Understanding the Gap: an Analysis of Research Collaborations in NLP and Language Documentation

Luke Gessler[1] **Alexis Palmer**[2] **Katharina von der Wense**[2,3]

[1]Indiana University Bloomington [2]University of Colorado Boulder
[3]Johannes Gutenberg University Mainz

lgessler@iu.edu, alexis.palmer@colorado.edu, katharina.kann@colorado.edu

Abstract

Despite over 20 years of NLP work explicitly intended for application in language documentation (LD), practical use of this work remains vanishingly scarce. This issue has been noted and discussed over the past 10 years, but without the benefit of data to inform the discourse. To address this lack in the literature, we present a survey- and interview-based analysis of the lack of adoption of NLP in LD, focusing on the matter of collaborations between documentary linguists and NLP researchers. Our data show support for ideas from previous work but also reveal the importance of little-discussed factors such as misaligned professional incentives, technical knowledge burdens, and LD software.

1 Introduction

Most of the world's languages may no longer be spoken by 2100 (Austin and Sallabank, 2011), and language documentation (LD) is the process of building a record of a language for purposes such as scientific analysis and language revitalization. In a typical LD project, many hours must be spent recording, transcribing, and analyzing texts. To increase their productivity, documentary linguists and language communities have often looked to language technologies, such as morphological parsers, for assistance. And in response, the natural language processing (NLP)[1] community has for over 20 years now been developing NLP systems specifically intended for LD.

However, despite great effort and interest in this area on both sides of the interdisciplinary boundary between LD and NLP, use of NLP in LD is still rare. While exact numbers are not known, it is easy to find both LD practitioners[2] who make little or no use of NLP in their work and NLP models and algorithms intended for use in LD which have never been used in a practical setting. This unrealized potential was identified early as an issue by those in the LD and NLP communities: indeed, the fact that "the technological landscape that supports [LD] remains fragmented, and the promises of new technology remain largely unfulfilled" was a major motivation for the organization of the inaugural ComputEL workshop (Good et al., 2014).

Subsequent works have further developed our understanding of this issue: e.g., Gessler (2022) cites software as a major impediment to adoption of NLP in LD, and Flavelle and Lachler (2023) emphasize the importance of fostering relationships between documentary linguists, communities, and NLP researchers. However, this literature is made mostly of position pieces—arguments are supported by the authors' authority as expert members of their respective academic communities rather than by data. We view this as a weakness: it is important to complement the existing literature with empirical evidence, especially in light of the stark lack of progress in this domain in the past decade.

In this work, we present an investigation of some central questions in the literature on NLP in LD, namely: (1) why interdisciplinary collaborations are so rare; (2) what motivates documentary linguists and NLP researchers to work with each other; and (3) what documentary linguists and NLP researchers think ought to be done in order to encourage more collaboration. These questions are ones that are repeatedly raised in the literature, and all three of them make reference to interdisciplinary partnerships, which have been identified as essential for progress. To investigate these questions, we take a mixed-methods approach, surveying documentary linguists and NLP researchers ($n = 49$) and further interviewing survey respondents ($n = 17$) in order to get a more detailed view of their experiences.

[1] We will not distinguish between NLP and computational linguistics in this work, calling both *NLP*. We will also regard speech processing technologies as belonging to *NLP*.

[2] Anyone involved in the LD process, whether they are, e.g., a documentary linguist or a language community member.

Findings of the Association for Computational Linguistics: ACL 2025, pages 867–877
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Our findings for the most part support previous discussions of issues in the domain of NLP in LD. However, they also bring to light important issues which have received less discussion, including the role of professional incentives in shaping collaborations, the degree to which NLP researchers underestimate the difficulty nontechnical users face in using their systems, and the unexpected way in which LD software is central in determining whether an NLP model will turn out to save users time in practice.

2 Previous Work

2.1 NLP in LD

Here we discuss some landmark works in NLP in LD; we must omit mention of many other works due to space. To our knowledge, the earliest work examining the application of NLP in LD is Kuhn and Mateo-Toledo (2004), in which a finite-state morphological parser, a part-of-speech tagger, and an *n*-gram language model were all evaluated for practical application in the documentation and revitalization of Q'anjob'al, a Mayan language of Guatemala, with mixed but promising results.

Two major developments in LD occurred in the next few years. First, two major apps used for conducting language documentation electronically were published: ELAN[3] (Wittenburg et al., 2006; Berez, 2007) and FLEx[4] (Moe, 2008; Rogers, 2010). Second, interest grew in the NLP community in assisting LD: one body of work epitomizing this interest is the tenth meeting of the Texas Linguistic Society[5] (Gaylord et al., 2006), organized on the theme of "computational linguistics of less-studied languages," featuring works exploring the utility of technologies such as grammar engineering, morphological parsers, and syntactic parsers in LD.

In the period that followed, ELAN and FLEx became established as the major apps for language documentation: one recent survey found that they were by far the most commonly used apps for LD among their respondents, with Toolbox being a distant third (Moeller, 2024). At the same time, interest in LD as an application domain continued to grow in the NLP community: while in 2006 NLP work on languages other than English was rare, it has become increasingly more common and accepted to work on other languages in the NLP academic community (see e.g. Joshi et al., 2020). An early landmark is Palmer et al. (2009)'s study of the practicality of performing interlinear glossing with the assistance of a morphological tagger.

NLP work on languages other than English has proliferated in recent years, both at general NLP venues and at purpose-specific venues such as AmericasNLP[6] (Mager et al., 2021) and ComputEL[7] (Good et al., 2014). However, many of these works have little or no relevance for LD: there are many non-English languages whose situations are not at all comparable to that of a language which is the subject of LD, and while NLP work which explicitly targets LD as an application has become somewhat more common, especially with the founding of an ACL interest group on endangered languages, SIGEL,[8] in 2021, it remains rare.

2.2 Commentary on NLP in LD

In light of the establishment of mainstream apps for LD and a flurry of recent work on NLP relevant for LD, one might expect that it would naturally follow that the LD and NLP communities would find ways of incorporating NLP into the documentary process. However, as many have noted, documentary linguists and language communities have not been able to use language technologies as much as they would like.

As we mentioned earlier, this was a major motivation for organizing the ComputEL workshop (Good et al., 2014), and many works since then have proposed ways forward for making NLP more accessible for documentary linguists and communities. These proposals have variously emphasized application software (Lane et al., 2021; Gessler, 2022); relationships between community members, linguists, and NLP researchers (Neubig et al., 2020; Liu et al., 2022; Flavelle and Lachler, 2023); and making existing NLP systems easier for documentary linguists to use (Foley et al., 2018; Esch et al., 2019; Neubig et al., 2019; Sheikh et al., 2024). Crucially, all these works do *not* attempt to empirically investigate the underutilization of NLP in LD—they present their solutions having made educated guesses at the nature of the problem they aim to solve. A major exception is Liu et al. (2022), in which 23 survey responses are collected from

[3] https://archive.mpi.nl/tla/elan

[4] https://software.sil.org/fieldworks/

[5] https://tls.ling.utexas.edu/2006/

[6] https://turing.iimas.unam.mx/americasnlp/

[7] https://computel-workshop.org/

[8] https://acl-sigel.github.io/

language teachers.[9]

We must also mention that another body of work adjoining this one reckons with the ways in which standard practices of documentary linguists and NLP researchers reproduce colonialist practices with indigenous language communities (Bird, 2020; Schwartz, 2022). One key theme in these works is that the overzealous deployment of technology where it is not wanted by communities can be harmful, and this can certainly be true. At the same time, as we will show in the present work, there are at least some language communities who have clear and unmet desires for language technologies and who want those technologies to be deployed in ways that would be routine and not require special consideration.

2.3 Language Documentation Apps

While FLEx and ELAN have been the most popular language documentation apps for the past 15 years or so, other apps have been created which attempt to serve the same basic need—primary data entry and analysis for LD—while addressing others as well. These apps were all motivated by perceived shortcomings of FLEx and ELAN that their creators wished to address.

Dunham (2014) presents an app which is focused on facilitating online collaboration among documentary linguists and allowing for integration of NLP systems such as morphological parsers which can partially automate the creation of interlinear glossed text (IGT). Bettinson and Bird (2017) describe prototype apps demonstrating the potential of purpose-specific apps which are characteristically deployed on mobile phones and target community members (instead of documentary linguists) as their intended users. Gessler (2022) presents a prototype app which aims to provide comprehensive integration of NLP models and extensible user interfaces. Hall (2022) develops an approach to software development which facilitates participatory design (Winschiers-Theophilus et al., 2010) in LD by making it easier for documentary linguists to build software. While none of these apps has gained widespread popularity, they demonstrate the conviction their authors have had that FLEx and ELAN could be substantially improved by the addition or enhancement of some other functionality.

3 Approach

How can we make NLP more practical in LD? While proposals vary, there is consensus on the fundamental importance of fostering relationships among documentary linguists, NLP researchers, and, most importantly, language communities. Flavelle and Lachler (2023) write of the "overall goal of bringing these three groups closer together, and strengthening the relationships that serve as the foundation to this work." But why are relationships so difficult to build between these populations to begin with? And how could we encourage them? While others have opined on these matters, we attempt to bring data to bear on them.

3.1 Scope

First, we limit our study's scope to only consider the relationship between two of the three key populations in LD: documentary linguists and NLP researchers. While this is perhaps the least important of the three relationships among the three populations, it is also likely the easiest to study: both NLP researchers and documentary linguists are considerably more homogeneous simply by virtue of their being members of academia. Examining the other two relationship pairs in this domain remains important, but we leave this for future work.

Additionally, in our study, we focus on NLP systems in the context of documentary linguist–NLP researcher relations and do not investigate the question of how the language communities they work with regard the use of NLP systems in their LD projects. Community attitudes towards NLP are of prime importance in this matter, of course, but the scope we have just defined precludes an investigation of them, and we expect that the dynamics at play between NLP researchers and documentary linguists must be at least somewhat separable.

In order to determine participant eligibility, we precisely define "documentary linguist" and "NLP researcher" as follows. A "documentary linguist" is anyone who has or is working towards a graduate-level degree in linguistics or a related discipline *and* has at some point taken part in "language documentation," which we in turn define classically as "collecting and analyzing language data with the production of dictionaries, grammars, archival texts, and other artifacts as a primary goal." An "NLP researcher" (for the purposes of this study)

[9] We would add to the above that there is one clear instance of a successfully-deployed language technology in LD which is not often discussed in this literature as such: the default morphological parser embedded in FLEx, XAMPLE (Black and Simons, 2006), a unification-based morphological parser.

is anyone who has or is working towards a graduate degree in computer science, computational linguistics, or similar and has published work on a system that was intended for use on either low-resource languages or in language documentation settings. We aim for our definitions to be as broadly inclusive as possible.[10]

3.2 Research Questions

In line with the consensus that has developed around the importance of relationships, we aim to address three main research questions (RQs) in this work: **RQ1:** What has impeded the formation of interdisciplinary collaborations between NLP researchers and documentary linguists? **RQ2:** What motivates documentary linguists and NLP researchers to work with each other? **RQ3:** What do documentary linguists and NLP researchers think ought to be done (if anything) in order to promote collaboration? While these questions are interrelated, we distinguish them by noting that RQ1 is focused on negative incentives and history, RQ2 is focused on positive incentives and the present, and RQ3 is focused on individual perspectives rather than observations of the field as a whole.

4 Methods

Our study has two components. First is a pair of surveys—one for NLP researchers and one for documentary linguists—which ask a mix of multiple-choice and free-response questions. Second is an interview: survey respondents are asked whether they would be interested in participating in a follow-up interview, and interview participants are recruited from respondents who indicated interest. Consent is obtained before both components of the study in line with our IRB's requirements. Participants are informed that their de-identified survey data and small portions of their interviews may be shared publicly. All materials used for conducting surveys and interviews may be found at github.com/lgessler/utg.

Survey Both surveys begin by prompting respondents to self-report on the eligibility criteria described in §3.1. If the respondent is eligible, then some additional background information is gathered (e.g., research interests) in order to contextualize the respondent's answers. Many of these questions are optional (e.g., country of residence, language families worked with) in order to give the participant the ability to remain anonymous.

Respondent	Surv.	Interv.
Ling., No System, No Collab	19	7
Ling., Yes System, No Collab	5	2
Ling., Yes System, Yes Collab	6	2
NLP, No LD Use	10	4
NLP, Yes LD Use	9	2

Table 1: Number of participants surveyed and interviewed. NLP researchers are grouped by whether they knew their systems being used. Documentary linguists are grouped by whether they used an NLP system, and those who did are grouped by whether they worked with an NLP researcher.

In order to ask different questions depending on whether a respondent has engaged in interdisciplinary work, we have two versions of each survey. For the NLP researchers, this criterion is whether any of their systems have been applied in any LD setting. For the documentary linguists, this criterion is whether they have ever used an NLP system in their work. Surveys are Google Forms and are distributed through professional networks such as the SIGEL mailing list.

Interview Survey respondents may indicate interest in a follow-up interview. Interested respondents are contacted and Zoom interviews with the lead author are scheduled. Interviews are structured around a topic guide (Knott et al., 2022) with around 10 questions prepared. For NLP researchers, just as in the survey, we ask different questions depending on whether the interviewee has had a system of theirs used in an LD setting. For documentary linguists, we ask different questions depending on whether (1) they have ever used an NLP system or worked with an NLP researcher in their work; (2) they have used an NLP system without collaborating with an NLP researcher; or (3) they have used an NLP system in collaboration with an NLP researcher.

5 Results

Here we describe the findings of the surveys. We synthesize these findings with the interview data in §6. Note that most questions are presented as Likert scale single-response items with 5 options ranging from "strongly disagree" to "strongly agree," with the middle option being neutral. We refer readers to the surveys for the full details of each question. The full, de-identified results of our surveys are

[10] While a few respondents meet the eligibility criteria for both populations, these cases are rare ($n = 4$), and the vast majority of respondents meet exactly one of the sets of criteria.

available at github.com/lgessler/utg. In total, 49 eligible respondents participated in the survey. See Table 1 for a summary of the responses.

Demographics The linguist respondents have over 40 languages of study represented among them, are based in 9 different countries, and are mostly (80%) employed as graduate students or tenure-track faculty members. They also have many years of experience doing fieldwork: the mean years of experience is 12.3 years, and the median is 11 years, with the most experienced respondent having 28 years of experience.

The NLP researcher respondents are based in 8 different countries and are likewise mostly (68%) either graduate students or tenure-track faculty members. 15 of the 19 respondents report 5 or more archival works in NLP venues. Expertise of the respondents is broad, representing 16 of the 25 research areas currently recognized by ACL.

5.1 Linguist Responses

Almost all linguists (29/30) demonstrate some awareness of the existence of NLP systems that can aid their LD work, and most (20/30) indicated that they could at least name one or two, with some (12/30) reporting even more familiarity.

Past System Use For linguists, systems used include finite-state morphological analyzers, FLEx's built-in morphological parsers, tokenizers, forced aligners, and speech recognition systems. System quality performed variably relative to expectations: 3/11 report having their expectations met, and the remaining 8 split evenly on whether their expectations were exceeded or disappointed. Asked whether their use of a system had "paid off" in terms of labor savings compared to doing the same work with no assistance from an NLP system, results are polarized: 1/11 choose the neutral option (break even), 6/11 report the system "appreciably" or "greatly" paid off, and 4/11 report that using the system yielded a net loss in productivity. It is interesting to note that in 4/6 of the positive cases, the respondent received assistance from an NLP researcher in using the system. Respondents agree that it was difficult to accommodate systems in their workflows: most say it was either "somewhat difficult" (6/11) or "very difficult" (2/11) to incorporate the NLP system into their workflow.

Prospective System Use For linguists who have not used an NLP system in their work, given a choice of common reasons why they have not done so, the most popular reasons are that they are unsure of how to set them up and integrate them into their workflows (13/19), that they lack appropriate hardware (6/19), and that they doubt it would be worth the effort (6/19). In an optional follow-on free response, several (4/10) respondents doubt that they had enough data to train an NLP model of sufficient quality.

Collaboration All linguist respondents are also asked about interest in collaboration with an NLP researcher. The response is very positive: 15/30 are "very interested", and 9/30 are "somewhat interested". Only 3/30 choose "neutral", and 3/30 choose "somewhat uninterested". No respondent chooses the strong "not at all interested".

5.2 NLP Researcher Responses

10/19 respondents report that their work was done in the context of a relationship with a language community that has lasted for more than one publication cycle. 14/19 respondents report having a documentary linguist co-author on at least one of their works, and only one respondent reports not being familiar with LD.

Outlook on NLP in LD Asked about the impact of NLP in LD up to the present, most respondents think NLP has had "little" (1/19), "limited" (9/19), or "moderate" (3/19) impact, with a minority (3/19) thinking it has achieved "sizeable impact in many settings". Asked the same about the future, responses are very positive: some (3/19) predict moderate impact, while the majority predicts "sizeable" (10/19) or "great" (6/19) impact.

Past System Use Of the respondents whose systems have been applied in real work, 7/8 were also responsible for the operation of the system in the application. Tellingly, when those 7 are asked whether the system could have been deployed without them, all 7 answered "no". In 6/8 cases, respondents report that manual intervention was required to move data between the NLP system and the software being used to support LD. Respondents are mildly positive about the impact of their system: 4/8 feel that, comparing time investment to time savings, the project broke even, and 2/8 feel that there was "moderate benefit" in excess of time invested. Of the remaining 2, one states that they do not know the answer, and one feels that investing in the system was a moderate net loss of time.

Prospective System Use Of the 11 respondents who have created a relevant NLP system that has never been applied, most (6/11) believe their systems would be usable by someone without a technical background, though others report it would be "fairly difficult" (3/11) or "quite difficult" (1/11), with only one expecting it would be "very easy." Asked how they expected data would flow between the NLP system and the LD app, 6/11 expect that it would be integrated into the LD app, while 4/11 expect that manual effort on the user's part would be needed to move data back and forth. The respondents are optimistic about potential benefits in terms of time invested vs. time saved: 1/11 expects the system would be a poor investment, while 3/11 expect a balanced return, 5/11 expect "moderate" benefit, and 2/11 expect benefit "well beyond" the time investment. Respondents are universally interested in collaborating with documentary linguists, with 9/11 choosing the strongest option "very interested," and 2/11 choosing "fairly interested."

6 Discussion

Five major themes emerge from the whole of our data. While some of these themes are in line with the existing, opinion-based literature, others are not yet well represented in existing literature. We structure our discussion by discussing each major theme, formulated as a thesis, in turn before concluding.

Thesis 1: NLP researchers and documentary linguists must find ways to align their professional incentives. The research activities that the fields of NLP and LD reward are very different. Small publications, each taken from start to finish in the span of months, are the norm in NLP research, while language documentation research spans years and even careers and is published in journal and book-length publications. A consequence of this is that NLP researchers are incentivized to work on many projects which are often discontinued after they result in a publication. The matter is exacerbated by the trend-driven nature of research in NLP—one NLP researcher interviewee, explaining the historical and enduring lack of research on endangered languages, puts it bluntly:

> *For computer scientists [...] it might not be very attractive to work with linguists because linguists are almost looked down upon a bit. And then, the languages themselves—no-one would say it, but in reality, it's like "oh, it's just a language that nobody cares about." [...] So it's not very attractive to say, as a computer scientist, "I work on languages that nobody cares about." And it's more attractive to work on ChatGPT for English.*

The relevance of an individual language to the research agenda of the NLP community has a strong effect on NLP researchers' interest in engaging with it, and, while in recent years the situation has improved, endangered languages still are often not as professionally rewarding to work on.

In turn, some documentary linguists described their difficulties in working with NLP researchers. Documentary linguists succeed by cultivating years-long collaborations studying under-studied (and therefore probably under-resourced) languages, and this timescale comes into tension with the goals of NLP researchers. One linguist survey respondent, asked how they would feel if an NLP researcher approached them about a collaboration, describes the lack of reciprocity in a past collaboration, which we view as a consequence (in part) of the NLP researcher's lack of professional incentive to provide things that are valuable to a documentary linguist in a collaboration:

> *The person approaching me would need to be prepared to do significant work—not just ask me to do a bunch of work for them, which is what almost always seems to happen when I get involved with [an] NLP project. I put way more into it than I get out of it.*

Another, describing interactions with NLP researchers, is quite positive on the whole, but notes that NLP researchers often have narrow interests:

> *Since I began interacting with colleagues in this field (in 2013) I have had quite a few contacts with NLP colleagues, many of them interesting. [...] Some quickly lead to the conclusion that I am not the right person to participate, as the NLP team looks for a very specific scenario in terms of amount of resources or type of resource and my dataset is not a good match.*

If the academic cultures of these fields are the problem, then it is beyond any individual's power to change, and the way forward is to envision how to frame interdisciplinary collaborations so that the professional expectations of each party's disciplines can be met, which then ought to make collaborations much more productive and durable. The first researcher we quoted above describes shared tasks as a potential avenue for this:

> *But I think there are computer scientists interested in [collaborating], if they think, "I understand the problem. I have ideas on how to do it." Because otherwise, if it's not very attractive, why should you care about it and think about*

> *how we could help them? [...] If you, for example, are a linguist, you could organize a shared task that goes into this direction. This is a way of connecting, because the computer scientists can just say, "Oh, I like this task!" and they don't really have to discuss it with others. And in the end you can see who's done something potentially useful.*

Beyond shared tasks, we expect there are other ways for either group to work creatively within the confines of their professional cultures to give the other group more of what it needs from "research activity." For example, some grant programs such as the NSF/NEH's DLI–DEL, while primarily serving documentary linguist PIs, are supportive of work that could also constitute NLP research. This is an opportunity for documentary linguists to think about how to structure grants for programs such as the DLI–DEL so that NLP researchers may participate and support the documentary effort while still meeting their professional needs.

Thesis 2: NLP researchers must treat social considerations as a primary concern. We find support for a position that Schwartz (2022) and others have argued for: that NLP researchers working in language documentation (and revitalization) are often unaware of the social and historical context of their activity with language communities. Within this study, we find that some (though not all) documentary linguists have reservations about the specific terms on which a collaboration with an NLP researcher would take place, with these reservations most often centering on possible harms that could come to the language communities that they work with.

Recall the respondent above who writes that their collaborator "would need to [...] not just ask me to do a bunch of work for them." Another respondent writes that they are interested in exploring collaborations, though with some "ethical concerns regarding the appropriation of Indigenous language and knowledge." Some linguist interviewees also affirm this position. One says:

> *I certainly sometimes feel a little bit of suspicion toward NLP researchers [...] about whether they understand the social circumstances of language documentation or revitalization. [...] It's very important to my participants that what I'm doing not be used for anything commercial, and I think that NLP researchers could probably understand that. But I don't necessarily know that they would engage in all the data protection I would like them to.*

NLP researcher interviewees also acknowledged the importance of this issue. One interviewee narrates a project in which they were assisting an indigenous community with gaining internet access:

> *They [the language community] had a conflict inside about whether their kids should access, or not, the internet. And because of that, they came to us and said, "I think we should stop what we're doing until the community figures out what the high-level thing they want to do is." And [while] we were trying to help them go to the internet [...], we said, "okay, you want us to stop, no problem."*

Explaining further that in the 7 months that had passed (at time of interview) since this happened, the community had still said nothing about resuming the project, the interviewee says that this was a setback, but that the community's will needs to be respected:

> *I'm employed by an organization that evaluates itself every 3 months. That's the culture of the New York Stock Exchange, okay? [But] it's part of how we work in this context. [...] You need to have not just a plan B, but a plan C, D, E—because it's complicated, like any interdisciplinary work.*

This interviewee's vignette emphasizes the importance, in assessing a project, of budgeting for setbacks—on all sides, but especially on the part of the NLP researchers, who often feel pressure to finish and publish work quickly.

Thesis 3: Both documentary linguists and NLP researchers must invest time into understanding each other. This broad theme has been discussed at length in previous work (e.g., Flavelle and Lachler, 2023), and in our data there is strong consensus on both sides of this interdisciplinary boundary that collaborations cannot succeed without conscious effort to understand the other parties' methods, needs, and values. One challenge within this domain many of our participants discuss is joint decision making. One NLP researcher interviewee says:

> *When we do interdisciplinary work, you have not only to score goals, but you have to discuss values. [...] The hardest disagreements happen when both sides will think they are doing the right thing—and they are, according to their values.*

One interviewee, asked how the two communities could better understand each other, says:

> *There are no easy answers. [...] It's really equivalent to the question of how you raise your kids and keep them all so that they're not constantly on their screens. So how do you fix that? I don't know. Get out of the water and get a boat. Just swimming against the current clearly isn't going to get you anywhere. [...] It's going to exhaust anybody you're trying to pull along with you.*

We view this as an indication that the conversations that have already been happening in LD and NLP venues on the challenges of fostering interdisciplinary understanding in these two fields ought to continue, as there is as yet still little idea on how, concretely, to achieve it.

Thesis 4: Documentary linguists cannot use many NLP models, and NLP researchers need to take downstream usability seriously. Surveyed linguists who have never used an NLP system in their work indicate in 13/19 cases that a lack of technical knowledge is preventing them from looking into how to use them. Moreover, of the 11 linguists who have used a system in their work, a majority (7/11) relied on a collaborator to set up the system, and a total of 8/11 report that it was "somewhat difficult" (6/11) or "very difficult" (2/11) to incorporate the system into their workflow.

One linguist, a mid-stage Ph.D. student, made an ultimately failed attempt to use an ASR system to transcribe their field recordings and details the difficulties they encountered attempting to get the system to run without computational expertise:

> *I ended up getting to the point where the error rate looked pretty decent [...] in TensorBoard. But when I exported that model, I used [...] a held-out audio file, and I tried to generate transcriptions for that, and then it just really, spectacularly failed. And I actually still don't know exactly why it just wasn't working on that new file. [...] I was reading more studies of people applying ASR to different endangered languages, and I saw that there's a decent number of people who [reported] it basically just didn't work.*

Two other linguist interviewees also report similar experiences of ASR models that were difficult to run and produced disappointing results.

NLP researcher respondents do demonstrate awareness of this difficulty, though they still seem to underestimate how hard it is for others to use their models. In the survey, most NLP researchers who have not had their systems used before believe that it would be "doable" (6/11) or "fairly easy" (1/11) for a documentary linguist to use their model. NLP researcher interviewees also express concern about the usability of their products in the hands of documentary linguists, but none of them discuss any steps that they would take to substantially improve the approachability of their models.

In sum, our data show that there is a significant disconnect between NLP researchers' perceptions of the amount of technical expertise required to use their tools and the realities of what happens when a lone documentary linguist attempts to use an NLP model. This disconnect has persisted despite the great increase in activity in this area from the NLP community in the past several years.

We would encourage NLP researchers to pause before claiming in their works that a given model could, for example, "allow documentary linguists to be more productive," or "provide better ASR for endangered languages." What are the exact conditions under which these claims of impact are true? And do these conditions necessitate the involvement of a highly-trained computer scientist? If so, is it right to claim nonetheless that they have great impact potential for language documentation?

Thesis 5: NLP models must be integrated into language documentation software in order to be practical to use. No matter how good an NLP model's outputs are, it will not be effective in a LD setting unless it can *save time*, considering both the initial cost of setup and the ongoing cost of interacting with a model in a documentary workflow. Early works on NLP in LD recognized this (e.g., Palmer et al., 2009) and accordingly evaluated models not just by the correctness of their outputs but also by the productivity (measured in terms of, e.g., morphemes glossed per minute) of the subjects using the model in a workflow.

Regrettably, we find much support in our data for the conclusion that NLP models are at present very impractical to use in LD settings, with one major reason being that it is not easy or in some ways even possible to integrate models into workflows to the extent that would be necessary to make the model-assisted workflow productive. As we saw earlier, a majority of linguists who have and have not used an NLP system in their work before are unsure of how to set them up and expect systems would be difficult to integrate into their workflows.

Interviewees also acknowledge this directly. Speaking about using a morphological parser in concert with ELAN, one linguist interviewee says:

> *My personal ideal solution to this is that ELAN integrates an actual parser that hooks up to a lexicon in the same way that FLEx does, but you don't need to move the data back and forth between applications in order to assign it a gloss.*

Critical to note here is the extra work of moving data between an app like ELAN and the NLP system. Indeed, with the exception of FLEx's integrated morphological parsers, this is what working with an NLP model must look like in most cases, and it incurs a significant time cost. A solution to this problem would be, as the interviewee suggests, to work towards a model where the LD app talks directly to an NLP system.

A second theme which indirectly supports this thesis is the quality of the LD software itself. If NLP models are to tightly integrate with software, then the model's ultimate impact could be dampened if the software itself were making workflows impractical. One linguist says:

> *Fundamentally what I need is a better user interface. I need FLEx to not suck. It'd be great if I didn't have to auto-approve the glosses. [...] Now, if the person that's approaching me is just like, "Hello, I'm a software engineer, and I would like to contribute." I have a lot of work for that person. Let's improve a shitty app.*

We would hasten to add that while it is easy to find myriad ways in which any piece of application software falls short of perfection, this sentiment nonetheless demonstrates the extent to which existing LD software fails to serve the basic workflow needs of documentary linguists even before NLP models enter into the equation.

We therefore believe, in line with some previous work in the literature, that improving the quality of LD software—both on its own terms, and in terms of facilities for integrating with NLP models—is of prime importance if NLP models are ever to see widespread adoption in LD.

7 Conclusion

We conclude by revisiting our research questions. On RQ1, we find support for the already well-discussed issues of: a) NLP researchers not demonstrating adequate awareness of the social context of LD and b) disciplinary and technical knowledge gaps. We also find two new issues which have received little previous discussion.

First, misaligned professional incentives inhibit collaborations between documentary linguists and NLP researchers. NLP researchers' focus on many small projects clashes with the community-centered approach of documentary linguists. Finding creative ways to accommodate this and other "cultural differences" between the two fields would do much to facilitate collaborations.

Second, we find support for the argument that NLP systems *cannot* be practical unless the software being used to support the LD workflow, such as FLEx and ELAN, harmonizes the three-way interaction between user, LD software, and NLP system. This matter has been raised before (Gessler, 2022), but as yet there is no software which provides such integration in a general way for any existing NLP system, which underscores the importance of developing such software in order to make the use of NLP systems in LD more practical.

For RQ2, our data confirm a strong mutual interest between NLP researchers and documentary linguists, despite the many obstacles which we have just outlined. Linguists' interest is driven by the hope of better productivity for their projects and useful language technologies for communities they work with, and NLP researchers are motivated by the application domain and by access to new languages and datasets. These motivations have already been noted in previous work, and our work provides an empirical basis for them.

On RQ3, the personal narratives we elicited in both interviews and survey responses mostly followed respondents' reflections about field-wide dynamics, though they also revealed that these trends do not always hold. For example, while many language communities have at least moderate reservations about how their data is used with language technologies, one interviewee noted that their community seemed to have almost none. The patterns we have identified are not universal, and so may not be their solutions.

Limitations

The robustness of our findings are limited by many factors endemic to survey- and interview-based research, including the subjectivity of analyzing interview transcripts. Additionally, while we believe we were successful in getting a fairly high response rate in our survey data, our two target populations are small in absolute numbers, which also led to a relatively small sample size and limits the robustness of the numerical trends we find in the survey.

Acknowledgments

First, we thank all 49 of our participants for the time they took to complete our survey and speak with us in interviews. Additionally, we thank the members of the NALA Group at CU Boulder for their helpful feedback on this work as it was progressing. Finally, we thank our ACL Rolling Review reviewers for their helpful comments on this work.

References

Peter K. Austin and Julia Sallabank. 2011. Introduction. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages*, Cambridge Handbooks in Language and Linguistics, pages 1–24. Cambridge University Press, Cambridge.

Andrea L. Berez. 2007. Review of EUDICO Linguistic Annotator (ELAN). *Language Documentation & Conservation*, 1. Publisher: University of Hawai'i Press.

Mat Bettinson and Steven Bird. 2017. Developing a suite of mobile applications for collaborative language documentation. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 156–164, Honolulu. Association for Computational Linguistics.

Steven Bird. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

H Andrew Black and Gary F Simons. 2006. The SIL FieldWorks Language Explorer Approach to Morphological Parsing. In *The Proceedings of the Texas Linguistics Society X: Conference Computational Linguistics for Less-Studied Languages*, Austin, Texas.

Joel Robert William Dunham. 2014. *The online linguistic database : software for linguistic fieldwork*. Ph.D. thesis, University of British Columbia.

Daan van Esch, Ben Foley, and Nay San. 2019. Future Directions in Technological Support for Language Documentation. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1.

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, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 205–209. ISCA.

Nick Gaylord, Alexis Palmer, and Elias Ponvert. 2006. TLSX: The Proceedings. In *The Proceedings of the Texas Linguistics Society X: Conference Computational Linguistics for Less-Studied Languages*, Austin, Texas. CSLI Publications.

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*, pages 119–126, Dublin, Ireland. Association for Computational Linguistics.

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.

Patrick J. Hall. 2022. *Participatory Design in Digital Language Documentation: A Web Platform Approach*. Ph.D. thesis, UC Santa Barbara.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Eleanor Knott, Aliya Hamid Rao, Kate Summers, and Chana Teeger. 2022. Interviews in the social sciences. *Nature Reviews Methods Primers*, 2(1):1–15. Publisher: Nature Publishing Group.

Jonas Kuhn and B'alam Mateo-Toledo. 2004. Applying computational linguistic techniques in a documentary project for Q'anjob'al (Mayan, Guatemala). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

William Lane, Mat Bettinson, and Steven Bird. 2021. A computational model for interactive transcription. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 105–111, Online. Association for Computational Linguistics.

Zoey Liu, Crystal Richardson, Richard 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)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors. 2021. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Association for Computational Linguistics, Online.

Ronald Moe. 2008. FieldWorks Language Explorer 1.0. *SIL Forum for Language Fieldwork*.

Sarah Moeller. 2024. Personal communication.

Graham Neubig, Patrick Littell, Chian-Yu Chen, Jean Lee, Zirui Li, Yu-Hsiang Lin, and Yuyan Zhang. 2019. Towards a General-Purpose Linguistic Annotation Backend. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, Honolulu, Hawaii. Association for Computational Linguistics. ArXiv: 1812.05272.

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 S 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)*, pages 342–351, Marseille, France. European Language Resources association.

Alexis Palmer, Taesun Moon, and Jason Baldridge. 2009. Evaluating Automation Strategies in Language Documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44, Boulder, Colorado. Association for Computational Linguistics.

Chris Rogers. 2010. Review of Fieldworks Language Explorer (FLEx) 3.0. *Language Documentation & Conservation*, 4. Publisher: University of Hawai'i Press.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. Association for Computational Linguistics.

Zaid Sheikh, Antonios Anastasopoulos, Shruti Rijhwani, Lindia Tjuatja, Robbie Jimerson, and Graham Neubig. 2024. CMULAB: An Open-Source Framework for Training and Deployment of Natural Language Processing Models. ArXiv:2404.02408 [cs].

Heike Winschiers-Theophilus, Shilumbe Chivuno-Kuria, Gereon Koch Kapuire, Nicola J. Bidwell, and Edwin Blake. 2010. Being participated: a community approach. In *Proceedings of the 11th Biennial Participatory Design Conference*, PDC '10, pages 1–10, New York, NY, USA. Association for Computing Machinery.

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).