed as romanticization, reducing systematic worldview patterns to noise.

The cumulative result is ontological compression: languages are digitized in ways that preserve surface structure but not the conceptual architectures generating meaning. A language may achieve formal digital survival while the conceptual systems it encodes remain inaccessible to computational tools.

3 Background and Related Work

3.1 Low-Resource Language Technology

Low-resource NLP has focused primarily on overcoming data scarcity for tasks such as machine translation, speech recognition, and corpus construction (Joshi et al., 2020), including documented challenges specific to indigenous languages of the Americas (Mager et al., 2018). Prior surveys demonstrate that global NLP remains concentrated on a small number of high-resource languages. Pava et al. (2025) document persistent asymmetries in evaluation and model coverage in low-resource language contexts.

van Rijn et al. (2026) demonstrate the scalability of automatically generated vocabulary-based proficiency assessments across a large number of languages. However, scalable language-level tools do not by themselves represent culturally specific conceptual distinctions: a model may process a language fluently while systematically underrepresenting the morphological, evidential, or relational dimensions through which it encodes meaning.

LLMs trained primarily on dominant linguistic and cultural contexts reproduce those structures when interpreting new material — reshaping it to fit familiar patterns. For under-documented cultures, this is particularly problematic: invented outputs can circulate as authentic representations in the absence of accessible reference corpora.

This work assumes languages encode distinctions not fully recoverable through translation alone, consistent with work on linguistic relativity, culturally structured meaning, and pragmatic interpretation (Lucy, 1992; Levinson, 1983; Wierzbicka, 1997), without strong claims about cognitive determinism. The key point is that these distinctions risk being lost in current documentation pipelines.

Our approach asks how meaning is structured before becoming data, capturing worldview patterns explicitly so they are not lost prior to annotation — shifting the task from labeling content to modeling how different systems organize meaning.

3.2 Annotation Frameworks and Related Work

Established annotation frameworks represent phonetics, morphology, syntax, discourse structure, and semantic roles. Architectures such as annotation graphs (Bird and Liberman, 2001) and the ISO Linguistic Annotation Framework (Ide and Romary, 2004) were developed to support interoper-

able representation across linguistic annotations and datasets. The issue is not that annotation is absent, but that most existing frameworks were designed to capture structural linguistic information rather than worldview-level distinctions. Table 1 summarizes how the proposed framework differs from other annotation approaches.

Existing semantic representation frameworks and lexical-semantic models—including FrameNet (Baker et al., 1998), AMR (Banarescu et al., 2013), UCCA (Abend and Rappoport, 2013), and OntoLex-Lemon (W3C Ontology-Lexica Community Group, 2016)—capture complementary aspects of lexical and utterance meaning, such as frames, concepts, participants, semantic relations, and links between lexical entries and ontological entities. WAS operates at a different dimension. Rather than modeling only the meaning expressed by an utterance, WAS represents evidence-grounded interpretations of the worldview patterns made salient by the utterance and its documented contexts of use. A FrameNet annotation would represent the event structure and participant roles evoked by the proverb. A WAS annotation instead represents its epistemic structure (signal $\neq$ outcome), ecological grounding (predator behavior), social function (calming fear in a hierarchical context), and boundary condition (inappropriate when harm has already occurred). These dimensions are not an extension of FrameNet. They operate at a different level of description, representing the worldview that semantic roles presuppose rather than the semantic roles themselves.

3.3 LLM-Based Cultural Annotation

Recent LLMs have enabled scalable annotation of narratives, values, and cultural discourse patterns. However, most such systems operate on observable outputs rather than the conceptual structures generating them. The standard pipeline is: Text $\rightarrow$ LLM $\rightarrow$ Annotation. This treats text and its meaning as already transparent. If culturally distinct meaning systems are interpreted only through dominant-language categories, the result may be efficient but reductive. WAS is designed to address this risk.

AlKhamissi et al. (2025) find that current cultural benchmarks reduce culture to static facts or homogeneous values, arguing that anthropological methods are largely absent from benchmark design. Khudayberganova et al. (2025) demon-

Approach	Preserves raw form	Captures text of use	con-	Controlled worldview tags	Structured output for downstream use	Annotator role in worldview capture
Linguistic glossing	Yes	Limited		No	Variable	No structured worldview role
Metadata archiving	Partial	Partial		Limited	Yes — metadata	Consultative; not a structured annotator
FrameNet	No	Limited — frame-evoking context, not community usage		No	Yes	Expert lexicographer/annotator; no community role
AMR	No	No — sentence-level meaning graph		No	Yes	Trained annotator; no community role
UCCA	No	Limited		No	Yes	Trained annotator; no community role
OntoLex	No	Limited		Limited — lexical-semantic relations, not worldview	Yes	Lexicographer/ontology engineer; no community role
This framework	Yes	Yes — usage event required		Yes — Tier-1 tagset	Yes — graph-compatible	First annotator: evidence precedes and constrains curator interpretation

Table 1: Comparison of annotation approaches.

strate community-grounded ontology annotation for Uzbek folklore preservation; their system begins with transcribed text, while WAS targets the pre-textual documentation stage. [Martinez Pandiani and Presutti \(2024\)](#) formalize annotation context through SituAnnotate, explaining how labels are produced but not how culturally specific conceptual systems shape what can be expressed. WAS targets this internal layer, encoding schema-level distinctions — personhood, time, agency, morality — as structured dimensions of meaning.

4 The Worldview Annotation Schema

In the version presented here, the Worldview Annotation Schema is a three-layer framework for capturing culturally grounded meaning before downstream AI analysis normalizes it. Its central claim is that language is a constrained evidence layer for worldview, not the worldview structure itself. The framework is intended as a practical first step that extends existing documentation practice. This approach is grounded in the sociology of culture and cognition: [Cerulo et al. \(2021\)](#) review how culture operates through cognitive processes, in-

cluding structured patterns of acquisition, storage, and retrieval. Building on this perspective, we argue that annotation limited to propositional content may miss aspects of the schema-level organization through which culturally grounded meaning is interpreted. [Alwajih et al. \(2025\)](#) demonstrate the value of culturally specific datasets and evaluation benchmarks, including attention to nuanced cultural variation that broad labels may obscure. We believe that non-equivalence should be treated as informative rather than as annotation failure — directly supporting the non-equivalence tolerance requirement built into the framework’s design. The Ukrainian diminutive asymmetry introduced in Section 1 illustrates why this matters even in a well-resourced language; the Maa case study below shows the same compression mechanism operating in a language with substantially fewer digital resources than Ukrainian, making such losses more difficult to detect and recover after the fact.

WAS is implemented through a two-tier workflow: a community-facing documentator layer and a technical curator layer, detailed below.

4.1 The Initial Three-Layer Architecture

Layer 1 — Raw Input: the documentator records what speakers say, exactly as spoken. No interpretation, no simplification, audio recording preferred, variants captured. The raw layer is immutable: nothing recorded here is discarded or corrected at later stages. This layer ensures linguistic survival regardless of how tags evolve.

Layer 2 — Analysis: divided into three sequential components. Layer 2A (Conceptual Extraction) identifies the domain, conceptual axes, and candidate tags — detecting where conceptual systems diverge from dominant-language defaults. Layer 2B (Language/Form Analysis) uses lexical, metaphorical, grammatical, and discourse-level signals as evidence for the worldview reading. Layer 2C (Evidence Integration) combines 2A and 2B outputs, checks for recurrence, validates tags against recorded evidence, and assigns a confidence level. Without 2C, annotation collapses into subjective labeling.

Layer 3 — Structured Output: clean, standardized representation encoding the final cluster, axes, tags, and confidence level. No interpretation happens here — only encoding. Output is graph-compatible, machine-readable.

4.2 The Two-Tier Workflow

The documentator is a community member or language worker with no required technical background. Using a structured field sheet, the documentator: (1) records the exact form with audio and variants; (2) collects one concrete remembered story of real use; (3) identifies 1–2 primary functions from a fixed list; (4) identifies the emotion targeted; (5) completes a metaphor map; (6) records boundary conditions. The documentator does not invent tag names.

The curator is a trained linguist or senior community worker who converts documentary evidence into structured annotation: normalizing form, creating usage event objects, extracting metaphor frames, encoding worldview assertions, and recording boundary conditions as explicit graph relations. The key constraint: all worldview assertions must be traceable to recorded documentary evidence. This traceability requirement ensures that curator interpretations remain constrained by community-provided evidence and open to community review.

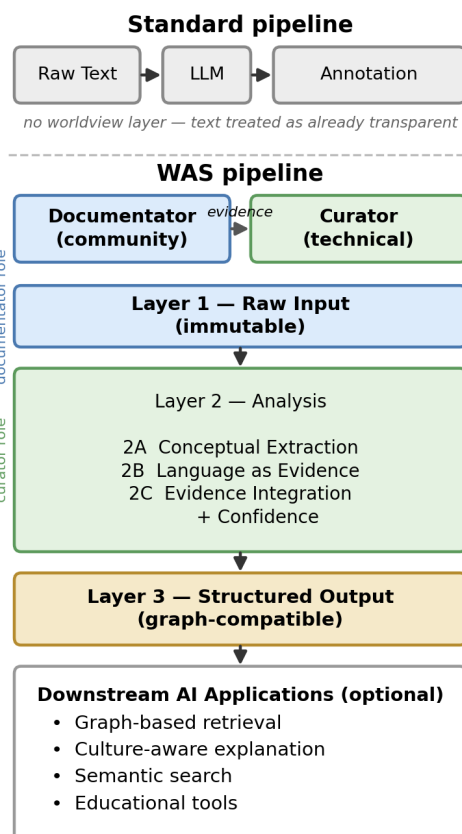


Figure 1: Standard pipeline versus the WAS pipeline. The graph-compatible structured output is designed to support future AI retrieval and comparison of worldview concepts alongside lexical and propositional content

4.3 Tag Governance

The governance model addresses the central tension: tags may not grow freely without losing interoperability, but cannot be locked too tightly without erasing worldview nuance. The solution is layered governance: field freedom in description, central discipline in vocabulary.

The Tier-1 tagset covers function (CalmFear, Warn, Correct, Encourage, AssertAuthority, ReduceConflict, TeachRespect, InduceShame, MotivatePatience, Humor), emotion (Fear, Anger, Shame, Pride, Grief, Envy, Impatience), domain (Conflict, Family, Leadership, Livestock, Marriage, Ritual, Community, Survival, YouthTraining), and epistemic hint (AppearancesMislead, WaitForEvidence, ActionsMatterMoreThanWords, TimeRevealsTruth, RespectHierarchy, CommunityBeforeSelf). New tags enter only after appearing in three independent cases, with a formal definition and two usage examples — mirroring controlled vocabulary evolution in FrameNet and WordNet.

Field	Content
Proverb ID	MAA-PROV-001
Language / Dialect	Maa (Eastern Africa)
Exact text	Me metan’ga olowuaru engutuk nijo kitoijoyie
Literal gloss	Me = not that Metan’ga = opens Olowuaru = a lion Engutuk = the mouth Nijo = you say Kitoijoyie = it has swallowed you
Free translation	It does not mean that when a lion opens its mouth that it has eaten or swallowed you
Variants	Single-speaker record (Alois Thomas Porokwa); cross-speaker variants not yet collected
Audio	Recorded with speaker consent; available upon request, subject to community approval.

Table 2: Layer 1 raw input record for MAA-PROV-001.

The system distinguishes Closed Mode (highly under-resourced languages: restricted to usage events and Tier-1 tags) and Open Mode (once recurring patterns are available: full Layer 2 pipeline with deeper conceptual extraction and vocabulary expansion).

5 Worked Example: A Maa Proverb

We illustrate WAS through a fully worked annotation of a proverb from Maa, a Nilotic language spoken in Eastern Africa: *Me metan’ga olowuaru engutuk nijo kitoijoyie* — “It does not mean that when a lion opens its mouth that it has swallowed you.” This example illustrates ecological grounding in local predator behavior, an epistemic structure distinguishing signal from outcome, and a socially situated function often omitted in summary translation.

5.1 Layer 1 — Raw Input

Table 2 presents the Layer 1 raw input record for MAA-PROV-001, documenting the proverb’s exact text, literal gloss, free translation, variants, and audio provenance exactly as recorded by the documentator, prior to any interpretation or analysis.

5.2 Layer 2A — Conceptual Extraction

Two independent usage events (Table 3) — a medical crisis and a livelihood crisis — show a speaker with cultural authority reorienting a younger person from despair toward action, suggesting the proverb encodes a portable cognitive structure for managing fear amid unresolved threat. This structure rests on three dimensions: an epistemic heuristic (wait

for evidence rather than inferring outcome from signal), a signal-outcome distinction (threat display vs. realized harm), and ecological grounding in predator behavior.

5.3 Layer 2B — Language as Evidence

Lexical and grammatical evidence: *olowuaru* (“lion”) evokes a culturally important predator; *engutuk* (“mouth”) foregrounds the display organ; *kitoijoyie* (“swallowed/eaten”) encodes irreversible harm as the feared outcome. The negation structure (*me metan’ga*, “it does not mean that. . .”) explicitly separates a perceived signal from its inferred consequence, a distinction confirmed by a substitution test: replacing “lion” with a generic threat weakens the display-versus-harm contrast, suggesting the proverb relies on shared knowledge of predator behavior rather than on danger alone.

5.4 Layer 2C — Evidence Integration

The lexical, grammatical, and contextual evidence converges on the same interpretation. Grammatically, the proverb distinguishes a threat signal from realized harm through explicit negation, while lexically it contrasts opening the mouth with swallowing, reinforcing the distinction between warning and outcome. Shared knowledge of predator behavior provides the ecological grounding, and recurrent usage suggests a social function of regulating fear under uncertainty. Confidence in the item-level interpretation is assessed as high based on convergence across evidence types and recurrence across the two documented usage events; however, data sufficiency for broader generalization remains limited.

5.5 Layer 3 — Structured Output

Example RDF graph relations:

```
MAA-PROV-001 hasFunction CalmFear;
MAA-PROV-001 hasEpistemicHeuristic Wait-
ForEvidence;
MAA-PROV-001 hasMetaphorSource Preda-
torThreatDisplay;
MAA-PROV-001 hasMetaphorTarget ThreatSit-
uation;
MAA-PROV-001 inappropriateWhen Actual-
HarmOccurred.
```

The Layer 3 representation supports graph storage and queries such as retrieving expressions that distinguish a threat signal from its outcome.

5.6 Pilot Field Application

To illustrate the practical application of the proposed workflow in an authentic documentation set-

Usage Field	Event	UE-001	UE-002
Usage Event ID		MAA-PROV-001-UE-001	MAA-PROV-001-UE-002
Speaker role		Maa community member (adult, cultural authority)	Elder (family leader, cultural authority)
Addressee role		Non-Maa interlocutor (younger, outside community)	Younger community member (adult male)
Context type		Medical crisis — serious illness, acute uncertainty	Livelihood crisis — near-total livestock loss due to drought
Trigger		Addressee expressing fear and distress	Addressee in hopelessness; considering abandoning pastoral livelihood
Speech act		Direct comfort; reassurance through ecological-epistemic reframing	Encouragement; reframing of crisis as non-terminal
Emotional climate		High anxiety; fear of irreversible harm	Despair, grief, fear of permanent loss
Outcome		Addressee reported proverb as meaningfully calming	Addressee reported renewed courage; rebuilt livestock holdings

Table 3: Usage Event records for MAA-PROV-001. Two independent usage contexts illustrate portability across threat types. UE-002 reported by Alois Thomas Porokwa, 28 April 2026.

ting, we conducted a pilot field application with Alois Thomas Porokwa, a native Maa speaker and community documentator. The pilot demonstrates the workflow end-to-end—from documentator-captured raw input through curator-produced Layer 3 structured output—for a single illustrative item.

The documentator independently completed the Layer 1 record for MAA-PROV-001 and assigned the Tier-1 tags Encourage and MotivatePatience (function), Fear (emotion), Survival (domain), and AppearancesMislead (epistemic hint), without reference to the curator’s analysis. These annotations were then compared with the curator’s independently produced annotation of the same raw evidence: CalmFear (function), Fear (emotion), Conflict/Leadership (domain), and WaitForEvidence (epistemic hint).

The two selections fully agree on emotion, partially agree on epistemic hint—representing adjacent framings of the same underlying logic—and diverge on function and domain in ways that appear context-sensitive rather than erroneous. The documentator’s tags foreground the encouragement given to a specific addressee, while the curator’s tags foreground the social-hierarchical frame of the exchange.

We report this comparison to illustrate feasibility in practice, not as an inter-annotator agreement study. A single comparison point cannot support a reliability statistic such as Cohen’s κ ; a formal agreement study involving multiple documentators remains future work. The comparison nonetheless provides a preliminary signal that a non-technical documentator using only the field cheat sheet can independently produce tags that meaningfully align

with expert curation, which is the minimum bar required by the proposed two-tier workflow.

6 Implications for AI and Low-Resource NLP

If culturally meaningful distinctions are systematically absent from training data, AI systems may perform well on surface tasks while remaining shallow in cross-cultural settings. Structured Layer 3 outputs are designed to support the downstream applications described below, complementing the scalable vocabulary-based language assessment demonstrated by [van Rijn et al. \(2026\)](#) by extending infrastructure from language coverage toward worldview coverage.

[AlKhamissi et al. \(2025\)](#) show that current cultural benchmarks often reduce culture to static facts or survey-derived values. WAS-based evaluation sets could instead test whether systems retain features such as evidential marking, relational morphology, or ecologically grounded metaphor — dimensions not explicitly evaluated in existing benchmarks.

Operationally, a Layer 3 record can serve as structured contextual knowledge for downstream AI applications. Rather than relying only on the literal wording of an expression, an AI system can use a Layer 3 record to explain the worldview it encodes. In the single worked example presented here, an assistant asked to explain the Maa proverb could present not only the literal translation (“a lion opening its mouth does not mean it has caught you”) but also an explanation of its documented epistemic stance (WaitForEvidence), the observation that appearances may be misleading

Field	Value
CLUSTER	Knowledge & Truth
AXES	Signal vs. Outcome; Appearance vs. Reality
FUNCTION (Tier-1)	CalmFear
EMOTION (Tier-1)	Fear → reduced / redirected
DOMAIN (Tier-1)	Survival, Conflict / Leadership
EPISTEMIC HINT (Tier-1)	WaitForEvidence; AppearancesMislead
METAPHOR MAPPING: SourceFrame	PredatorThreatDisplay (Agent: Lion; Action: OpenMouth; PotentialOutcome: Swallow)
METAPHOR MAPPING: TargetFrame	ThreatSituation (Agent: PowerfulPerson or ExternalThreat; Action: ThreatDisplay; PotentialOutcome: Harm)
Derived Assertion: EncodesNorm	DoNotInferHarmFromThreatSignal
EncodesEpistemic-Heuristic	WaitForEvidence
EcologicalAnchor	Savannah predator behavior (lion threat display)
ReinforcesHierarchy	Contextual — acknowledges power differential while moderating fear response
BoundaryCondition	Inappropriate when actual harm has already occurred; may be used sarcastically to minimize real danger (documentator flag required)
CONFIDENCE	High (based on convergence across independent evidence types; see §5.6 documentator–curator comparison)

Table 4: Layer 3 structured output for MAA-PROV-001. All fields are traceable to documentary evidence recorded at Layers 1 and 2.

(AppearancesMislead), its ecological grounding, and its conditions of use. These concepts are explicitly represented in the Layer 3 record rather than remaining implicit in free-text explanation. The same structured representation also supports graph-based retrieval and provides a reference for detecting cases where distinct worldview concepts have been collapsed into an apparent equivalent during translation or AI processing. For example, a query such as “find expressions that distinguish a threat signal from its outcome” would return MAA-PROV-001 because these worldview concepts are encoded as searchable fields rather than embedded only in prose. This example illustrates the type of AI-supported explanation and retrieval enabled by WAS on a single annotated item. As larger collections of WAS-annotated records become available, the same representation could support retrieval and comparison of analogous worldview structures across languages. Building and evaluating such a multilingual corpus remains future work.

Regarding scalability: a single proverb annotation took about 15 minutes with no technical training, and the framework fits existing documentation practice, with trained curators able to support multiple community projects remotely. Deployment must respect restricted or ceremonial knowledge through informed consent.

7 Limitations

The framework is illustrated through one worked example; broader validation across typologically diverse languages remains future work. The documentator–curator comparison reported in the pilot field application (§5.6) involves a single community documentator rather than multiple independent annotators; a full inter-annotator agreement study with several documentators per language remains future work. The workflow assumes access to a trained curator, which may limit applicability in severely under-resourced settings. The Tier-1 tagset is based on a limited empirical base and may require expansion. The governance model has not yet been tested in situations where it could be misused. The architecture presented here is an initial conceptual model and may be refined as its provenance relations, and governance procedures are tested across additional languages and documentation settings.

8 Conclusion

Low-resource language technology increasingly preserves speech, text, and computational access, but linguistic form alone does not capture how speakers represent evidence, personhood, agency, time, and moral responsibility. These distinctions can be reduced across transcription, translation, and annotation.

This paper introduced an initial three-layer annotation framework with a two-tier community/curator workflow, illustrated through a worked Maa proverb. Non-technical community members act as first annotators, providing the evidence from which curators build controlled, structured representations.

As AI systems increasingly learn from recorded language data, the question is not only which languages are present, but which meaning-encoding features remain representable. WAS supports collaborative annotation of alternative interpretations without forced equivalence or cultural collapse.

References

- Omri Abend and Ari Rappoport. 2013. [Universal Conceptual Cognitive Annotation \(UCCA\)](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.
- Mai AlKhamissi, Yunze Xiao, Badr AlKhamissi, and Mona Diab. 2025. [Hire your anthropologist! rethinking culture benchmarks through an anthropological lens](#).
- Fakhraddin Alwajih et al. 2025. [PEARL: A multimodal culturally-aware Arabic instruction dataset](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23048–23079, Suzhou, China. Association for Computational Linguistics.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. [The Berkeley FrameNet project](#). In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. [Abstract Meaning Representation for sembanking](#). In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Steven Bird and Mark Liberman. 2001. [A formal framework for linguistic annotation](#). *Speech Communication*, 33(1–2):23–60.
- Karen A. Cerulo, Vanina Leschziner, and Hana Shepherd. 2021. [Rethinking culture and cognition](#). *Annual Review of Sociology*, 47:63–85.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. [Glottolog Database 4.3 as CLDF](#).
- Nancy Ide and Laurent Romary. 2004. [International standard for a linguistic annotation framework](#). *Natural Language Engineering*, 10(3–4):211–225.
- Roman Jakobson. 1959. [On linguistic aspects of translation](#). In Reuben A. Brower, editor, *On Translation*, pages 232–239. Harvard University Press, Cambridge, Massachusetts.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Daniel Jurafsky. 1996. [Universal tendencies in the semantics of the diminutive](#). *Language*, 72(3):533–578.
- Durdona Khudayberganova, Nilufar Tosheva, Hafiza Aslanova, Olim Kayumov, Gulnoz Khallieva, Otabek Mukhitdinov, I. B. Sapaev, and Dilfuza Tursunova. 2025. [Hybrid ontology and NLP-based annotation model for indigenous lexicon and folklore preservation](#). *Indian Journal of Information Sources and Services*, 15(2):333–339.
- Stephen C. Levinson. 1983. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.
- John A. Lucy. 1992. *Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis*. Number 12 in Studies in the Social and Cultural Foundations of Language. Cambridge University Press, Cambridge.
- Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. [Challenges of language technologies for the indigenous languages of the Americas](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Delfina Sol Martinez Pandiani and Valentina Presutti. 2024. [Situating ground truths: Enhancing bias-aware AI by situating data labels with SituAnnotate](#).
- Juan N. Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T. Truong, Daniel Zhang, Elena Cryst, Vukosi Marivate, and Sanmi Koyejo. 2025. [Mind the \(language\) gap: Mapping the challenges of LLM development in low-resource language contexts](#). White paper, The Asia Foundation and Stanford Institute for Human-Centered Artificial Intelligence.
- Pol van Rijn, Yue Sun, Harin Lee, Raja Marjeh, Ilia Sucholutsky, Francesca Lanzarini, Elisabeth André, and Nori Jacoby. 2026. [One test, many tongues: Surveying language proficiency across the globe](#). *Proceedings of the National Academy of Sciences*, 123(13):e2420179123.
- UNESCO. 2022. [International decade of indigenous languages 2022–2032](#).
- W3C Ontology-Lexica Community Group. 2016. [Lexicon model for ontologies](#). Final community group report, World Wide Web Consortium.
- Anna Wierzbicka. 1997. *Understanding Cultures through Their Key Words: English, Russian, Polish, German, and Japanese*, volume 8 of *Oxford Studies in Anthropological Linguistics*. Oxford University Press, New York.