Annotation Tools for Language Documentation: A Survey of Capabilities, Gaps, and Morphological Support

Changbing Yang[1], PT Anderson[2], Godfred Agyapong[3], Sarah Moeller[3]
[1]University of British Columbia
[2]Revitalization Technology [3]University of Florida
cyang33@mail.ubc.ca, smoeller@ufl.edu

Abstract

Annotation tools are foundational infrastructure for language documentation, yet few comprehensive surveys have evaluated the tool landscape specifically from a documentary linguistics perspective. We survey 98 annotation tools across dimensions critical to language documentation workflows: annotation support, collaboration features, active learning, cost and openness, and institutional sustainability. Of the 44 tools both free and accessible for evaluation, only 15 support morpheme segmentation and glossing, and only 6 combine morphological annotation with remote collaboration at no cost. We identify a structural gap between the current tools and the requirements of field linguists working with endangered and Indigenous languages. While many NLP tools prioritize scalable annotation for high-resource settings, documentary linguists need interlinear glossed text (IGT) support and community-accessible interfaces. We taxonomise the tool landscape, present a multi-dimensional feature matrix, suggest current tools for language documentation, and conclude with concrete recommendations for tool developers and the documentary linguistics community.

1 Introduction

Language documentation is an urgent scholarly and humanitarian endeavor. Of the roughly 7,000 languages spoken today, a substantial proportion are endangered (Krauss, 1992; Eberhard et al., 2026), making the creation of annotated linguistic records essential not only for scientific research, but also for community-based language maintenance and revitalization (Himmelmann, 1998; Woodbury, 2003). A central part of documenting those endangered languages involves enriching texts with detailed linguistic annotations. This is a multi-step process involving: 1. phonetic and orthographic transcription, 2. translation into a high-resource language like English, 3. morpheme segmentation and glossing, and 4. other grammatical annotation. Traditionally, these tasks have been carried out manually, a process that is thorough but extremely labor intensive. To reduce this burden, linguists often use specialized annotation software tools which are few in number. ELAN (Auer et al., 2010) and FLEx (Rogers, 2010) are widely used for annotation tasks such as time-aligned transcription, free translation, and morpheme analysis, have significant drawbacks: Both ELAN and FLEx were originally developed in the 1990s and reflect an earlier technological era. These tools were developed before the recent rise of text-based machine learning and generative AI, and they are not always well aligned with modern AI-assisted annotation workflows. In particular, they offer limited support for machine-in-the-loop interaction, where model predictions can be incorporated into annotation and iteratively corrected by human users to reduce manual effort. Substantially updating such legacy platforms can also be difficult and costly.

At the same time, the growing importance of annotation in both linguistic documentation and NLP has led to the development of many newer annotation tools. The sheer number and diversity of available tools makes it difficult for linguists and community language workers to determine which ones are actually suitable for documentation tasks. Despite this, such tools have received comparatively little systematic evaluation from the perspective of language documentation and basic linguistic analysis.

Our work presents a evaluation of language annotation tools, aiming to provide insights that help linguists and community language workers make informed choices and strengthen connections between their efforts and helpful AI. We design rubrics and use them to systematically evaluate 98 annotation tools, focusing on their functional capabilities for linguistic analysis, sustainability, and graphical interface design. We established a list

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 80–92

of criteria broken into specific questions to guide the evaluation, including active learning support[1], since such support may help reduce annotation effort in low-resource documentation workflows. We do not include speech-oriented transcription tools in our evaluation because they address a substantially different stage of the documentation workflow and require different criteria than text annotation, such as audio handling, time alignment, and speech recognition performance. Our focus is on tools for text-centered annotation tasks, particularly those that may support AI-assisted workflows for translation, segmentation, glossing, and grammatical analysis. A central question driving our research is whether tools designed for NLP are adaptable for endangered language documentation. Our contributions are:

- A multi-dimensional feature evaluation rubrics with particular attention to morphological annotation support (one of the most consistently underserved requirements for language documentation by NLP).
- A quantified gap analysis revealing that only 15 accessible tools support morpheme segmentation, and only 1 combine morphology with active learning.
- Suggestions for the development of annotation tools targeting the documentary linguistics community.
- A curated recommendation of the currently available free tools for different types of need of linguists. For example, we list tools supporting morpheme segmentation, with active learning noted as a bonus criterion.

2 Background and Related Work

Here we more fully describe the text annotation tasks of basic linguistic analysis. Then we compare our work to similar surveys, noting our contribution from the language documentation perspective.

2.1 Language Documentation and Annotation Needs

Language documentation involves the creation of a comprehensive, multi-layered record of a language, including audio and video recordings, transcriptions, translations, and morphological analyses (Himmelmann, 1998; Bird and Simons, 2003). A central output format is the interlinear glossed text (IGT), in which each word and morpheme is annotated with its grammatical gloss. A Gitksan (ISO 639-3 `git`) example is shown below:

Orthography:	Ii hahla'lsdi'y goohl IBM
Segmentation:	ii hahla'lst-'y goo-hl IBM
Gloss:	CCNJ work-1SG.II LOC-CN IBM
Translation:	And I worked for IBM.

Endangered language documentation introduces constraints that separate it from the mainstream NLP annotation efforts: limited annotator pools (often community members rather than trained linguists), non-standardized orthographies, polysynthetic or highly agglutinative morphological systems (contrasted with simpler isolating or fusional systems among populous Indo-European and Sino-Tibetan languages), offline fieldwork contexts, and ethical obligations around data sovereignty and community ownership (Rice, 2011). These conditions place specific demands on annotation tools, which must minimally support sub-word segmentation, tier alignment, and flexible schema definition.

2.2 Prior Surveys of Annotation Tools

Several surveys have catalogued annotation tools (Neves and Ševa, 2021), but these studies largely focus on general NLP or corpus annotation settings and do not evaluate tools from the perspective of documentary linguistics. This gap has been noted from another direction in work on language documentation tools themselves. Thieberger (2009) argues that IGT requires specialized tooling and highlights the lack of modern, usable systems for creating well-formed, standardized, and reusable IGT. He further emphasizes a broader disconnect between the computational agendas of language technology research and the practical needs of field linguists. More recently, Gessler et al. (2025) show that the limited adoption of NLP in language documentation is not simply a matter of model quality, but also of software infrastructure: documentary linguists face substantial technical burdens, and existing language documentation software often does not integrate smoothly with NLP systems. Our survey addresses this gap by examining annotation tools through the lens of documentary practice, with particular attention to linguistic functionality, sustainability, interface design, and AI-assisted workflows.

[1]The ability of a tool to leverage partial model predictions to accelerate annotation

3 Methodology

Our approach to surveying 98 annotation tools focused on their functional capabilities, sustainability, and graphical interface design. Functional capabilities refers to features that support language documentation tasks, such as morpheme or subword segmentation, morpheme glossing, remote collaboration, interoperability, and NLP-assisted workflows. We evaluate the tools in seven key categories.

3.1 Tool Selection

We compiled an initial list of 98 tools by aggregating from four sources: (1) prior annotation tool surveys and reviews; (2) tools recommended in language documentation literature and community wikis; (3) websites recommending commercial NLP annotation platforms active as of 2025; and (4) tools cited in ComputEL and LREC proceedings. Tools were included regardless of scope (general NLP vs. linguistic annotation) or development status. A full list of our investigated tools is available through the Google spreadsheet[2] and Table 6.

3.2 Feature Schema

To select the ideal annotation tool, users must navigate diverse specifications and purposes. Although the work of NLP and linguistics overlap, they differ in the exact subtasks, workflows, and priorities, particularly in ways that align with linguistic or community-based goals. Not all NLP annotation tools support the tasks or data needed by academic or community linguists. We developed a 32-dimensional feature schema organized into seven categories. The details of all features are listed in Table 5.

1. Cost A tool's financial model plays a critical role in its adaptability, as academic or community users often have limited funding. On the other hand, paid tools may provide better technical support and longevity.

2. Sustainability and Longevity A tool's long-term viability depends on active maintenance and the nature of the entity maintaining it. Proprietary tools tend to have more consistent maintenance, but some open-source projects thrive thanks to dedicated developer communities.

[2] https://docs.google.com/spreadsheets/d/1o-IQTC7vIK1xRqdOoIzdgAlrsedkJeIswAeFyeiS93M/edit?usp=sharing

3. Portability Given the varied needs of linguistic projects, it is unlikely that a single tool will meet all needs. Therefore, data portability ensures seamless workflows across different apps. Data portability depends on export/import capabilities to commonly used data schemas that thoroughly represent the IGT data model.

4. User Friendliness When we consider the uneven technical expertise involved in language documentation, usability is critical. The installation process should not require advanced technical expertise. The graphical interface should allow users who are familiar with the tool's purpose to get started without needing detailed instructions.

5. Sensitivities Working with endangered languages involves unique ethical considerations related to privacy, access rights, and data ownership (Brinklow, 2021). Software specifications should be clear where uploaded data is stored (if not on the user's computer) and who has access to the data, and how that data may be used. These considerations are especially important when working with data collected from minority communities, where ethical and privacy standards may differ from those in commercial and some research settings.

6. Linguistic Annotation Capabilities Rather than focusing on the specific tasks (e.g. named entity recognition or dependency parsing) that a tool was designed for, we assess its capacity to adapt to basic documentary tasks such as word-by-word glossing, morpheme segmentation and glossing, as well as linguistic tasks that are more common in NLP such as part-of-speech (POS) tagging and translation.

Here, we differentiate morpheme segmentation support from IGT support: the former requires only sub-word annotation capability, while the latter additionally requires aligned interlinear tier display in the documentary linguistics format.

7. Active Learning Producing annotated data can be costly in terms of time and resources. There's a common goal in NLP and linguistics to minimize costs. One strategy that minimizes human labor and maximizes the utility of computer–annotated labels is Active Learning (AL), sometimes referred to as machine-in-the-loop in linguistics settings (Bird and Yibarbuk, 2024; Moeller and Arppe, 2024). In the AL paradigm, the machine learning model actively selects data points from which to learn, rather than being passively

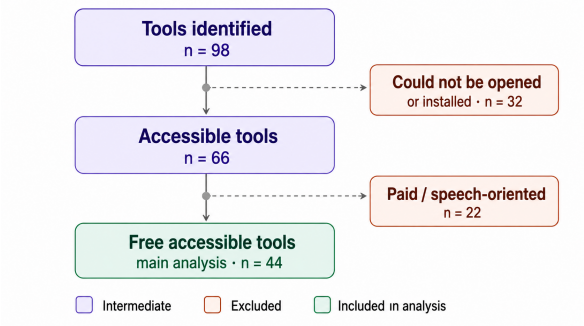

Figure 1: Flowchart of tool selection for the survey. Of the 98 annotation tools initially identified, tools that could not be opened or installed were excluded, and the main quantitative analysis was restricted to the subset of free accessible tools.

trained on a fixed dataset. This prioritizes the most informative examples for human annotation, reducing the overall cost and effort. In anticipation that AL can be integrated as AI assistance to language documentation, we assess whether a tool offers active learning functionalities. This includes whether users are provided with feedback on computer predictions and can add their own pre-annotations. For instance, does the tool allow the user to import annotations with confidence scores from a machine learning model or does it have functionalities itself to train and test models?

3.3 Coding Procedure

Of the 98 tools surveyed, 32 could not be accessed or installed, rendering them inaccessible to current practitioners. As mentioned earlier, our focus is on text annotation following transcription, so we did not include primarily speech or transcription tools. Our project budget did not allow for the evaluation of tools behind paywalls. Therefore, our quantitative analysis was necessarily limited to the 44 free tools. Given that many linguists and community partners face similar resource constraints, we consider this limitation to be consistent with real-world conditions and therefore not detrimental to the validity of our findings. The tool selection procedures are shown in Figure 1.

Our analysis is based primarily on official documentation, including user manuals, project websites, published papers, and other materials provided by the tool developers. We use these sources because they represent the most complete and authoritative descriptions of each tool's intended functionality, supported features, and design goals. However, this also means that our feature comparisons are based on self-reported information[3] rather than systematic installation and hands-on testing of every platform (due to time and budget limitation).

Feature values were coded as Yes/No/Partial or free text where applicable[4], with qualitative notes retained. Morpheme segmentation and glossing support was coded as *Yes* only where the tool provides explicit sub-word annotation tiers or morpheme-level labeling functionality (not merely word-level annotation), although open-source tools might allow adaptation to morpheme segmentation and glossing.

Feature coding was conducted in two stages by four annotators. All four coders have expertise in both language documentation and NLP. One annotator carried out the primary coding for each tool based on the tool's official documentation and, where necessary, direct inspection of the tool itself. The other three annotators reviewed the coding decisions and supporting notes. Any disagreements or unclear cases were discussed collectively until a consensus judgment was reached.

4 Taxonomy of Annotation Tools

We explore in more detail the 44 of the 98 tools that are freely accessible. We organize them into two functional categories based on primary purpose of design, institutional origin, and typical use case. Table 1 summarizes the taxonomy. While recognizing the constraints faced by many linguists, we are not encouraging them to only use free software. This survey of free tools can assist identifying which for-cost software should be explored further, within budget constraints.

NLP/industry tools (e.g., INCEpTION (Klie et al., 2018), Label Studio (Tkachenko et al., 2020), Doccano (Nakayama et al., 2018), Brat (Stenetorp et al., 2012), ALToolbox (Tsvigun et al., 2022)) account for 18 of the 44 free tools. They are predominantly designed for high-resource text annotation pipelines for tasks in high demand in NLP research or industry: named entity recognition and docu-

[3]We therefore acknowledge that some reported features may be incomplete, outdated, or no longer functional, especially for tools whose documentation or code has not been actively maintained. Future work should complement this documentation-based survey with deployment-based evaluation, including testing whether each tool can still be installed and used successfully in a contemporary computing environment.

[4]For example, when coding the cost feature, we find that Labelbox has a free tier, but advanced features and larger scale usage require a subscription.

Category	#	Examples	Use Case
NLP / Industry	18	INCEpTION (Klie et al., 2018), Label Studio (Tkachenko et al., 2020), Doccano (Nakayama et al., 2018), Brat (Stenetorp et al., 2012)	General NLP; entity, relation
Linguistic Corpus	26	EXMARaLDA (Schmidt and Wörner, 2014), TEITOK (Janssen, 2016), CATMA (Horstmann, 2020), MMAX2 (Müller and Strube, 2006), PACTE (Ménard and Barrière, 2017)	Corpus; morphology, syntax

Table 1: Taxonomy of the 44 free accessible annotation tools surveyed.

ment classification. Several offer active learning support and web-based collaboration, but none of the 18 free NLP tools support morpheme segmentation, reflecting their design focus on word- and span-level annotation for standard NLP tasks.

Linguistic corpus tools form the largest free category with 26 tools (e.g., EXMARaLDA (Schmidt and Wörner, 2014), TEITOK (Janssen, 2016), CATMA (Horstmann, 2020), MMAX2 (Müller and Strube, 2006), PACTE (Ménard and Barrière, 2017), UAM CorpusTool (O'Donnell, 2008)). Developed primarily in academic linguistics contexts, they offer richer morphological and syntactic annotation support: 14 of the 26 free corpus tools support morpheme segmentation, accounting for all but one of the free tools with this capability. However, collaboration infrastructure and active learning do not often co-occur, many being desktop-only applications with limited interoperability or portability.

5 Main Findings and Analysis

Here we summarizes the survey's findings, analyzing the gaps documentary linguists are likely to discover when adapting annotation tools that were designed for NLP.

5.1 No Single Tool Meets All Linguistic Needs

No one tool provides a comprehensive solution for all linguistic tasks, but many tools offer complementary functionalities to established linguistic software like ELAN (Auer et al., 2010) and FLEx (Rogers, 2010). Our work reduces the decision space by eliminating clearly irrelevant choices and inaccessible tools. The ideal choice depends on the user's priorities—whether cost, ease of use, sustainability, or AI support. Recognizing which criteria are essential for a given project will enable a more focused selection.

Table 2 presents a feature matrix for a representative selection of the 44 free tools, prioritising those with morphological annotation support alongside key NLP tools for comparison.

5.2 Criteria and Trade-offs

The value of each criteria depend on the user's needs. Each criterion introduces notable variability and often present trade-offs with another criteria, reinforcing the realization that the lack of a one-size-fits-all solution is partly due to the difficulty of addressing all needs sufficiently in one tool. We illustrate this with two specific examples.

Example 1: Cost, Sustainability and Longevity. Of these 44 free tools, 31 are also open-source. Open-source tools offer full functionality without licensing fees, making them convenient for projects with limited funding. Of the 44 free tools, 9 tools with morpheme support are confirmed as actively maintained: EXMARaLDA (Schmidt and Wörner, 2014), TEITOK (Janssen, 2016), CATMA (Horstmann, 2020), MMAX2 (Müller and Strube, 2006), Praaline (Christodoulides, 2018), SLATE (Kummerfeld, 2019), UAM CorpusTool (O'Donnell, 2008), WebLicht (Ljubešić et al., 2017), and Kratylos (Kaufman and Finkel, 2018). Several tools in this set show signs of abandonment or uncertain long-term maintenance. Dexter (Trani et al., 2014) and Emdros (Lowery, 2008) appear to be abandoned, while The Simple Corpus Tool (Weisser, 2016) and Lexonomy (Měchura and Rychlý, 2017) have unclear or stalled maintenance histories. Such uncertainty raises data preservation concerns for language documentation projects that depend on these platforms. In contrast, for-profit companies tend to offer consistent tool updates, ensuring long-term usability. Open-source tools depend on community involvement for maintenance, leading to variability in update frequency. However, some open-source tools such as INCEpTION (Klie et al., 2018) and Label Studio (Tkachenko et al., 2020) benefit from dedicated developer communities and see consistent improvements. In contrast, proprietary tools may risk discontinuation if a startup fails or subscriptions lapse, undermining long-term reliability.

Example 2: User-friendliness and linguistic customisability. Ease of use is critical for adaptability, particularly for non-technical users, but

Tool	Open	Free	Morph	IGT	Collab.	AL	Maint.
EXMARaLDA (Schmidt and Wörner, 2014)	✓	✓	✓	∼	✓	×	✓
TEITOK (Janssen, 2016)	✓	✓	✓	∼	✓	×	✓
CATMA (Horstmann, 2020)	✓	✓	✓	×	✓	×	✓
Interlinear Text Ed. (Hughes et al., 2004)	✓	✓	✓	✓	×	×	✓
Lexonomy (Měchura and Rychlý, 2017)	✓	✓	✓	×	✓	×	×
Praaline (Christodoulides, 2018)	✓	✓	✓	∼	×	×	✓
MMAX2 (Müller and Strube, 2006)	✓	✓	✓	×	×	×	✓
SLATE (Kummerfeld, 2019)	✓	✓	✓	×	×	×	✓
Emdros (Lowery, 2008)	✓	✓	✓	×	×	×	×
PACTE (Ménard and Barrière, 2017)	×	✓	✓	×	✓	✓	✓
INCEpTION (Klie et al., 2018)	✓	✓	×	×	✓	✓	✓
Label Studio (Tkachenko et al., 2020)	✓	✓	×	×	✓	✓	✓
ALToolbox (Tsvigun et al., 2022)	✓	✓	×	×	×	✓	✓
Rubrix (https://rubrix.readthedocs.io/en/v0.4.1/#)	✓	✓	×	×	×	✓	✓
Brat (Stenetorp et al., 2012)	✓	✓	×	×	✓	×	✓
Doccano (Nakayama et al., 2018)	✓	✓	×	×	✓	×	✓

Table 2: Feature matrix for selected free tools. **Morph** = morpheme segmentation/glossing; **IGT** = interlinear glossed text support (∼ = partial); **Collab.** = remote collaboration; **AL** = active learning; **Maint.** = actively maintained. × = no. Top block: free tools with morpheme support; middle block: free tools with AL but no morpheme support; bottom block: general-purpose popular free NLP tools for comparison. Full table can be seen in https://docs.google.com/spreadsheets/d/1o-IQTC7vIK1xRqdOoIzdgAlrsedkJeIswAeFyeiS93M/edit?usp=sharing.

advanced functionality and customisability often reduce simplicity of the user interface. Some tools prioritise advanced functionality over simplicity. Among the 44 free tools, 12 are web-based because they support remote collaboration. This means they also require minimal installation. For example, Doccano (Nakayama et al., 2018)'s web-based interface appeals to non-technical users but its limited flexibility makes it less suitable for handling complex linguistic tasks. In contrast, tools like TEITOK (Janssen, 2016)offer broader functionality and customisation but require a more involved server setup. Projects with limited IT resources might favour user-friendly, web-based tools, while more technically complex projects could benefit from tools that, although harder to set up, support rich and customised annotation workflows.

5.3 Gap Analysis

Table 3 quantifies the key gaps between free tool capabilities and the requirements[5] of language documentation practice.

5.3.1 Morphological Annotation Support Gap

Morpheme-level annotation is one of the critical and consistently underserved features for language documentation (Klimek et al., 2021; Gromann et al., 2024; Rice et al., 2025). Among the 44 free accessible tools, 15 support morpheme segmentation and glossing. This number drops further when combined with other requirements relevant to documentary workflows.

Requirement	Tools	%
Morpheme segmentation	15/44	34%
Morpheme + collaboration	6/44	14%
Morpheme + AL	1/44	2%
Open + morpheme + collaboration	4/44	9%
Morpheme + adjudication	3/44	7%
IGT / interlinear glossing	∼3/44	<7%

Table 3: Gap analysis: proportion of free accessible tools meeting key language documentation requirements.

Of the 15 free tools with morpheme support[6], the majority are linguistic corpus tools: EXMARaLDA (Schmidt and Wörner, 2014), TEITOK (Janssen, 2016), CATMA (Horstmann, 2020), Praaline (Christodoulides, 2018), MMAX2 (Müller and Strube, 2006), and SLATE (Kummerfeld, 2019). The Interlinear Text Editor (part of SIL's FLEx ecosystem) (Hughes et al., 2004) is the only explicitly purpose-built IGT tool in this set.

[5] As mentioned in Section 3.2, Morpheme feature requires sub-word annotation capability, while the IGT feature additionally requires aligned interlinear tier display in the documentary linguistics format.

[6] Label Studio (Tkachenko et al., 2020) is not included here because although it supports arbitrary text-span annotation, including partial-word spans, but does not provide native support for morpheme segmentation and glossing as a first-class annotation workflow.

5.3.2 The IGT Gap

Interlinear glossed text is the standard output format of language documentation, yet its importance is not reflected in the supported annotation types among free tools. This reflects a mismatch between the dominant design assumptions of current NLP annotation tools and the practical requirements of language documentation. Although modern NLP tools are typically optimized for span-level or token-level annotation in standardized text classification or sequence labeling tasks, documentary workflows require persistent alignment across multiple linguistic tiers. The very minimal IGT support therefore represents a distinct and foundational gap in the current tool landscape.

5.3.3 The Collaboration Gap

Remote collaboration is essential for language documentation projects involving geographically dispersed teams or a combination of linguists and community members fulfilling different roles. Of the 44 free tools, 12 support remote collaboration. Among free tools with morpheme support, this figure is lower: 6 of 15 offer collaborative functionality, including EXMARaLDA (Schmidt and Wörner, 2014), TEITOK (Janssen, 2016), CATMA (Horstmann, 2020), Lexonomy (Měchura and Rychlý, 2017), PACTE (Ménard and Barrière, 2017), and WebLicht (Ljubešić et al., 2017). Three tools–TEITOK, CATMA, and EXMARaLDA–offer both web-based collaboration and morpheme-level annotation. They are free but each carries significant technical setup requirements that may be prohibitive for community-based projects without dedicated infrastructure.

5.3.4 The Active Learning Gap

Active learning (Settles, 2012), the ability of an annotation tool to leverage a partially trained model to prioritize uncertain examples and accelerate annotation, has high pragmatic value in low-resource NLP settings if its functionality can be made accessible to non-technical users in their workflow. Of 44 free tools, 10 report some active learning functionality, and of these, PACTE (Ménard and Barrière, 2017) supports morpheme segmentation. The remaining 9 AL-capable free tools (Label Studio (Tkachenko et al., 2020), INCEpTION (Klie et al., 2018), YEDDA (Yang et al., 2018), ALToolbox (Tsvigun et al., 2022), Rubrix[7], Markup (Dobbie et al., 2021), PAL (Skeppstedt et al., 2017), CVAT[8], Hugging Face (Jain, 2022)) are all oriented toward standard NLP tasks such as named entity recognition, and do not support morpheme-level annotation. This near-complete absence of AL support for morphological annotation among free tools represents the sharpest gap between the potential contribution of NLP and current documentary linguistics needs.

5.3.5 Data Accessibility Gap

Unfortunately, the Sensitivities criteria outlined in Section 3.2 reveal a further structural gap: many free tools fail the accessibility and data sovereignty requirements of community-based documentation. Internet-dependent architectures present barriers for communities working in low-connectivity contexts. Among free tools with morphological support, Unicode coverage is generally good (>90%), but offline operation, multi-language UI support, and self-hostable server architectures are available in a small subset—principally TEITOK (Janssen, 2016)and CATMA (Horstmann, 2020).

6 Suggestions for Tool Development

Given the features and gaps identified above, we offer the following suggestions for the development of annotation tools targeting the documentary linguistics community.

Design a modern IGT editor. We articulate the most pressing unmet needs in the documentary linguistics tool landscape is a user-friendly, open-source, browser-based IGT editor with morpheme segmentation, gloss lookup, interlinear alignment, and real-time collaborative editing. User-friendly graphic interfaces accommodate users who are not software developers. Open-source is amenable to development sustained by the relatively small, short-term budgets common in academia and community organizations. Browser-based tools are independent of the user's operating system. Existing open-source tools like INCEpTION (Klie et al., 2018) or Label Studio (Tkachenko et al., 2020) could be extended with IGT-specific annotation schemas; the Ligt (Ionov, 2025) data model provides a tested and extensible IGT ontological reference architecture for a web-native reimplementation.

[7] https://github.com/rasbt/rubrix?tab=readme-ov-file

[8] https://github.com/cvat-ai/cvat?tab=readme-ov-file

Integrate active learning for morphological annotation. Investment in active learning for morphological annotation, such as building on purpose-trained morphological segmentation models or LLM-assisted IGT annotation, could significantly reduce the annotation bottleneck for documentation projects. Tools like INCEpTION (Klie et al., 2018) or TEITOK (Janssen, 2016) might be adapted to prioritise morpheme-level active learning as an extension module, given that only one free tool currently supports both AL and morpheme annotation.

Prioritise self-hostable, offline-capable architectures. Tools designed for community-based documentation should support offline-first operation or self-hosted server deployment, addressing both fieldwork connectivity constraints and data sovereignty requirements. Examples are TEITOK (Janssen, 2016) and CATMA (Horstmann, 2020) which already support self-hosted deployment and could serve as a default design principle for new tools intended for the documentation community.

Extend adjudication for community co-annotation. For community language documentation, where multiple stakeholders, ranging from linguists, community members, to heritage speakers may annotate the same data, adjudication is not merely a quality control mechanism but a collaborative practice that respects community expertise. Adjudication features could be designed for collaborative negotiation to allow flexible annotator role models and transparent conflict resolution that treat community input as a first-class form of linguistic knowledge. These functionalities are highly valuable for projects involving community members and linguists with different types of expertise.

Address tool attrition and data portability. Data portability is a crucial issue raised compellingly over 20 years ago (Bird and Simons, 2003; Simons and Bird, 2003). The inaccessibility of 32 tools and unclear maintenance status of several free tools is a concern for data preservation. Tool repositories should be archived with initiatives such as Software Heritage, and documentation projects should include explicit export plans in interoperable formats (e.g., ELAN's EAF, FLEx's LIFT/FLEXTEXT, Ligt's RDF vocabulary).

The unclear maintenance status of many free tools underscores the drawbacks of depending on free tools, but also points to an advantage that NLP annotation might provide academic and community linguists. Subscribing to a commercial tool that supports important IGT tasks may be to be cheaper and more sustainable long-term than a custom-built, open-source tool. Commercial companies may provide upon inquiry free subscriptions for educational teams and others might be happy to hear how their tools could better support scientific and community efforts. Such interactions should be approached with very clear understandings about financial, time, or storage costs and the ownership or allowable uses of the data.

7 Recommended Tools for Language Documentation

We further provide a curated list of free tools most suitable for language documentation workflows, filtered to those supporting morpheme segmentation and glossing. Active learning support is noted as a bonus criterion. Among the tools in Table 4, the following stand out for specific use cases:

Best for collaborative web-based documentation: TEITOK (Janssen, 2016) and CATMA (Horstmann, 2020) are the strongest candidates. Both are web-based, actively maintained, open-source, and support morphological annotation with server self-hosting for data sovereignty. TEITOK (Janssen, 2016) offers IGT-adjacent tier support suited for transcription-linked morphological annotation; CATMA (Horstmann, 2020)provides flexible free-form tagset definition useful for under-described languages with non-standard grammatical categories.

Best for IGT-centred workflows: The dominant free tools for IGT-centred workflows is the Interlinear Text Editor (part of SIL FLEx) (Hughes et al., 2004). The Interlinear Text Editor is purpose-built for interlinear glossing, though FLEx has increasingly shifted toward lexicon management via IGT rather than serving as a primary annotation environment. This tool is desktop-only with limited collaboration support, and none integrates an active learning component. Despite these limitations, they remain the de facto standard for IGT workflows due to the absence of any modern, web-based alternative.

Best for active learning: PACTE (Ménard and Barrière, 2017) is the only free tool combining morphological annotation with active learning. However, it is closed-source and its AL component is

Tool	Type	Open	Collab.	AL*	Maint.	IGT	Notes
EXMARaLDA (Schmidt and Wörner, 2014)	Corpus	✓	✓	×	✓	∼	Multi-tier XML; self-hosted server option; strong interop with ELAN
TEITOK (Janssen, 2016)	Corpus	✓	✓	×	✓	∼	Web-based; server self-hostable; TEI-XML; good for transcription + morphology
CATMA (Horstmann, 2020)	Corpus	✓	✓	×	✓	×	Web-based; free-form tagsets; suitable for team annotation projects
Lexonomy (Měchura and Rychlý, 2017)	Lexicon	✓	✓	×	×	×	Web-based; lexicographic focus; morpheme-level lexical entries
Interlinear Text Ed. (Hughes et al., 2004)	IGT	✓	×	×	✓	✓	Part of SIL FLEx; purpose-built for IGT; desktop only
Praaline (Christodoulides, 2018)	Corpus	✓	×	×	✓	∼	Desktop; phonetic + morphological tiers; strong prosody support
MMAX2 (Müller and Strube, 2006)	Corpus	✓	×	×	✓	×	Desktop; multi-level annotation; XML-based
SLATE (Kummerfeld, 2019)	Corpus	✓	×	×	✓	×	Lightweight; command-line friendly; morpheme span annotation
UAM CorpusTool (O'Donnell, 2008)	Corpus	×	×	×	✓	×	Multi-layer annotation; flexible schema; desktop only
PACTE (Ménard and Barrière, 2017)	NLP	×	✓	✓*	✓	×	Closed-source; **only free tool combining AL and morpheme support**

Table 4: Recommended free tools for language documentation with morpheme segmentation support. **AL*** = active learning (bonus criterion); ✓* = supports AL. **IGT** = interlinear glossed text support (∼ = partial). Tools are ordered from most to least suitable for collaborative community-based documentation.

designed for parallel corpus annotation rather than the small, single-language datasets typical of endangered language fieldwork.

Tools to watch: INCEpTION (Klie et al., 2018), while not currently supporting morpheme segmentation, is open-source, actively developed, and has a plugin architecture that makes it the most promising candidate for future extension toward IGT and morphological active learning.

8 Conclusion and Future Work

We have presented a systematic survey of annotation tools evaluated from a language documentation perspective, focusing on the 44 free tools accessible to practitioners with constrained resources. Our analysis across 32 feature dimensions reveals that although several NLP tools can be recommended for language documentation tasks a notable misalignment exists between the NLP/industry tool ecosystem and documentary linguistics needs. The ecosystem has developed powerful annotation infrastructure for high-resource settings, but the morphological annotation capabilities, community-accessible architectures, and offline-ready designs required for endangered and Indigenous language documentation remain underserved.

Of 44 free tools, 15 support morpheme segmentation and glossing; 6 combine this with remote collaboration; and 1 adds active learning (that tool is closed-source). We provide a curated recommendation table (Table 4) to help linguists and community language workers navigate this landscape and we describe the categories, criteria, and trade-offs we considered to assist informative evaluations of future options. It should be noted that with the recent rapid rise of AI in the form of LLMs, this landscape may change precipitously. But we intend for the rubrics we designed to be used to evaluate existing tools and to guide the development of improved platforms in the future.

Limitations

Our evaluation is limited to tools accessible as of 2025; tool features and maintenance status may have changed. A thorough user testing of every tool was not feasible. Feature coding was based on public documentation and secondary sources rather than direct inspection, and may therefore not fully reflect actual tool capabilities. We did not compute inter-annotator agreement statistics; disagreements were resolved through discussion, which may introduce subjective bias in borderline cases. The free/paid classification is based on publicly stated pricing at time of evaluation and may not capture all licensing nuances (e.g., freemium tiers or institutional agreements).

References

Eric Auer, Albert Russel, Han Sloetjes, Peter Wittenburg, Oliver Schreer, S. Masnieri, Daniel Schneider, and Sebastian Tschöpel. 2010. ELAN as flexible annotation framework for sound and image processing detectors. In *Proceedings of LREC 2010*, pages 890–893.

Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582.

Steven Bird and Dean Yibarbuk. 2024. Centering the Speech Community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta. Association for Computational Linguistics.

Nathan Thanyehténhas Brinklow. 2021. Indigenous language technologies: Anti-colonial oases in a colonizing (digital) world. *WINHEC: International Journal of Indigenous Education Scholarship*, (1):239–266.

George Christodoulides. 2018. Praaline: An open-source system for managing, annotating, visualising and analysing speech corpora. In *Proceedings of ACL 2018, System Demonstrations*, pages 111–115, Melbourne, Australia. Association for Computational Linguistics.

Samuel Dobbie, Huw Strafford, W Owen Pickrell, Beata Fonferko-Shadrach, Carys Jones, Ashley Akbari, Simon Thompson, and Arron Lacey. 2021. Markup: a web-based annotation tool powered by active learning. *Frontiers in Digital Health*, 3:598916.

David M. Eberhard, Gary F. Simons, and Alison J. Robinson, editors. 2026. *Ethnologue: Languages of the World*, twenty-ninth edition. SIL International, Dallas, Texas.

Luke Gessler, Alexis Palmer, and Katharina Von Der Wense. 2025. Understanding the gap: an analysis of research collaborations in NLP and language documentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 867–877, Vienna, Austria. Association for Computational Linguistics.

Dagmar Gromann, Elena-Simona Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Liudmila Mockiene, Michael Rosner, Ineke Schuurman, Gilles Sérasset, Purificação Silvano, Blerina Spahiu, Ciprian-Octavian Truică, Andrius Utka, and Giedre Valunaite Oleskeviciene. 2024. Multilinguality and LLOD: A survey across linguistic description levels. *Semantic Web*, 15(5):1915–1958.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1).

Jan Horstmann. 2020. Undogmatic literary annotation with CATMA. *Annotations in Scholarly Editions and Research: Functions, Differentiation, Systematization*, page 157.

Baden Hughes, Catherine Bow, and Steven Bird. 2004. Functional requirements for an interlinear text editor. In *LREC*.

Maxim Ionov. 2025. Ligt: Towards an Ecosystem for Managing Interlinear Glossed Texts with Linguistic Linked Data. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 100–105. Unior Press.

Shashank Mohan Jain. 2022. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043.

Daniel Kaufman and Raphael Finkel. 2018. Kratylos: A tool for sharing interlinearized and lexical data in diverse formats. *Language Documentation and Conservation*, 12:124–146.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Bettina Klimek, Markus Ackermann, Martin Brümmer, and Sebastian Hellmann. 2021. MMoOn Core – the Multilingual Morpheme Ontology. *Semantic Web*, 12(5):813–841.

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

Jonathan K Kummerfeld. 2019. SLATE: a super-lightweight annotation tool for experts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 7–12.

Nikola Ljubešić, Tomaž Erjavec, Darja Fišer, Erhard Hinrichs, Marie Hinrichs, Cyprian Adam Laskowski, Filip Petkovski, and Wei Qui. 2017. Multilingual text annotation of slovenian, croatian and serbian with weblicht. In *Proceedings of the CLARIN Annual Conference 2017*, pages 1–4, Budapest, Hungary.

Kirk E Lowery. 2008. Review of Emdros: The database engine for analyzed or annotated text.

Kovář Vojtěch Měchura, Michal Boleslav and Pavel Rychlý. 2017. Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, pages 19–21.

Pierre André Ménard and Caroline Barrière. 2017. PACTE: a collaborative platform for textual annotation. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Sarah Moeller and Antti Arppe. 2024. Machine-in-the-Loop with Documentary and Descriptive Linguists. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 27–32, St. Julians, Malta. Association for Computational Linguistics.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Mariana Neves and Jurica Ševa. 2021. An extensive review of tools for manual annotation of documents. *Briefings in bioinformatics*, 22(1):146–163.

Mick O'Donnell. 2008. Demonstration of the UAM corpustool for text and image annotation. In *Proceedings of the ACL-08: HLT Demo Session*, pages 13–16.

Enora Rice, Katharina von der Wense, and Alexis Palmer. 2025. Interdisciplinary Research in Conversation: A Case Study in Computational Morphology for Language Documentation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11284–11296, Suzhou, China. Association for Computational Linguistics.

Keren Rice. 2011. Documentary linguistics and community relations.

Chris Rogers. 2010. Review of Fieldworks Language Explorer (FLEx) 3.0. *Language Documentation & Conservation*, 4:78–84.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

Burr Settles. 2012. *Active Learning*. Morgan & Claypool.

Gary Simons and Steven Bird. 2003. The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing*, 18(2):117–128.

Maria Skeppstedt, Carita Paradis, and Andreas Kerren. 2017. PAL, a tool for pre-annotation and active learning. *Journal for Language Technology and Computational Linguistics*, 31(1):91–110.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Nick Thieberger. 2009. Culture clash – Humanities research and computing: a case study of Interlinear Glossed Text (IGT). Sydney, Australia.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label Studio: Data labeling software. `https://github.com/HumanSignal/label-studio`.

Salvatore Trani, Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. 2014. Dexter 2.0: an open source tool for semantically enriching data. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track*, volume 1272, pages 417–420.

Akim Tsvigun, Leonid Sanochkin, Daniil Larionov, Gleb Kuzmin, Artem Vazhentsev, Ivan Lazichny, Nikita Khromov, Danil Kireev, Aleksandr Rubashevskii, Alexander Panchenko, Olga Shahmatova, Dmitry Dylov, Igor Galitskiy, and Artem Shelmanov. 2022. ALToolbox: A set of tools for active learning annotation of natural language texts. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 406–434, Abu Dhabi, UAE. Association for Computational Linguistics.

Martin Weisser. 2016. DART–the dialogue annotation and research tool. *Corpus Linguistics and Linguistic Theory*, 12(2):355–388.

Anthony C Woodbury. 2003. Defining documentary linguistics. *Language documentation and description*, 1.

Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018. YEDDA: A lightweight collaborative text span annotation tool.

A Full evaluation schema

See Table 5.

B Full list of annotation tools

See Table 6.

Feature	Evaluation question	Response type
Metadata		
Tool name	Name of the tool	Free text
Source	Where was the tool identified? (survey, workshop, wiki, etc.)	Free text
Link	Main website, repository, or download page	Free text
Cost		
Cost	What is the cost model? (free / freemium / paid)	Categorical
Adoption rationale	Why would (or wouldn't) a linguist adopt this tool?	Free text
Sustainability and Longevity		
Open source	Is the source code publicly available?	Binary
Maintenance status	Is the tool actively maintained?	Binary
Maintaining org.	Type of maintaining organization (academic / non-profit / for-profit)	Categorical
Portability		
Export formats	What file formats does it export?	Free text
Import options	What file formats does it import?	Free text
User Friendliness		
Technical setup	What are the installation requirements and difficulty?	Free text
Ease of starting	Can a user get started without reading documentation?	Binary
Ease of instructions	How easy is it to follow the documentation?	Free text
UI language	What languages is the UI available in?	Free text
Unicode support	What Unicode/font coverage does it provide?	Free text
Guideline support	Can annotation guidelines be uploaded and displayed in the UI?	Binary
Annotation lookup	Can users look up prior annotations of the current token?	Binary
Documentation quality	How robust and accessible is the software documentation?	Free text
Sensitivities		
Data hosting	Where is data hosted? Who owns it? Privacy/IPR concerns?	Free text
Linguistic Annotation Capabilities		
Adjudication	Does it support adjudication or consensus voting across annotators?	Binary
Word glossing	Token-level selection with open label list?	Binary
Morpheme segmentation	Sub-word segmentation and glossing?	Binary
Syntactic/semantic	POS, SRL, or syntactic role labeling?	Binary
Transcription/translation	Sentence-aligned transcription or translation?	Binary
Corpus processing	Can it process multiple files or corpora simultaneously?	Binary
Remote collaboration	Is web-based or server-synced collaboration supported?	Binary
Active Learning		
Any AL functionality	Does the tool offer any active learning support?	Binary
Built-in training	Does it support built-in model training and testing?	Binary
Import predictions	Can it import model annotations and confidence scores?	Binary
AL usability	How intuitive is the AL interface for non-technical users?	Free text
Prediction feedback	Can users rate, comment on, or correct model predictions?	Binary
AL comments	Additional observations on the AL interface	Free text

Table 5: Full evaluation schema used for tool assessment, grouped into seven categories. In addition to the seven analytic categories described in Section 3.2, we also recorded basic metadata for traceability and reproducibility. Binary features were coded Yes/No/Partial where applicable.

Included (n = 44)		Excluded (n = 54)		
#	**Tool**	**#**	**Tool**	**Reason**
1	INCEpTION	1	Labelbox	Not free
2	NLP Lab (JSL)	2	LightTag	Not free
3	Label Studio	3	TagTog	Not free
4	Doccano	4	Prodigy	Not free
5	Brat	5	UBIAI	Not free
6	CVAT	6	Labellerr	Not free
7	PIAF Platform	7	Amazon SageMaker GT	Not free
8	Hugging Face	8	Dataturks	Not free
9	Sloth	9	Superannotate	Not free
10	Atomic	10	Google Cloud AutoML NL	Not free
11	BFSU Qualitative Coder	11	Playment	Not free
12	CATMA	12	Appen	Not free
13	CorefAnnotator	13	@nnotate	Inaccessible
14	Corpona	14	ACTRES Corpus Manager	Inaccessible
15	DART	15	AMALGAM	Inaccessible
16	Dexter	16	ANVIL	Inaccessible
17	DISCO	17	DisMo	Inaccessible
18	Emdros	18	PALinkA	Inaccessible
19	EXMARaLDA	19	Sketch Engine	Not free
20	Lexonomy	20	SPPAS	Inaccessible
21	MMAX2	21	SPre	Inaccessible
22	Praaline	22	VideoAnt	Inaccessible
23	RSTTool	23	WebAnno	Inaccessible
24	SLATE	24	Worldbuilder	Inaccessible
25	Synpathy	25	QualCoder	Inaccessible
26	The Simple Corpus Tool	26	Embedding Viewer	Inaccessible
27	TreeTagger	27	Grammar Explorer	Inaccessible
28	UAM CorpusTool	28	Kura	Inaccessible
29	WebLicht	29	NooJ	Inaccessible
30	YEDDA	30	OneClick Terms	Not free
31	TEITOK	31	Systemics	Inaccessible
32	Sanchay	32	Encord	Not free
33	Text Feature Analyser	33	SysAm	Inaccessible
34	EEVEE	34	TATOE	Inaccessible
35	Markup	35	Tgrep2	Inaccessible
36	MedTAG	36	TIGERSearch	Inaccessible
37	PAL	37	XTrans	Inaccessible
38	ALToolbox	38	Kili	Not free
39	Rubrix	39	ATLAS	Inaccessible
40	Interlinear Text Editor	40	Emu Speech DB System	Speech scope
41	Kratylos	41	ToBI Labelling	Inaccessible
42	AGTK	42	MATE Workbench	Inaccessible
43	CLaRK	43	NITE XML Toolkit	Inaccessible
44	PACTE	44	PRAAT	Speech scope
		45	SACODEYL Transcriptor	Speech scope
		46	SignStream	Inaccessible
		47	SoundScriber	Inaccessible
		48	TalkBank	Inaccessible
		49	TASX-Annotator	Inaccessible
		50	Transana	Speech scope
		51	Transcriber	Inaccessible
		52	UCSB Disc. Transcription	Inaccessible
		53	VOCALE	Inaccessible
		54	wavesurfer	Speech scope

Table 6: All 98 tools surveyed, split by inclusion status. Left: 44 tools included in the main analysis. Right: 54 excluded tools with reason.