# Computational Language Documentation: Designing a Modular Annotation and Data Management Tool for Cross-cultural Applicability

**Alexandra O'Neil** and **Daniel Swanson** and **Shobhana Lakshmi Chelliah**
aconeil, dangswan, schellia @iu.edu
Indiana University
Bloomington, IN, USA

## Abstract

While developing computational language documentation tools, researchers must center the role of language communities in the process by carefully reflecting on and designing tools to support the varying needs and priorities of different language communities. This paper provides an example of how cross-cultural considerations discussed in literature about language documentation, data sovereignty, and community-led documentation projects can motivate the design of a computational language documentation tool by reflecting on our design process as we work towards developing an annotation and data management tool. We identify three recurring themes for cross-cultural consideration in the literature - Linguistic Sovereignty, Cultural Specificity, and Reciprocity - and present eight essential features for an annotation and data management tool that reflect these themes.

## 1 Introduction

Although rapid advances in language technology have been made in the last few decades, these advances have largely benefited speakers of global majority languages (Brinklow, 2021). In addition to population-based divides in technology availability, the delineation between well-resourced and low-resourced languages is connected to modern and historical socio-economic power dynamics, with resources for languages being reflective of the relative dominance of groups at the expense of others (Kuhn et al., 2020). In addition to the disproportionate availability of language technology for documented languages, advances in language technology have yet to significantly benefit those working on documenting languages, meaning that access to language technology is minimal, if not nonexistent, for languages that are currently undergoing the process of documentation. Language technology is used as an inclusive term that describes both the technology that can help with the documentation and analysis process and the technology that the community can use to interact with, support, and teach their language.

While language documentation processes vary vastly amongst different communities, the prototypical process normally involves a linguist and one or more members of a language community. The linguist works with the language community to gain a better understanding of the language by collecting data from the speakers. This data normally includes recordings from the speakers and annotations of the recordings, often as transcriptions in IPA or the language's orthography. Throughout the process, there is typically a multitude of tools used to make the recordings, annotate the data regarding various features, analyze these annotations, and create a resource for the community. The process of transcribing audio is usually identified as one of the most time-consuming parts of the process, a problem referred to as the "transcription bottleneck." However, the next steps of analysis and resource development are equally, if not more, time-consuming. During analysis and resource development, the linguist often continues to consult with the language community to ensure correct analysis of the language and applicability of the developed resource. With the advent of computational linguistics, computational linguists are now often included in these last steps of annotation, analysis, and resource development.

Recent advances in language technology present an opportunity for expediting the process of language documentation and reducing the inequity of access to language technology, specifically through the development of language documentation tools. In order to maximize the utility of such a tool, it is essential to consider the varying cultural considerations that are present in the different contexts in which the tool might be used. We identify three integral steps in the process of creating a cross-culturally applicable tool for language documen-



*Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 107–116

tation: design, collaboration in communities, and feedback integration. While the rest of this paper focuses on the design step of the process, the effectiveness of an intentional design is mitigated if it is not followed by collaboration in communities and feedback integration. Collaboration in communities should involve discussions with community members, activists, and language documentarians from various language communities about ethics, functionality, and risks of language technologies, and project outcomes. The subsequent step to collaboration is the integration of this feedback into the developing tool.

We intend to further cover and demonstrate all three of these steps in future work, but this paper details our experiences with the first step: design. This paper provides a case study for designing a cross-culturally applicable tool by presenting how this process has been realized in the design phase of our own language documentation tool. Our approach demonstrates how innovative research in language documentation, data sovereignty, and community-led technology development can be used as the foundation for the design of an annotation and data management tool. In section 2 we describe existing annotation and data management tools and how our tool compares. Section 3 uses discussions of linguistic sovereignty, cultural specificity, and reciprocity to frame critical cross-cultural considerations that inspire the eight features that are described in section 4. In section 5 we conclude by discussing the benefits of integrating cross-cultural considerations into a project during the design process.

## 2 Related Work

The field of language documentation currently includes tools for assisting with transcription, annotation, and data management, as well as a series of recent attempts at developing more advanced versions of these tools. This section briefly describes the most popular tools, including the strengths and limitations of the various features. Next, we describe novel approaches and further elaborate on the motivation for prioritizing cross-cultural considerations in the design process. The goal of this section is to provide a better understanding of what an annotation and data management tool encapsulates, before describing how cross-cultural considerations (section 3) motivate particular features in the design of such a tool (section 4).

### 2.1 Popular Annotation and Data Management Tools

While there are a multitude of tools and derivations of annotation and data management tools available, we highlight the two most popular: Fieldworks Language Explorer (FLEx) and EUDICO Linguistic Annotator (ELAN). While we offer a critical review of the platforms, both provide an exceptional example for the future of language documentation, as they promote accessibility through free and accessible applications. Other improvements of these tools are available but often include an associated fee and proprietary code, which diminish their utility in the language documentation, as discussed in section 4.6.

#### 2.1.1 Fieldworks Language Explorer

FLEx is a commonly used lexicography tool in language documentation (Black and Simons, 2006), likely due to the fact that it is both free and includes an adequate graphical user interface. The tool allows for the creation and refinement of a lexicon, as well as glossing and analysis of texts. The lexicon section offers a large, but predetermined, selection of tiers for providing additional information about an entry, such as the inclusion of multiple senses, allomorphs, variants, and usage notes. The texts and words section allows users to import stories and other narrative transcriptions with the ability to analyze the text by providing nested morphological segmentation and derivation, bilingual glossing, and part-of-speech tagging.

FLEx has features to help with the generation of a language's grammar and various other levels of linguistic description, like customizable lists detailing dialectal variation, morpheme types, and semantic domains. However, the interface of FLEx is complicated for non-linguists and those without extensive training in lexicography tools. Additionally, advanced, but extremely useful features, like automatic parsing using existing segmentations, often cause the tool to crash and importation of other non-FLEx formats is lossy. For example, FLEx is not consistently able to import morphological segmentation encoded in other linguistic annotation file formats, like SFM files, without prior explicit cross-references in the lexicon. Further, collaboration between multiple parties requires cumbersome sending and receiving of database backup files and cannot be done synchronously. That being said, automatic parsing suggestion, querying of texts by feature, and intricate layers of annotation are

notable contributions of FLEx that should offer inspiration for future data management tools.

The tool supports automatic export into web and dictionary platforms, well-aligned with the ideas of reciprocity discussed in 3.3. However, as FLEx was developed for the purpose of bible translation, it has extremely limited functionality for integrating audio during the analysis process. The data from speakers is transcribed (annotated in IPA or an orthography) and then moves into FLEx for analysis. In order to contribute to analysis in this step, the contributor needs to be able to understand the written transcription of the language and the features presented in FLEx. Failing to account for cultural specificity by confining the representation of the language to a written form excludes the involvement of many speakers from oral language cultures. For example, speakers may not participate if they feel uncomfortable with the abstraction of their language into an unfamiliar writing system with no auditory representation.

#### 2.1.2 EUDICO Linguistic Annotator

ELAN is a documentation tool focused on speech transcription, the process of representing a speech signal with writing,[1] and includes the ability to flexibly create multiple tiers with customizable hierarchical relations while playing a recorded segment of audio (Wittenburg et al., 2006). Additionally, users can configure the view to focus annotation efforts. Particularly useful for those with phonetic training, the audio clip can be displayed alongside a spectrogram, a visual representation of speech that encodes speech signal frequencies and can be used for phoneme identification and analysis (Zue and Cole, 1979). As ELAN was originally created for transcription of signed speech with multiple interlocutors, data management on a self-referencing language-documentation level is minimal. However, flexible tier creation, configurable displays, and the spectrogram presentation and replay of recorded audio are indispensable aspects of the tool for many with a background in linguistics.

The user interface of ELAN is well-suited to linguists and those with high computer literacy, but otherwise requires training. The flexible tier creation of the tool and representation of audio support the ability of users to develop culturally specific projects. However, the tool presents issues for linguistic sovereignty and reciprocity due to the challenging interface. Linguistic sovereignty, further defined in 3.1 encapsulates the ability of community members to understand and participate in the research that is being done on their language, but the interface of ELAN is designed for a user with high computer literacy and a background in linguistics. This further endangers the ability of a project to be reciprocal, as it prioritizes academic access and understanding of annotated language data over community access.

[1]Transcription is commonly performed using the international phonetic alphabet (IPA) or an orthography of the language

### 2.2 Novel Approaches

While there have been many attempts to create improved language documentation tools, we present two projects that are working towards an annotation and data management tool but are still developing. These two projects are noteworthy in that both are open-sourced and provide a demo version that allows interested individuals to participate and comment on the development of the tools. We hope that these similar developments of computational language documentation tools can support each other and work together to positively impact those working in language documentation. The cross-cultural applicability of these approaches is not evaluated as the projects are still developing. That being said, the utility of this paper lies in the explication of how cross-cultural considerations define the features that are prioritized in the development of our tool.

#### 2.2.1 Linguistic Field Data Management and Analysis System

The Linguistic Field Data Management and Analysis System (LiFE), is a language documentation annotation and data management program with a user interface aimed at linguists, with the goal of aiding language documentation efforts by integrating various NLP libraries (Singh et al., 2022). This tool focuses on making various advancements in computational linguistics available to documentary linguists without a computational background. The research also provides extensive background on the development of language documentation tools and offers conversion of in-tool annotation to facilitate integration with other NLP tools.

#### 2.2.2 Glam

Glam is another annotation and data management tool aimed at improving the experience of those in the field of language documentation while inte-

grating advancements in NLP (Gessler, 2022). The presentation of this tool defines two features intrinsic to the design of a successful annotation tool: interlinear text annotation and lexicon development. The project also highlights the importance of cross-discipline collaboration in the development of an annotation tool.

### 2.3 Designing a Tool

Similar to the other projects presented here, we recognize the limitations of existing language documentation tools and the great potential of developments in the field of computational linguistics. Existing approaches center two contributors: computational linguists and documentary linguists. However, the field is currently neglecting who should be acknowledged and prioritized as the main contributor in language documentation: the language community. This is evidenced by the marginalization of the role of language communities in the presentation of these tools. Our remedy to this problem is proposing a novel, yet simple, approach that consults existing literature in the fields of language documentation, data sovereignty, and computational linguistics, with a focus on highlighting research by Indigenous scholars. This approach in developing language documentation tools is not sufficient without further consultation with language communities but provides a basis for design prior to the necessary steps of collaboration in communities and feedback integration.

## 3 Cross-Cultural Considerations

Recent work on computational language documentation has attempted to understand how documentary linguists, community members, and computational linguists can best support each other (Flavelle and Lachler, 2023; Lu et al., 2024; Wiechetek et al., 2024). Collaborative work between these three groups has great potential to be mutually beneficial, as expertise from each group can guide the development of tools and documentation to maximize their impact. However, existing scholarship prioritizes the role of documentary linguists and computational linguists in the design of technology, which often results in either minimizing cross-cultural differences in a way that neglects recording information that is important to a community or produces a tool that works for a specific purpose, but is hard to extend to use in other communities.

The challenge in designing an annotation and data management tool lies in the ability to support linguistic sovereignty, flexibly adapt to varying needs and ethics of language communities, and establish reciprocity as the basis for documentation. These themes are essential for a language documentation tool to integrate to the design, but their inclusion in the final project output is also dependent on project stakeholders conducting research in an ethical fashion that supports the outlined themes.

### 3.1 Linguistic Sovereignty

Amongst those working on language documentation, the importance of the work is often discussed either in terms of data preservation or cultural preservation. While both motivations are interested in the knowledge contained in language, data preservation focuses on how knowledge stored in all of the languages of the world can inform research. In one such example of language as data, Himmelmann (2006) describes the importance of language documentation as it secures current and future researchers' access to information from various language communities and allows others to validate claims made in such research by cross-referencing records in the language. The utility of language in research is evidenced by current research movements in a variety of fields, such as the integration of Indigenous knowledge in sustainability research (Ferguson and Weaselboy, 2020; Zidny et al., 2020).

Discourse emphasizing the importance of language documentation for cultural preservation is especially prevalent in language communities, as the ability of language to store important cultural practices motivates community members to participate in language documentation. Further, the importance of empowering languages within communities is intensified by research connecting the health of speakers to linguistic engagement in the community, such as reports showing significant correlations between decreased youth suicide rates in Indigenous communities wherein at least half of the community members had some proficiency in their native language (Hallett et al., 2007). Motivation based on cultural preservation highlights the role of language in supporting and empowering a community, as language documentation efforts can assist in community projects that build on culturally appropriate practices to address community needs (Barker et al., 2017; Brady, 1995).

This section uses the broad phrase "linguistic

sovereignty" to encapsulate both the dichotomy between data preservation and cultural preservation and the importance of data sovereignty. When describing the passing of data to another party, questions of responsible data practices arise, particularly as they pertain to data sovereignty. Kukutai and Taylor (2016) define data sovereignty as "managing information in a way that is consistent with the laws, practices and customs of the nation-state in which it is located." Data has been described by many as the new medium for colonialism (Bird, 2020; Leonard, 2018; Ricaurte, 2019), and thus those working on language documentation projects must ensure that the data practices being used in the project are aligned with the community's ideals.

Suggestions for how to best protect a community's data sovereignty include the development of ethical research standards in the field of computational linguistics (Schwartz, 2022) and language documentation (Belew and Holmes, 2023), defining data sovereignty and privacy practices within communities (Leonard, 2018), and ensuring transparency in research through continuous collection of informed consent (Austin, 2010). However, as laws, practices, and customs of various language communities differ drastically, a well-designed tool must account for both restrictions to access and collaboration between individuals, as desired by whichever community is using the tool.

Language documentation projects also protect data sovereignty by ensuring community members understand how and for what their data is being used. If an annotation and data management tool only allows for an abstract representation of linguistic meaning that is outside of a culture's epistemological construction of their language, it threatens the ability of community members to understand how their language is being used and minimizes their agency in the documentation project. One example of the success of using culturally appropriate epistemological constructions for language is demonstrated by (tonh et al., 2018) in their work detailing the successful use of the root-word method in teaching community members the Kanyen'kéha language.

Clearly indicating the intended purpose for the data is also essential to data sovereignty, especially as advances in NLP permit the use of data in novel ways that may not be easily interpretable to contributors in language communities. For example, providing consent to use recordings as audio for entries in an online dictionary is markedly different from providing consent to use a series of recordings for speech synthesis.

### 3.2 Cultural Specificity

In the development of language documentation tools, there is a delicate balance between linguistic specificity and cross-linguistic extendibility. A tool developed specifically for one language produces a project outcome that is more detailed and accurate to the context of the community, while a tool built for general use with many languages produces project outcomes that may be useful to many communities, but often fall short in including all of the information that is important to the community. While cultural specificity and cross-cultural applicability may appear to be in conflict, a cross-culturally applicable tool can account for cultural specificity by allowing users to access features to customize the storage, presentation, and annotation of the data based on the preferences of the community.

As Brinklow (2021) suggests, a broad approach is not the responsibility of a language community and the development of language technology in the community's language should be the priority, as Indigenous-led projects have found success in starting with a language-specific approach before considering crosslinguistic extendibility (Kuhn et al., 2020). However, for computational linguistics working on the development of a tool, there is a responsibility to design a tool that can work in multiple cultural contexts, while still allowing for community-specific customization that accounts for the inclusion of data that marks culturally relevant phenomena in the language.

In addition to accounting for varying needs and interests in integrating technology, differences in community ethics necessitate the development of a community-based definition of ethics. While an academic researcher may be bound to a code of conduct or ethical framework from their field or another governing body, this code is unlikely to comprehensively address the community's definition of ethical research (Bow and Hepworth, 2019). Further, collaboratively defining ethical research within a language community is conducive to fostering a relationship between those working on a project (Belew and Holmes, 2023), thus supporting the next key consideration: reciprocity.

### 3.3 Reciprocity

Reciprocity in language documentation is fundamental to ensuring ethical research (Austin, 2010). Maiter et al. (2008) define reciprocity as an "ongoing process of exchange with the aim of establishing and maintaining equality between parties." In the context of language documentation, this exchange can be seen as the community providing a researcher with linguistic data in exchange for resource creation. The creation of new language resources helps to mitigate disparity in resource availability and thus contributes to the process of establishing equality. However, language documentation has historically prioritized the access of other researchers to research output (Henke and Berez-Kroeker, 2016), with the creation of community resources posed as the secondary goal (Austin, 2006). When the motivation for language documentation is the function of language as data, the natural result is a prioritization of a resource output.

Belew and Holmes (2023) discuss the importance of reciprocity through the role of relationship in "A Linguist's Code of Conduct: Guidelines for Engaging in Linguistic Work with Indigenous Peoples." This publication suggests ethical standards for language documentation and was written by a non-Indigenous and an Indigenous researcher. Belew and Holmes encourage researchers to view their methodology and approach to research by centering their relationship with the community. Listening is foundational to building and maintaining a relationship with the community and results in culturally appropriate research that addresses the needs and interests of the community. Extractive research is avoided by focusing on the relationship with the community and the reciprocal nature of the research.

## 4 Tool Features

The three themes - linguistic sovereignty, cultural specificity, and reciprocity - identified in section 3 relate and intersect in various ways to motivate the 8 annotation and data management tool features described below.

### 4.1 User Management

An annotation and data management tool must allow for control over project contributors in order to protect linguistic sovereignty. This can be accomplished through a user management feature which allows for a user that has been designated as a project administrator to add other users to an existing project. Linking login credentials to user profiles secures the data in the project and ensures that the community is able to control who accesses the annotated data. Further, project administrators should have the ability to select permissions on a tier-by-tier basis, as the skillset of different contributors determines the relevancy of different tiers. For example, it would be nonsensical to give edit permission on the "IPA transcription" tier to a contributor without experience in IPA.

### 4.2 Collaborative Editing

Collaborative editing allows for more than one individual to provide updates to a project at the same time. In both subsections 3.1 and 3.3, the role and importance of contributions from various members of the community in a language documentation project is discussed. Once a language community has decided on appropriate contributors for a project and which permissions various users should have, the efficiency of the language documentation work can be increased by allowing multiple parties to work on the annotation project at once. Collaborative editing allows for more contributors, which presents the opportunity to benefit from input from more community members.

### 4.3 Edit History

Edit history works in tandem with collaborative editing and user management to ensure participation and control over the project, thus supporting linguistic sovereignty. Edit history maximizes the ability to include multiple contributors by providing a time-ordered list that details the changes made to an entry and who made the change. A time-ordered list that tracks changes allows other users to collaboratively review each other's work and move towards consensus linguistic descriptions or the development of best practices for representing natural variation. Edit conflicts can be avoided by allowing users to check out the sentence or lexicon they are annotating or editing.

### 4.4 Customizable View

A feature allowing customizable views allows different contributors to see and engage with the parts of the language documentation project appropriate for their contribution and best suited to their skills. This feature supports linguistic sovereignty by ensuring that the community members participating in the project feel ownership, agency, and

confidence in the way their language is being represented. For example, although language speakers have implicit knowledge of their language, demonstrated by their ability to produce and comprehend syntactically complex phrases, some speakers may not be familiar with explicit linguistic knowledge, such as dependency structures, parts of speech, or semantic roles (Bowles, 2011). Further, being presented with this abstraction constantly may make them feel alienated from the documentation process. Additionally, asking for such information without explicit training could result in inaccurately annotated data and frustration from speakers.

In a study discussing the open issues posted about accessibility on GitHub, Bi et al. (2021) find that the user interface (UI) is the most mentioned issue. Thus, presenting speakers with a UI that has many tiers and fields for detailed annotation could be overwhelming, as is a concern when introducing common annotation and recording tools like FLEx or ELAN (Moeller, 2014). Allowing language projects to define views for various contributors also protects data sovereignty by assuring information is displayed to community member contributors in an accessible format that reflects the cultural understanding of linguistic representation.

### 4.5 Compatibility with Other Platforms

Creating a tool that supports cultural specificity necessitates an understanding of communities with diverse documentation histories. While some communities may be starting projects from scratch, others may have existing materials from previous projects that they want to reference. Further, different members of a documentation team may have strong preferences for continuing their contribution in a platform that is familiar to them. For this reason, the tool should allow users to export to and import from a variety of popular and historically common linguistic tools. Compatibility with other platforms via proper file conversion has the ability to support the integration of new technologies from the field of computational linguistics. For example, a complete text analysis in FLEx includes the lemma of word forms, part of speech tags, and additional morphological information for sentences that could easily be used as the basis for a CoNNL-U file, which is the standard format for syntactic annotation in the Universal Dependencies project (de Marneffe et al., 2021).

### 4.6 Open-source

Open-source development is integral to ensuring reciprocity in a language documentation tool. An open-source license allows for access to the source code of the tool and grants permission to modify and redistribute the produced code while specifying rules for licensing the derivative code (Sen et al., 2008), allowing a community to directly access the output of the project. Open-source licenses are especially popular in the field of computational linguistics as these licenses are reported to improve the success of projects by interesting more contributors (Stewart et al., 2006), alleviating restrictions placed on projects with limited data (Streiter et al., 2006), and ensuring the reproducibility of empirical research (Wieling et al., 2018).

While accounting for cross-cultural considerations in the design of a tool promotes cultural specificity in a project and improves the baseline utility for as many language communities as possible, further integration of the specific cultural context of a project has the potential to improve a developed tool. Thus, open-souring a project allows for further customization of the tool and encourages language documentation projects to further reflect on how technology can best serve their goals.

### 4.7 Modular Integration of Computational Linguistic Technologies

The integration of computational linguistic technologies has the potential to greatly aid language documentation projects, but not all tools will be of interest to all communities. For example, a language community focusing on oral language documentation is unlikely to be interested in using finite-state transducers (Pirinen and Lindén, 2014) or long-short term memory neural networks (Etoori et al., 2018) to develop a spell-checker. Therefore, users should have the option to integrate the tools they feel best align with their project goals. By allowing users to decide on which tools they will integrate into their project based on community needs, this feature supports linguistic sovereignty and cultural specificity.

The order in which technologies are integrated into the tool should be influenced by cross-cultural considerations. There will always be more technologies to integrate, but it is important to ensure that certain project applications are not being favored over others through their prioritization. For example, written documentation of language has

long been prioritized in the field of language documentation, which exacerbates the well-established promotion of literacy at the expense of orality (Vansina, 1985). As many languages are primarily oral, a cross-culturally applicable tool that supports language documentation must ensure the ability of a community to use oral methods of language documentation.

Current NLP tools have a variety of applications for language documentation, both in support of the development of the understanding of a language and in the creation of pedagogical resources. As discussed in section 3.1, prioritizing language documentation for data preservation over cultural preservation often results in different goals for a project. For example, those working towards data preservation may be more interested in developing linguistic theory for the language while those working towards cultural preservation may prefer to prioritize the use of NLP tools that can help build pedagogical tools, such as the Kawennón:nis verb conjugator developed by Kazantseva et al. (2024) for Kanyen'kéha learners.

### 4.8 Transparency of Data Policies

Transparency of how data is stored and shared with others is an integral part of protecting the linguistic sovereignty of communities. An annotation and data management tool is tasked with clearly communicating how it is ensuring secure handling of a project's data and communicating any risks associated with passing the data through third-party NLP tools. Notifications should be presented to users to clearly indicate when data is being processed through another platform and consent should be requested if the data is being stored by the platform in any way.

While open-sourcing is common in computational linguistics, it is not appropriate in all cultural contexts. Ensuring data sovereignty necessitates that language communities decide who should have access to their data (Kukutai and Taylor, 2016). As the licensing of the tool as open-source is separate from the licensing of any linguistic data, projects are able to select licenses for their data based on ethical and cultural considerations within their community (Moshagen et al., 2013).

## 5 Conclusion

Advances in computational language documentation have the potential to support community-led initiatives by designing tools with cross-cultural considerations as the foundation. While cross-linguistic extensibility often comes at the expense of cultural specificity, designing modularity and customization into the tool's features and the user interface empowers users to shape the tool to their specific cultural and linguistic context. Existing research in language documentation, data sovereignty, and community-led research initiatives should inform those working on designing computational language documentation tools. Following the intentional design of a cross-culturally applicable tool, the tool should be further developed in consultation with multiple language communities.

## Acknowledgments

We would like to acknowledge the makers of previously developed language documentation tools for their contributions to language documentation and credit them as a source of inspiration in our work. Additionally, we would like to thank language community members and scholars working to define the best approaches to linguistic sovereignty, reciprocity, and ethical language work, as their work is of utmost importance in guiding our research and the direction of the field.

## Limitations and Ethics Statement

The main limitation of this project is the lack of direct language community involvement. This paper attempts to address cross-cultural considerations by referring to research that encourages and demonstrates involvement by and feedback from a variety of language communities, but subsequent research should include consultation with multiple language communities.

## References

Peter K. Austin. 2006. *Chapter 4 Data and language documentation*, pages 87–112. De Gruyter Mouton, Berlin, New York.

Peter K. Austin. 2010. Communities, ethics and rights in language documentation. *Language Documentation and Description*, 7(00).

Brittany Barker, Ashley Goodman, and Kora DeBeck. 2017. Reclaiming indigenous identities: Culture as strength against suicide among indigenous youth in canada. *Canadian Journal of Public Health = Revue Canadienne De Sante Publique*, 108(2):e208–e210.

Anna Belew and Amanda Holmes. 2023. A linguist's code of conduct: Guidelines for engaging in linguistic work with indigenous peoples.

Tingting Bi, Xin Xia, David Lo, and Aldeida Aleti. 2021. A first look at accessibility issues in popular github projects. In *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 390–401.

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. *Computational Linguistics for Lessstudied Languages: Texas Linguistics Society*, 10.

Catherine Bow and Patricia Hepworth. 2019. Observing and respecting diverse knowledge traditions in a digital archive of indigenous language materials. *Journal of Copyright in Education & Librarianship*, 3(1).

Melissa A. Bowles. 2011. Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in Second Language Acquisition*, 33(2):247–271.

Maggie Brady. 1995. Culture in treatment, culture as treatment. a critical appraisal of developments in addictions programs for indigenous north americans and australians. *Social Science & Medicine*, 41(11):1487–1498.

Nathan Thanyehténhas Brinklow. 2021. Indigenous language technologies: Anti-colonial oases in a colonizing (digital) world. *WINHEC: International Journal of Indigenous Education Scholarship*, 16(1):239–266.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Pravallika Etoori, Manoj Chinnakotla, and Radhika Mamidi. 2018. Automatic spelling correction for resource-scarce languages using deep learning. In *Proceedings of ACL 2018, Student Research Workshop*, pages 146–152, Melbourne, Australia. Association for Computational Linguistics.

Jenanne Ferguson and Marissa Weaselboy. 2020. Indigenous sustainable relations: considering land in language and language in land. *Current Opinion in Environmental Sustainability*, 43:1–7.

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.

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.

Darcy Hallett, Michael J. Chandler, and Christopher E. Lalonde. 2007. Aboriginal language knowledge and youth suicide. *Cognitive Development*, 22(3):392–399.

Ryan E Henke and Andrea L Berez-Kroeker. 2016. A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation*, 10.

Nikolaus P. Himmelmann. 2006. *Chapter 1 Language documentation: What is it and what is it good for?*, pages 1–30. De Gruyter Mouton, Berlin, New York.

Anna Kazantseva, Brian Maracle, Owennatékha, Ronkwe'tiyóhstha Josiah Maracle, and Aidan Pine. 2024. Kawennón:nis: the wordmaker for kanyen'kéha - nrc publications archive.

Roland Kuhn, Fineen Davis, Alain Désilets, Eric Joanis, Anna Kazantseva, Rebecca Knowles, Patrick Littell, Delaney Lothian, Aidan Pine, Caroline Running Wolf, Eddie Antonio Santos, Darlene A. Stewart, Gilles Boulianne, Vishwa Gupta, Brian Maracle Owennatékha, Akwiratékha' Martin, Christopher Cox, Marie-Odile Junker, Olivia Sammons, Delasie Torkornoo, Nathan Thanyehténhas Brinklow, Sara Child, Benoit Farley, David Huggins-Daines, Daisy Rosenblum, and Heather Souter. 2020. The indigenous languages technology project at nrc canada: An empowerment-oriented approach to developing language software. In *COLING*, pages 5866–5878. International Committee on Computational Linguistics.

Tahu Kukutai and John Taylor. 2016. *Indigenous Data Sovereignty: Toward an agenda*. ANU Press. Accepted: 2017-02-17.

Wesley Y. Leonard. 2018. *Reflections on (de)colonialism in language documentation*. University of Hawai'i Press.

Yanfei Lu, Patrick Littell, and Keren Rice. 2024. Empowering Oneida language revitalization: Development of an Oneida verb conjugator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5757–5767, Torino, Italia. ELRA and ICCL.

Sarah Maiter, Laura Simich, Nora Jacobson, and Julie Wise. 2008. Reciprocity: An ethic for community-based participatory action research. *Action Research*, 6(3):305–325.

Sarah Ruth Moeller. 2014. Review of saymore, a tool for language documentation productivity.

Sjur N. Moshagen, Tommi Pirinen, and Trond Trosterud. 2013. Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 343–352, Oslo, Norway. Linköping University Electronic Press, Sweden.

Tommi A. Pirinen and Krister Lindén. 2014. State-of-the-art in weighted finite-state spell-checking. In *Computational Linguistics and Intelligent Text Processing*, page 519–532, Berlin, Heidelberg. Springer.

Paola Ricaurte. 2019. Data epistemologies, the coloniality of power, and resistance. *Television & New Media*, 20(4):350–365.

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.

Ravi Sen, Chandrasekar Subramaniam, and Matthew L. Nelson. 2008. Determinants of the choice of open source software license. *Journal of Management Information Systems*, 25(3):207–239.

Siddharth Singh, Ritesh Kumar, Shyam Ratan, and Sonal Sinha. 2022. Towards a unified tool for the management of data and technologies in field linguistics and computational linguistics - LiFE. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 90–94, Marseille, France. European Language Resources Association.

Katherine Stewart, Anthony Ammeter, and Likoebe Maruping. 2006. Impacts of license choice and organizational sponsorship on user interest and development activity in open source software projects. *Information Systems Research*, 17:126–144.

Oliver Streiter, Kevin P. Scannell, and Mathias Stuflesser. 2006. Implementing nlp projects for non-central languages: instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4):267–289.

tonh, Jeremy Green, and Owennatékha Brian Maracle. 2018. *The Root-Word Method for Building Proficient Second-Language Speakers of Polysynthetic Languages: Onkwawén:na Kentyókhwa Adult Mohawk Language Immersion Program*. Routledge.

J. Vansina. 1985. *Oral Tradition as History*. James Currey.

Linda Wiechetek, Flammie A. Pirinen, Børre Gaup, Trond Trosterud, Maja Lisa Kappfjell, and Sjur Moshagen. 2024. The ethical question – use of indigenous corpora for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15922–15931, Torino, Italia. ELRA and ICCL.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in Computational Linguistics: Are We Willing to Share? *Computational Linguistics*, 44(4):641–649.

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).

Robby Zidny, Jesper Sjöström, and Ingo Eilks. 2020. A multi-perspective reflection on how indigenous knowledge and related ideas can improve science education for sustainability. *Science & Education*, 29(1):145–185.

V. Zue and R. Cole. 1979. Experiments on spectrogram reading. In *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 116–119.