Beyond ELAN

Toward Open, Modular Infrastructure for Multimodal Annotation

Babajide Owoyele

June 2026 | *Preprint*

ELAN has been the de facto standard for multimodal language annotation for over two decades. It was designed, however, for a fundamentally different research context: small corpora, entirely manual workflows, desktop-local data, and no expectation of cross-institutional sharing. Today, the field operates at a scale and with requirements—large video corpora, AI-assisted analysis, GDPR compliance, real-time collaboration—that ELAN's monolithic architecture cannot accommodate. We survey the current annotation tool landscape, identify three structural gaps, and argue that the path forward is not a new ELAN but an *open annotation stack*: a community-governed, modular infrastructure in which privacy, AI integration, and interoperability are first-class design primitives.

Introduction

Multimodal language research—spanning gesture studies, sign language linguistics, conversation analysis, and child language acquisition—has long depended on video as its primary data source. For most of the past two decades, annotating that video has meant one thing: ELAN.

Developed at the Max Planck Institute for Psycholinguistics in the late 1990s, ELAN gave researchers a structured way to time-align annotations across multiple tiers (Wittenburg et al. 2006). It became indispensable not because it was the best imaginable tool but because it was the first tool good enough to become shared infrastructure. Researchers built workflows around it, trained graduate students on it, and deposited EAF files in discipline archives. ELAN is, in the deepest sense, path-dependent. [1]

[1] The ELAN Annotation Format (EAF) is an XML schema that encodes tiers, time slots, and media linkages. Its stability as a format—the same files round-trip across two decades of ELAN versions—has been a genuine asset. It is the format's surrounding tool ecosystem, not the format itself, that has stagnated.

The problem is that the world around ELAN has changed substantially, and ELAN has not. Three transformations in particular have opened a gap between what the tool provides and what researchers now need:

1. **Scale.** Corpora that once comprised dozens of hours of video now routinely comprise thousands. Manual annotation at that scale is economically and temporally untenable without AI assistance.
2. **AI.** Automated transcription, body-pose estimation, and gesture segmentation have matured from research prototypes to deployable components. ELAN provides no interface for them.

3. **Privacy and regulation.** GDPR and its equivalents have made the unprocessed sharing of research video legally fraught. Researchers either forgo sharing entirely or assemble ad-hoc anonymization scripts upstream of ELAN. Privacy remains structurally outside the tool.

These are not feature requests. They are structural mismatches between a tool's design assumptions and the field's current reality. Incremental patches to ELAN will not close them.

The Annotation Landscape: Three Structural Gaps

Closed, Monolithic Architecture

ELAN is desktop software. It reads and writes EAF—an XML schema with no query API, no streaming interface, and no mechanism for external processes to read or modify annotation state during a session. Integrations are possible only by reading and writing EAF files on disk, a pattern that treats ELAN as a file format rather than a platform. [2]

Modern research pipelines are compositional. A researcher might use Whisper for speech transcription, ViTPose for skeletal tracking, a custom segmentation model for gesture boundaries, and ELAN for manual correction. Each step should produce outputs that feed cleanly into the next. EAF-on-disk integration makes this possible in principle but brittle in practice: format changes break parsers, file locking causes conflicts, and there is no shared notion of provenance. The tool that standardized annotation practice has standardized a *format* while leaving the surrounding workflow fragmented.

This is not merely a developer inconvenience. It means that the choices made by one lab—which parser, which conversion script, which naming convention—cannot be shared with another. Every lab that wants to add an AI component to its ELAN workflow must reinvent the integration from scratch. The cumulative cost to the field is substantial.

No AI Integration Path

The past five years have produced a set of AI capabilities directly relevant to multimodal annotation. Whisper (Radford et al. 2023) transcribes speech across dozens of languages with accuracy rivaling human transcription on clean audio. ViTPose (Xu et al. 2022) and its successors estimate full-body pose from video at real-time speed. Gesture segmentation models, gaze estimators, and emotion classifiers now exist as open-source releases. These tools do not eliminate the need for annotation; they change its character.

Human annotators are no longer asked to produce first-pass transcriptions from silence, but to correct, adjudicate, and interpret outputs that machines produce cheaply. The bottleneck shifts

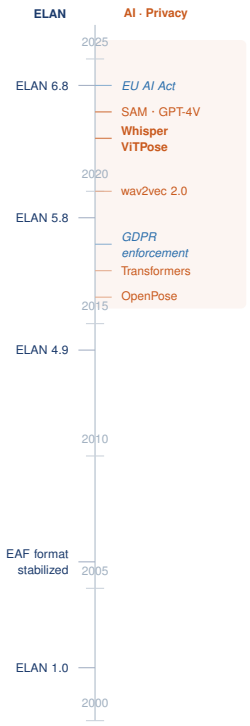

Figure 1: ELAN releases (left) have been sparse and incremental. AI capabilities (orange) and privacy regulation (blue) have accelerated sharply since 2016. The divergence is structural, not incidental.

[2] A small ecosystem of conversion utilities exists—`pympi`, `elan-parser`, various lab-internal scripts—but each is maintained by a different group, none share a common abstraction layer, and all are fragile to ELAN version changes.