The Missing Middle: Language Documentation Needs Better Infrastructure, Not Better Models

Luke Gessler[1*] **Antonios Anastasopoulos**[2] **Sandra Auderset**[3] **Timotheus Bodt**[4] **Shobhana Chelliah**[1] **Sebastien Christian**[5] **Maxime Fily**[6] **Santiago Herrera**[7] **Eva Huber**[8] **Sharid Loáiciga**[9] **Marieke Meelen**[10] **Robert Östling**[11] **Alexis Palmer**[12] **Eline Visser**[13]

[1]Indiana University Bloomington [2]George Mason University [3]University of Bern
[4]Trinity College Dublin [5]University of French Polynesia
[6]National Institute for Oriental Languages and Civilizations (INALCO)
[7]Université Sorbonne Paris Nord, LIPN, CNRS [8]University of Cologne
[9]University of Gothenburg, FLoV [10]University of Cambridge [11]Stockholm University
[12]University of Colorado Boulder [13]Uppsala University
[*]lgessler@iu.edu

Abstract

Despite decades of progress in human language technology (HLT) and growing research interest in endangered languages, practical uptake of HLT in documentary linguistics workflows remains rare. In this opinion piece, we report on a structured dialogue among approximately twenty academics convened to diagnose why this gap persists. Across all topics, we identify a recurring structural problem, which we call the missing middle: despite the existence of many potentially useful HLTs, the connective infrastructure necessary to make them genuinely accessible to linguists and language communities does not exist. We report the details of our discussion and make four specific recommendations for how those active in language documentation and HLT research might orient their future work.

1 Introduction

Since the emergence of documentary linguistics as a distinct enterprise in the late 1990s (Himmelmann, 1998), many languages have lost their last speakers (Evans, 2009). Over roughly the same period, human language technology (HLT)[1] has also advanced dramatically, producing systems which are in principle capable of facilitating language documentation by automating some language documentation tasks. Still, practical uptake of HLT in documentary workflows remains rare (Gessler et al., 2025), and this observation is by now a familiar and long-standing one (Bird, 2009; Good et al., 2014) despite a considerable amount of attention to it. Various accounts of *why* this situation persists have been offered (e.g., Flavelle and Lachler, 2023), but the field has yet to converge on a unified diagnosis or a concrete plan of action.

We are a group of approximately twenty researchers in documentary linguistics and human language technology who gathered to discuss this matter (see Acknowledgments). Over the course of three days of structured discussion, we found that every topic we examined—scientific outputs, community outputs, ethical issues, and data analysis issues—kept revealing the same structural problem. We have come to call it the **missing middle**: we lack the necessary medium to bridge what HLT researchers produce and what documentary linguists actually need in their workflows.

In this opinion piece, we summarize the four main sessions of our discussion, each of which illuminates a different facet of the missing middle, and then offer a set of four concrete recommendations for how to address it.

2 Related Work

The disconnect between HLT research and documentary practice has been recognized for well over a decade. Bird (2009) identified the mismatch between NLP research agendas and the practical needs of linguistic fieldwork as early as 2009, and the intervening years have seen a steady stream of papers revisiting the problem. Neubig et al. (2020) reported on a workshop that brought together com-

[1]We use "HLT" to refer broadly to any technology which computationally processes language data, encompassing work done in natural language processing, speech processing, computational linguistics, and other fields.

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 136–147

munity members, documentary linguists, and technologists to build prototypes for nine indigenous languages in an attempt to create HLT that is more relevant for the latter two groups. More recently, Gessler and von der Wense (2024) provided empirical evidence that scarce graduate training and low rates of interdisciplinary collaboration contribute to the gap, and Gessler et al. (2025) presented survey and interview data showing the role of misaligned professional incentives, technical knowledge burdens, and limitations in existing language documentation software. Rice et al. (2025) argue that centering the user is essential to achieve practical uptake.

Several proposals have targeted the structural dimensions of the problem. Gessler (2022) argued that the core bottleneck is not model quality but software infrastructure, and presented a system for integrating NLP into documentary workflows. Zariquiey et al. (2022) proposed making NLP-ready annotated data a standard deliverable of documentation projects, aiming to bridge the two fields at the level of data practice.

3 Sessions

Discussions at the workshop were divided along four major topics, and we summarize our discussions in each subsection below.

3.1 Scientific Outputs

The first session addressed scientific outputs. Participants observed that in many ways, current norms around what are considered valuable scientific outputs are hindering the delivery of HLT with practical impact for documentary work, even in cases where this is an explicit goal.

Academic and Practical Goals A central issue identified was the difference between "academic" and "practical" HLT products. Participants noted that creating practical tools, user interfaces, annotated datasets, or plugins is rarely rewarded in academia, as these are mostly not regarded as intellectually meritorious. Consequently, HLT systems are typically developed just up to the point where experimental evidence of their excellence and novelty may be published. Afterwards, they are abandoned, as any further work on making the system easier for others to use or more usable across a wide range of datasets would, from a career perspective, constitute wasted effort.

This constitutes a major obstacle for language documentation efforts, as projects often lack budgets for dedicated engineers, and the necessary integration and UI/UX work falls outside the scope of a typical linguist's technical background and research goals. This issue extends beyond software to other valuable outputs, such as dictionaries and annotated corpora, which are of great practical utility but are rarely incentivized by academic bodies for tenure or promotion.

Shared Tasks Shared tasks were identified as a promising means for engaging HLT researchers in problems in language documentation. The competitive and time-bounded aspects of shared tasks are highly motivating for technologists, who are eager to work on novel and extrinsically motivated problems with a guaranteed publication after a few months. Many shared tasks have already been organized specifically for issues in language documentation at venues such as AmericasNLP (Mager et al., 2021; Ebrahimi et al., 2024) and SIGMORPHON (Ginn et al., 2023), and some have begun targeting community-facing outputs directly, such as the generation of educational materials for indigenous languages (Chiruzzo et al., 2024).

Participants viewed shared tasks with mixed feelings. On the negative side, many issues with shared tasks stem from their transitory nature. The systems produced are often abandoned as soon as the shared task is finished, as researchers are not incentivized to provide the amenities which would grant their systems true practical utility. Just as importantly, shared tasks do not clearly facilitate long-term relationship-building required for community-led projects.

However, participants also noted some benefits of shared tasks when designed thoughtfully. First, they are an effective way to direct the attention of the HLT community towards the particular problems endemic to language documentation, which might have otherwise gone unstudied in a computational setting. As has been noted before, it is difficult for HLT researchers and linguists to communicate with each other (Flavelle and Lachler, 2023; Gessler et al., 2025, *inter alia*), and predigesting a problem from language documentation by describing it in familiar language and providing a clean accompanying dataset can be very effective for directing HLT researcher interest to a problem. Compare this to an alternative, where a linguist and an HLT researcher must struggle to

understand each other for unclear ultimate benefit.

Second, while shared tasks do not directly facilitate long-standing collaborations, participants noted that any such collaboration must begin with some kind of contact, and shared tasks appear to be the only obvious way to facilitate encounters between language documenters and HLT researchers in a scalable way. To this end, participants also speculated that shared tasks might be improved if they required some contact between the two groups *during* the shared task, rather than staking all hopes of contact on the day or two during the conference when the shared task results are to be presented. This could come in the form of, for example, qualitative evaluation of system outputs, and could perhaps even involve community members, thereby also addressing potential concerns that have been raised regarding treating language as data for machine exploitation (Bird and Yibarbuk, 2024).

Participants therefore viewed shared tasks as flawed, but the best means available for ultimately building relationships between HLT researchers and language documenters. One suggestion was to host them at language documentation venues such as ComputEL, which could promote more meaningful contact between language documenters and HLT researchers. Participants felt that we have not yet found the most productive form of a shared task for language documentation, and that more thought ought to be put into how to go further in using shared tasks to facilitate intellectual exchange and social connection between the two fields.

Data Culture Beyond the mechanics of academic incentives, participants identified cultural differences in how data is regarded between the two fields. Documentary linguists often spend years building relationships with communities to create datasets, making them understandably hesitant to collaborate with HLT researchers without a clear understanding of the benefits and risks. Conversely, HLT researchers are not incentivized to go "digging in archives for data" and can have a tendency to treat complex linguistic information as a commodity, divorced from its context, biases, and the nuances of its collection.[2] This mismatch in temperament presents a significant barrier to collaboration. For example, such decontextualized work on the part of an HLT researcher can lead to technologically novel but practically irrelevant systems, undermining the efforts of both researchers and failing to deliver meaningful benefits to the language communities in question.

Attaining Practical Success Finally, participants emphasized the immense difficulty of practical deployment of systems—the "last mile" problem: the successful deployment of an HLT system into a real documentary workflow. One important requirement for attaining this more productive workflow is integrating such a system into existing language documentation apps (like ELAN, Wittenburg et al. 2006, or FLEx, Butler and van Volkinburg 2007), which was noted to be quite challenging. Without dedicated engineering effort for deployment, training, usability tuning, and maintenance for these systems, even successful AI models are unlikely to have a real-world impact.

We noted a few cases where HLT has been successfully integrated into documentary workflows on a small scale. Michaud et al. (2018), for instance, embedded automatic phonemic transcription into a workflow for Yongning Na, producing transcripts that served as a useful "canvas" for linguist correction. But such successes remain isolated, and they notably tend to involve ASR and transcription rather than text-based NLP tasks.

3.2 Community Outputs

The second session turned to the question of what kinds of outputs are actually useful for language communities, as opposed to what researchers might assume is useful.

Diversity A recurring theme throughout this session was the sheer diversity of language communities and their relationships to technology. Some communities have considerable technical capacity—participants noted one community in Dharamshala, India, whose members are pursuing graduate degrees in computer science and working directly with LLMs. Others have strong oral traditions and limited literacy but are enthusiastic users of video and voice messaging on mobile phones. Still others place high value on paper as a tangible, lasting object, and digital products which do not also lead to products on paper may be of limited interest, not least because internet access is limited. This diversity means that there is no single

[2] Nevertheless, calls to "mobilise the archive" (Bird, 2020) have, to a small extent, been answered (Agarwal and Anastasopoulos, 2025; Agarwal et al., 2025).

answer to the question of what a useful community output looks like: what is transformative for one community may be irrelevant or even unwelcome to another.

Practical HLT Participants observed a frequent mismatch between what HLT researchers build and what communities request (cf. Liu et al., 2022). Some of the most consistently desired technologies are mundane by HLT standards: keyboards, for instance, are frequently requested by communities and are genuinely useful, yet even these can fail to gain traction because users cannot always set them up on their own devices.

A striking example of how communities actually engage with technology came from a participant who noted that many of their speakers do not read or write, but watch videos constantly, using voice assistants in a lingua franca to search for content. In such communities, people are already navigating technology in pragmatic, diglossic ways, and the question of whether they "need" an interface in their language is not straightforward. Voice messages were also noted as hugely popular in many indigenous communities, though no formal studies of their impact and potential were known to participants.

Nontraditional Outputs and Organic Traction Several participants observed that the language technology outputs which gain the most traction are often not the products of funded research projects at all. One participant described a Facebook page they had created with the sole requirement that all interaction take place in the language; this had become one of their most impactful contributions. Another mentioned an online dictionary maintained in someone's spare time. These informal, community-facing outputs often succeed precisely because they are lightweight, immediately usable, and embedded in the social fabric of daily life—qualities that more ambitious, research-driven tools often lack.

"Old" Tech Participants challenged the common assumption that the most sophisticated available HLT is necessarily the most useful. One concrete example was discussed: a community speaking Mapudungun needed teaching tools, and after consultation it was determined that a finite-state transducer was more appropriate than a neural model (Ahumada et al., 2022). More broadly, participants argued that the right tool for a given community's needs might be as simple as an app, a keyboard, or a pedagogical grammar illustrated by a local artist (cf. Cruz 2022), not a large language model (Claus et al., 2026).

Technologists as Consultants One participant proposed a useful framing: technologists working with language communities should think of themselves as consultants whose job is to address their client's concerns, however difficult or unglamorous. This framing was broadly endorsed, though participants noted that technologists may only do so within the hard constraints of what is required to maintain their careers. Further, community members may not know the full scope of what is technologically possible, and so it may be difficult for them to know what to ask for from technologists.

Several participants pointed to existing models for this kind of engagement, including ELAN workshops run at universities and in villages, language documentation stations in Guatemala and Peru, and programs which bring students to field sites for combined training and capacity building.

The Limits of "Helping" Finally, participants grappled with the tension between wanting to be useful and the risk of overstepping. Building relational networks with communities is valuable—one participant noted that something as simple as charging people's phones during fieldwork can establish trust—but the line between genuine partnership and unwanted intervention is not always clear. This came up concretely in a discussion about whether researchers should assist communities with health-related information: while some forms of assistance seemed straightforward, others were judged too complex to provide responsibly, and participants acknowledged that external aid could risk undermining local practices. Participants felt that decisions like these can only be made on a case-by-case basis, accounting for community-specific considerations and individual ethical judgments.

3.3 Ethical Issues

The third session addressed ethical dimensions of applying HLT to language documentation, with a focus on data.

Data Sovereignty and Consent Communities and technologists have fundamentally different relationships to linguistic data. For many communi-

ties, language data is not fungible: some knowledge must be earned, and speakers may be selective about which data they share, with whom, and under what circumstances. This stands in contrast to the default orientation in HLT research, where data is a commodity to be collected, packaged, and distributed as efficiently as possible—often under frameworks like FAIR (Wilkinson et al., 2016) that assume openness as a default, in tension with Indigenous data governance frameworks like CARE (Carroll et al., 2020) and OCAP (First Nations Information Governance Centre, 2014).

We join others before us in observing that this mismatch has consequences. Depositors to linguistic archives have signed digital rights agreements without fully understanding the ramifications—effectively permitting, for example, the training of ASR models on their speech without their knowledge. Some of us have updated our consent forms to explicitly address the possibility that data may be used for model training, but this remains ad hoc and inconsistent across the field. Consent forms in general do not come close to covering every ethical concern raised by current AI capabilities, and there is considerable variation in whether they are even reviewed by institutional bodies. We also note that consent forms serve a communicative function beyond their legal role: they signal intent to community members, and poorly written or overly broad forms can erode trust even when technically permissive.

A further complication is that some AI applications are much harder to explain to non-technical audiences than others. Translation is relatively intuitive; syntactic parsing or language modeling is not. When community members cannot meaningfully evaluate what they are consenting to, the ethical weight of that consent is degraded. More fundamentally, some language communities conceive of language in relational terms that make generative AI systems difficult to reason about in the way a human speaker can be reasoned about and held accountable—and it is not always clear who, if anyone, bears responsibility for what an LLM produces with community data (cf. Bird, 2024).

Open Access We find ourselves caught in tension between the scientific value of open data and the risks of making linguistic data freely available. On the one hand, open access enables reproducibility, promotes language visibility, and accelerates collective progress. On the other, once a dataset is formatted for easy use—say, for a shared task—it tends to be reused far beyond its original purpose, including for applications that may not have been anticipated at the time of collection. As one of us put it: "if data was sensitive five years ago, it is even more sensitive now" (cf. Junker, 2024).

This tension is sharpened in contexts involving oppressive governments, where linguistic data could be weaponized against minority communities, for instance by identifying individuals as speakers of a certain language or even revealing information that could be seen as politically sensitive. We acknowledge that such governments typically have other means of suppression available to them, but this does not make researchers less accountable for the data they make accessible. We also note that not every language needs an open-source dataset or a full data release: once HLT systems have been developed and validated, they can often be applied to new data without requiring that data to be publicly released. Keeping data local or on secure institutional infrastructure can mitigate some concerns—recent work on access control frameworks for language collections offers promising models (Foley et al., 2024)—but this imposes technical barriers: configuring local compute environments is nontrivial and may exclude precisely those researchers and communities who most need access.

We discussed but did not endorse the proposal that communities could sell their data as a way of gaining agency over its use. While superficially empowering, this places the burden of managing a complex business relationship on the community, presumes a single coherent community for each language, and risks reproducing the logic of allotment—atomizing collective resources for piecemeal extraction.

Synthetic Data One concrete proposal that generated significant debate was the use of synthetic data as a substitute for sensitive authentic data. The idea is appealing in principle: generate data that preserves the structural properties needed for model training while removing personally identifiable or culturally sensitive information, analogous to practices in medical informatics. However, we identify some serious risks.

Synthetic data could be confused for authentic data, and since it is likely to be of lower quality in at least some respects, it risks poisoning the already small data pools available for endan-

gered languages. It also threatens to take humans out of the equation in a domain where human involvement is often precisely the point—in some communities, speakers want to be cited and recognized for their words, and replacing their contributions with machine-generated approximations undermines this. We further note that synthetic data could be used for outright harmful purposes, such as generating fake Wikipedias or fraudulent language learning materials. While synthetic data may serve certain narrow ML objectives, we find that linguists are generally much less interested in it, because it is not natural human language, and its risks in a language documentation context are not yet well understood.

3.4 Data Analysis Issues

The fourth session turned to the question of how HLT systems—and LLMs in particular—should actually be used in the analysis of linguistic data.

Interpretability We discussed the familiar "black box" criticism of LLMs, but find that it matters less than one might expect for many language documentation tasks. For relatively constrained applications like annotation or glossing, what matters is whether the output is correct, not whether the system's internal reasoning is transparent. For tasks involving original linguistic analysis, however, the lack of interpretability becomes a real problem: if an LLM proposes a morphological parse or a syntactic generalization, the linguist needs to be able to evaluate not just the output but the basis for it, including the underlying data, and identify potential biases. We note that this is not unique to LLMs—most ML methods lack strong explainability—but the scope of what LLMs are being asked to do in language documentation is expanding rapidly, and the interpretability question becomes more pressing as these systems are applied to more analytically sensitive tasks.

We argue that at a minimum, any use of automated analysis in language documentation should clearly state what has and has not been manually checked, and explicitly acknowledge its limitations. Faulty automatic documentation is not merely unhelpful—it can cause harm. The literature on automation bias—the well-documented tendency to over-accept computer output as a heuristic replacement for careful evaluation (Goddard et al., 2012)—suggests that linguists working under time pressure may uncritically accept NLP-generated annotations, especially when outputs are fluent and plausible. We need explicit risk mitigation for both false positives and false negatives.

Should LLMs Write Grammars? One of the most spirited debates concerned the proposal that LLMs could, given enough data, produce an entire descriptive grammar end-to-end (Spencer and Kongborrirak, 2025). Some of us find this prospect exciting: language documentation faces a severe shortage of trained linguists and time, and even a rough automatically generated grammar might be better than no grammar at all. Others are deeply skeptical. Grammar is sufficiently nuanced that one could reasonably doubt whether an LLM, even a few years from now, would be able to notice everything a trained linguist would. And if such a grammar is produced, does it become the official grammar of the language? Can it even become a reference grammar if it is not the result of work between communities of speakers and scholars? In other words, who validates it?

We do not reach consensus on this question, but we note that the answer may depend on the community. Some speaker communities might welcome even a defective grammar for the visibility and legitimacy it confers—a poor grammar, like a dictionary, can have social and political value beyond its linguistic accuracy. Others might find it unacceptable. In general, there is a risk that automatically generated grammars may be mistaken for expert linguistic analysis if they are not clearly labeled as such; more subtly, they may also shape patterns of language use by implicitly legitimizing particular variants in communities with multiple dialects, especially under conditions of automation bias. What we do agree on is that humans must remain in the loop: an automatically generated grammar without human validation is not a scholarly output, and presenting it as one would be irresponsible.

Automating Is Not LLMizing It is common to suppose that "automating language documentation" amounts to "applying LLMs to language documentation", and participants took this to be an unwarranted conflation. Although recent LLM-based models have made progress in automated segmentation and glossing of languages with very limited data availability (Ginn et al., 2026), many tasks in language documentation are still well served by

non-LLM and even non-neural methods. For instance, linear-chain CRFs (Moeller and Hulden, 2018) and finite-state transducers (Beesley and Karttunen, 2003) are competitive for morphological analysis for many languages, and memory-based machine learning, Bayesian models, rule-based methods, and small (e.g. n-gram-based) language models can be effective in some settings (cf. Chirkova et al., 2025; Christian, 2025; Meelen and Griffiths, 2026).

Participants noted that this principle also applies to automated grammatical analysis. One participant described work employing hybrid methods that combine "black box" models with interpretable statistical and machine learning techniques, allowing more reliable partial grammatical descriptions and pedagogical materials from sparse or annotated data. Compared to LLMs, these methods can be less computationally intensive, more interpretable, and comparable in performance on the small datasets common in documentary work.

Yet the field's publication incentives push in the opposite direction. We observe that papers using "old" ML methods are increasingly difficult to publish: NLP venues consider them boring, and some participants described having work desk-rejected by computational linguistics journals for the mere fact that it did not involve neural models, irrespective of results. Participants found this to be regrettable, as performance is of primary concern for applications in language documentation, not internal mechanisms. While alternative publication venues are available (e.g. AI4CHIEF, ComputEL), publications in these venues may not be considered to be as relevant or prestigious for academic researchers in HLT, which presents a problem for career development. Participants expressed concern that the field is chasing fashion at the expense of practical utility, and that language documentation is particularly ill-served by this tendency, since the communities involved cannot afford to wait for the trendiest method to be made practical.

Choosing Models Responsibly The environmental cost of LLMs reinforces the case for methodological sobriety. Training and deploying large models consumes substantial energy (Strubell et al., 2019; Luccioni et al., 2023), and not all of that energy is clean—the carbon footprint of a model depends heavily on the electrical grid powering the hardware. We believe researchers working in language documentation should report their computational costs and select the smallest model adequate for the task. Few of the documentary linguists among us run models locally, but doing so is increasingly feasible and may help with both environmental impact and the data sovereignty concerns discussed earlier. More broadly, we see a need for practical guidelines on model selection for language documentation workflows—something like a Pareto analysis of performance against resource consumption, so that researchers can make informed choices rather than defaulting to the largest available model. Ideally, these practical considerations should be embedded in teaching HLT and NLP courses as well to raise awareness at an early stage.

4 Discussion

We notice one recurring pattern throughout our dialogue: on one side are the human language technologies, ripe for application in language documentation and revitalization, and on the other are communities and linguists willing and eager to apply them in their work. But what is missing lies in the middle: we lack a standard integration layer between HLT systems and documentary software tools; we lack resources and relationships necessary to deploy shared task systems for real use. The incentive structures of academia actively discourage the time-consuming, but crucial, bridging work: building an ELAN plugin does not earn a PhD, maintaining a keyboard does not lead to tenure, and publishing a glossing system based on a CRF rather than a transformer risks poor reviews. The people who could do this work are either not trained for it, not rewarded for it, or both.

A comprehensive solution to this problem might begin at the root, starting with incentive structures. However, we single out software as the most tractable issue to focus on in the short term. While people and incentive problems are important, we think the current software landscape is where the structural failure is most tangible and addressable.

ELAN and FLEx, the workhorses of documentary linguistics, were designed as standalone desktop applications with bespoke file formats and no native interface for external models. Integrating an HLT system into either tool today requires writing custom glue code, maintaining it against version changes, and distributing it outside any package manager—work that falls to whoever happens to

care enough, and that is abandoned the moment they move on.

Difficult as this problem is, we believe that if HLT researchers pooled their efforts under the right leadership, they could create a shared integration layer, providing facilities that make it as easy as possible for models to interoperate with these existing applications by means such as plugin APIs (e.g. ELAN's recognizer API) or direct file modifications. Such a shared integration layer, once realized, would allow any HLT researcher's model to reach any linguist's workflow without requiring each pair to reinvent the connection. It would also create a virtuous cycle: HLT researchers would gain a credible claim to real-world impact, because their systems would actually be reaching users, and linguists would no longer need to become or enlist a technological consultant in order to benefit from HLTs.

Moreover, we believe that in the long term, it may be preferable to build an "HLT-native" successor to apps such as ELAN and FLEx. As others have argued (Gessler, 2022), ELAN and FLEx are fundamentally limited in the extent to which they can interoperate with the full range of extant HLTs, and it is not feasible to take on the great task of retrofitting these apps for such capabilities. We therefore also submit that, even as human language technologists make efforts to integrate with ELAN and FLEx, they ought also to consider how they could combine efforts to build a new generation of apps to realize the full potential of HLTs in language documentation and revitalization.

5 Recommendations

Here, we synthesize our own observations with those of others and make a few concrete recommendations for how to address the missing middle.

Invest in reusable integration infrastructure. The field needs practically useful software products—not prototypes, not proofs of concept—that connect HLT systems to documentary workflows. This means building and maintaining the connective tissue: plugins, APIs, data pipelines, and user interfaces that make it possible for a documentary linguist to use an HLT system without becoming an HLT researcher. Presently, as noted above, the most impactful targets are ELAN and FLEx, which together constitute the de facto standard toolkit for language documentation. Both currently lack any native mechanism for invoking external models; a plugin architecture or standardized API that allowed, say, an automatic glossing model to be called from within FLEx's interlinearization workflow would immediately lower the barrier for dozens of existing HLT systems.

We therefore also recommend that the HLT community consider the question of how they might contribute to developing the successor(s) to these apps, which are now decades old, along the lines of proposals such as those outlined by Gessler (2022). While much more expensive to develop, a completely new design could thoroughly address the matter of how to stitch HLTs into a documentary workflow, while also addressing other perennial pain points, such as FLEx's poor support for platforms other than Windows.

We recognize that this work is expensive and unglamorous, and that grant-funded software often dies when the grant ends. We suggest that grant proposals for HLT research targeting language documentation should explicitly budget engineering effort for integration and deployment, and that funding bodies should consider supporting long-term software maintenance as a distinct funding category, analogous to infrastructure grants in the natural sciences. In many countries, the currently-available funding opportunities for linguists are limited to small grants only, but a maximum of, e.g. 10,000 GBP (for the British Academy Small Grant in the UK), is inadequate for covering engineering costs for development and maintenance of the required tools.

Redesign shared tasks for sustained engagement. Shared tasks are arguably the primary mechanism for directing HLT researcher attention toward problems in language documentation, but their transitory nature limits their impact. We recommend that future shared tasks be designed to require sustained contact between HLT researchers and the documentary linguists who provide the data—not just at the workshop where results are presented, but during the task itself. For example, a shared task on interlinear glossing could require participants to submit outputs for qualitative evaluation by the linguist who produced the training data, with a structured feedback round before the final submission deadline. The AmericasNLP shared task on educational materials (Chiruzzo et al., 2024) already points in this direction by targeting community-facing outputs rather than purely technical benchmarks; future itera-

tions could go further by embedding community evaluation into the task design and/or making using the least amount of computational (and therefore environmental) resources part of the task's aim. The overall goal is to make shared tasks a beginning of collaboration, not a substitute for it.

Push to recognize tools and software as research contributions. The academic incentive structure will not change overnight, but we can push for incremental progress. Building a tool that enables research is itself research, and making an existing system usable across a wider range of datasets and users is a genuine intellectual contribution. We urge tenure and promotion committees, journal editors, and conference organizers to treat well-engineered, well-documented software as a first-class research output—not a lesser category of work that must be laundered through a system-description paper to count.

Develop practical guidelines for model selection. Not every task in language documentation requires a large language model. Many tasks are well served by simpler, cheaper, more interpretable methods, and the field would benefit from practical guidance on when to use what. We envision something like a Pareto analysis of performance against resource consumption for common documentary tasks, so that researchers and practitioners can make informed choices rather than defaulting to the largest available model. Such guidelines would also help address the environmental costs of HLT research, which are nontrivial and unevenly distributed.

6 Conclusion

We are under no illusion that these recommendations are easy to implement. But we believe that the current trajectory—in which HLT for language documentation produces an ever-growing pile of research papers and an essentially static set of practical tools—is not sustainable. The communities whose languages are at risk cannot wait for academic incentive structures to reform themselves. The most useful thing we can do right now is start addressing the missing middle.

Limitations

This piece reflects the perspectives of approximately twenty researchers who participated in a structured discussion over three days. While the group included documentary linguists and human language technologists, it was not designed to be a representative sample of any of these fields, and notably did not include language community members as participants. Our observations are further shaped by the particular languages, regions, and institutional contexts with which we have experience. The "missing middle" diagnosis is offered as a unifying framework, not as an empirical claim validated by systematic evidence; we hope it proves useful for orienting future work, but acknowledge that others may diagnose the problem differently.

Ethical Considerations

We discuss ethical issues surrounding linguistic data, consent, and community engagement at length in the body of this paper. We note here that the recommendations we offer—particularly those concerning integration infrastructure and shared task design—carry their own ethical implications. Making it easier to connect HLT systems to documentary tools also makes it easier to apply those systems to data without adequate community consultation, and any integration infrastructure must therefore embed meaningful access controls and consent mechanisms rather than treating them as an afterthought. We also acknowledge the irony of a paper advocating for community voice that was written without direct community co-authorship, and we view this as a limitation of the present work rather than a model to follow.

Acknowledgments

The meeting *Automating language documentation*, held on September 17–19, 2025 in Uppsala, Sweden, was funded by the Riksbankens Jubileumsfond Research Initiation grant F24-0293 to Eline Visser. Antonios Anastasopoulos was partially supported by the US National Science Foundation under awards 2109578 and 2439202. Sharid Loáiciga has been supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. Marieke Meelen has been supported partially by the European Union (ERC, PaganTibet, 101097364) and the Endangered Language Documentation Programme (ELDP SG 0716).

We thank the other non-author participants

at this workshop for their contributions to the ideas in this piece: Harald Berthelsen, Rolando Coto-Solano, Harald Hammarström, Tatiana Korol, Joakim Nivre, Philipp Rönchen, and Daan van Esch.

References

Milind Agarwal and Antonios Anastasopoulos. 2025. AILLA-OCR: A first textual and structural post-OCR dataset for 8 indigenous languages of Latin America. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 120–127, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Milind Agarwal, Antonios Anastasopoulos, and Daisy Rosenblum. 2025. Developing a mixed-methods pipeline for community-oriented digitization of kwak'wala legacy texts. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 133–138, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Cristian Ahumada, Claudio Gutierrez, and Antonios Anastasopoulos. 2022. Educational tools for mapuzugun. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 183–196, Seattle, Washington. Association for Computational Linguistics.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford, CA.

Steven Bird. 2009. Natural language processing and linguistic fieldwork. *Computational Linguistics*, 35(3):469–474.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Steven Bird. 2024. Must NLP be extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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.

Lynnika Butler and Heather van Volkinburg. 2007. Review of FieldWorks Language Explorer (FLeX). *Language Documentation & Conservation*, 1(1):100–106.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):43.

Katia Chirkova, Rolando Coto-Solano, Rachael Griffiths, and Marieke Meelen. 2025. Comparing efficacy of ipa vs pinyin romanisation transcriptions for complex tonal languages: A case study in baima. In *The Eighth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–181.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Sebastien Christian. 2025. Enhancing grammatical documentation for endangered languages with graph-based meaning representation and loopy belief propagation. 12:100164.

Hannah Claus, Songbo Hu, Emre Isik, Anna Korhonen, Kitty Wenying Liu, and Marieke Meelen. 2026. Revitalising Endangered Languages and Cultural Heritage through Language Technology: A Pilot Study for Dzardzongke. In *Proceedings of the ComputEL workshop*.

Hilaria Cruz. 2022. Chatino Tonal Books Project. `https://ir.library.louisville.edu/chatino/`. Accessed March 30, 2026.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Nicholas Evans. 2009. *Dying words: Endangered languages and what they have to tell us*, volume 6. John Wiley & Sons.

First Nations Information Governance Centre. 2014. Ownership, control, access and possession (OCAP):

The path to First Nations information governance. Technical report, First Nations Information Governance Centre, Ottawa.

Darren Flavelle and Jordan Lachler. 2023. Strengthening relationships between indigenous communities, documentary linguists, and computational linguists in the era of NLP-assisted language revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 25–34, Dubrovnik, Croatia. Association for Computational Linguistics.

Ben Foley, Peter Sefton, Simon Musgrave, and Moises Sacal Bonequi. 2024. Access control framework for language collections. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 113–121, Torino, Italia. ELRA and ICCL.

Luke Gessler. 2022. Closing the NLP gap: Documentary linguistics and NLP need a shared software infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

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*.

Luke Gessler and Katharina von der Wense. 2024. NLP for language documentation: Two reasons for the gap between theory and practice. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

Michael Ginn, Lindia Tjuatja, Enora Rice, Ali Marashian, Maria Valentini, Jasmine Xu, Graham Neubig, and Alexis Palmer. 2026. Massively multilingual joint segmentation and glossing.

Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2012. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127.

Jeff Good, Julia Hirschberg, and Owen Rambow, editors. 2014. *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1):161–195.

Marie-Odile Junker. 2024. Data-mining and extraction: the gold rush of AI on indigenous languages. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Zoey Liu, Anneliese Richardson, Emily Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research*, 24(253):1–15.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Marieke Meelen and Rachael M. Griffiths. 2026. Historical Tibetan Normalisation: rule-based vs neural & n-gram LM methods for extremely low-resource languages. In *Proceedings of the AI4CHIEF conference, Paris, France - April 2026*.

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, 12:393–429.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 84–93, Santa Fe, New Mexico, USA.

Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, Shirley Anugrah Hayati, Antonios Anastasopoulos, Olga Zamaraeva, Emily Prud'hommeaux, Jennette Child, Sara Child, Rebecca Knowles, Sarah Moeller, Jeffrey Micher, Yiyuan Li, Sydney Zink, Mengzhou Xia, Roshan Sharma, and Patrick Littell. 2020. A summary of the first workshop on language technology for language documentation and revitalization. In *Proceedings of the 1st Joint Workshop on*

Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL).

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 11273–11285, Suzhou, China. Association for Computational Linguistics.

Piyapath T. Spencer and Nanthipat Kongborrirak. 2025. Can LLMs help create grammar?: Automating grammar creation for endangered languages with in-context learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10214–10227, Abu Dhabi, UAE. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Roberto Zariquiey, Arturo Oncevay, and Javier Vera. 2022. CLD-squared: Language documentation meets natural language processing for revitalising endangered languages. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.