Design factors and data sovereignty in developing tools for Kanien'kéha language learning*

Wíshe Mittelstaedt
Ionkwahronkha'onhátie'

Austin Kraft
McGill University

Abstract: Of interest to formal syntacticians and language teachers, we present our extensions to the open-source LingView project for visualizing language documentation files (Pride, Tomlin, and AnderBois 2020). The project specifically supports the community-led revitalization of the Northern Iroquoian language Kanien'kéha (a.k.a. Mohawk), but the interface and our learner-centred approach are relevant for language documentation at large. We integrate language-specific morphosyntax and classroom input to design a user-friendly interface for time-aligned transcriptions and their accompanying media. This project offers a case study in how contemporary frameworks for Indigenous data sovereignty, such as OCAP® (First Nations Information Governance Centre 2014), are technologically feasible for linguists who are involved in the study of under-documented languages.

Keywords: Kanien'kéha, language revitalization, language documentation, pedagogy, inflectional morphology

1 Introduction

Through the specific example of a language documentation interface for Kanien'kéha (Northern Iroquoian), we seek to explicitly relate formal linguistic description to revitalization — particularly in how the former can service the latter. This paper is concerned with the means to make language documentation usable for the various stakeholders in language documentation, with an emphasis on leveraging granular linguistic detail to support the language goals of the speaker-learner communities themselves. Because of the wide diversity of backgrounds and approaches in language documentation, we make our definitions of the italicized keywords explicit in (1).

(1) Working definitions of key terms in this discussion of language documentation:

a. *Language documentation*: "mono- or multi-modal records of language in use"

* We thank the Kanien'kéha speakers, teachers, and learners for sharing their knowledge with us, both about the Kanien'kéha language and about ways to improve the LingView interface to better serve learners. In particular, we thank Mary Onwá:ri Tekahawáhkwen McDonald and Akwiratékha' Martin for their central role in making LingView possible, first as a pilot project and now as an ongoing resource. Thanks also to the students of McGill's Field Methods course in Fall 2023, particularly Jinny Choi, Blair Jackson, Alex Ito, and Xindi Zhang. We are grateful to the audience at WSCLA 27 as well as Chase Boles, James Crippen, Jessica Coon, and the Roti'nikonhrowá:nens group at McGill University for their feedback. Any shortcomings in this paper are our own.

In *Proceedings of the Workshop on Structure and Constituency in Languages of the Americas* 27,
University of British Columbia Working Papers in Linguistics,
Ella Hannon and Bruce Oliver (eds.). Vancouver, BC: UBCWPL, 2025.

b. *Stakeholders in language documentation*: "the various and often overlapping groups of people with an interest in the production and/or usage of language documentation, including but not limited to: first-language (L1) speakers, advanced second-language (L2) speakers, language students, language teachers, community cultural institutions, linguistics professors, linguistics students, and software developers"

c. *Language goals of speaker-learner communities*: "the fixed and ongoing goals held by a group of language speakers and learners regarding the future vitality, use, and status of their language"

Intended for formal linguists and teachers of any language, this article spares most of the programming-related details of the open-source LingView project. Where appropriate below, we direct interested readers to Pride, Tomlin, and AnderBois's (2020) introduction to LingView. The Kanien'kéha code repository is also available at https://kanienkeha-lingview.lingspace.org/.

Section 2 introduces the Kanien'kéha language and some community-led revitalization efforts in recent decades. Section 3 identifies cross-linguistic and Kanien'kéha-specific approaches to language documentation in service to revitalization, highlighting the approaches' respective successes and challenges. With this context, we conclude the section with an overview of the LingView interface for language documentation files. Section 4 presents a case study on Kanien'kéha language resource design, focusing on two areas of Kanien'kéha inflectional morphology challenging for formal linguists and L2 learners alike: pronominal prefixes and aspect suffixes. After describing how we have bridged different terminology systems using LingView, we describe feedback from Kanien'kéha learners and teachers. We mention this feedback not only to outline our next steps but also to demonstrate an ongoing process of community consultation in directing the future of the resource. In Section 5 we frame this project in relation to Indigenous data sovereignty, discussing our efforts to uphold the OCAP® framework (First Nations Information Governance Centre 2014) as an example of how linguists can integrate culturally relevant data governance into their language documentation projects. Section 6 concludes.

2 Overview of Kanien'kéha

2.1 Kanien'kéha background

Kanien'kéha, the Mohawk language, is one of five languages in the Five Nations Iroquois branch of the Northern Iroquoian language family (Mithun 2014). It is spoken in eight communities in Ontario, Quebec, and New York: Ohswé:ken (Six Nations of the Grand River), Wáhta, Kenhtè:ke (Tyendinaga), Ahkwesáhsne, Kahnawà:ke, Kanehsatà:ke, Ganienkeh, and Kana'tsioharè:ke. According to DeCaire (2023), there are approximately 562 L1 speakers living on-territory and 77 advanced L2 speakers.

Kanien'kéha is a polysynthetic language that has been described as having free or discourse-configurational word order, which has made it a focus of linguistic researchers for its puzzling syntax (Baker 1996; Mithun 1999; DeCaire, Johns, and Kučerová 2017). Due to the polysynthetic nature of the language, theoretical and descriptive approaches extensively discuss possible morpheme categories used in the language (Bonvillain 1973; Mithun 1999, 2005; Baker 2003; Michelson 2023; Martin 2023). There are three primary lexical classes in Kanien'kéha — particles, verbs, and nouns — with verbs constituting the vast majority of forms. The combinatorics of Kanien'kéha verbs are vast. Verb roots can incorporate up to one noun root in a word, and there are at least 11 verb prefixes, seven verb suffixes, and three post-aspectual suffixes (Mithun 2005).

The language makes use of three main aspects: the Habitual, the Punctual, and the Stative. Counting their allomorphs, there are upwards of 300 pronominal prefixes, which vary based on the initial sound of the noun or verb root to which they attach.

2.2 Kanien'kéha vitality

Kanien'kéha is a language that is primarily spoken by elders (DeCaire 2023). There are, however, a number of organizations actively working towards language revitalization, with teaching offered to both adults and children. The main five adult immersion programs are: Onkwawén:na Kentyóhkhwa, in Six Nations; Shatiwennakará:tats, in Tyendinaga; the Kanien'kéha Language Program, in Ahkwesáhsne; Ratiwennahní:rats, in Kahnawà:ke, and Ratiwennenhá:wi, in Kanehsatà:ke. There are immersion programs and language nests for school-aged children, operating in nearly every community.

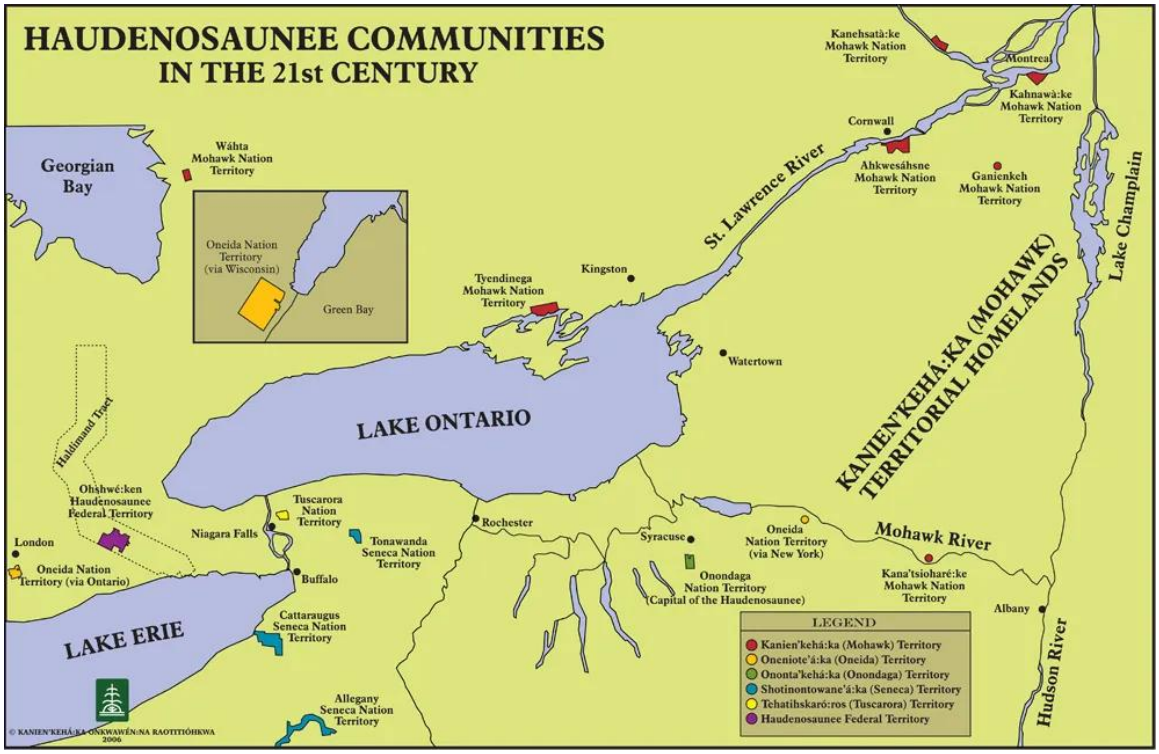


Figure 1: Map of Haudenosaunee Communities (Kanien'kehá:ka Onkwawén:na Raotitióhkhwa, 2006).

Despite ongoing revitalization efforts across communities, Kanien'kéha remains a critically endangered language. In the 2016 Canada Census, 2,350 individuals self-reported as speakers of the Mohawk language (Statistics Canada 2016), and in 2021 that number was 1,435 (Statistics Canada 2021). Because census data for language use is self-reported, these numbers could represent anyone from a fluent L1 speaker to somebody who has learned how to say a few scripted sentences. For a much more detailed account of the language's vitality, see DeCaire (2023). Therein, DeCaire estimates the number of living, advanced speakers for each reserve community, and evaluates the vitality of Kanien'kéha using a number of different metrics. In Ahkwesáhsne, DeCaire identifies

350 L1 and 10 L2 speakers, and rates the language vitality as Severely Endangered (UNESCO), or 8a Moribund (EGIDS). For Kahnawà:ke, DeCaire identifies 150 L1 and 40 L2 speakers, and rates the language vitality as Severely Endangered (UNESCO), or 8a Moribund (EGIDS). In Kanehsatà:ke, DeCaire identifies 60 L1 and five L2 speakers, and rates the language vitality as Severely Endangered (UNESCO), or 8a Moribund (EGIDS). In Ohswé:ken, DeCaire identifies zero L1 and 15 L2 speakers, rating the language vitality as Critically Endangered (UNESCO), or 8b Nearly Extinct (EGIDS). For the reserve Tyendinaga, DeCaire identifies zero L1 and six L2 speakers, rating the language vitality as Critically Endangered (UNESCO), or 8b Nearly Extinct (EGIDS). Finally, in Wáhta, DeCaire identifies two L1 speakers and one L2, rating the language vitality as Critically Endangered (UNESCO), or 8b Nearly Extinct (EGIDS). For all of the communities that DeCaire (2023) examined, he scored them an 8 on the GIDS vitality metrics rating: "The only remaining speakers of the language are members of the grandparent generation" (Lewis and Simons 2010, adapted from Fishman 1991 as cited in DeCaire 2023, pg. 37).

With this context in mind, we present the LingView interface as a tool for language learners, both in the classroom and outside of it. As we discuss in the following sections, one particular benefit of LingView is that it shares the speech of the elderly L1 speaker population with those who may be from communities where they no longer have elders to speak with.

3 Challenges and proposals for accessible language documentation

Many of the considerations in our efforts to produce language documentation of Kanien'kéha are ubiquitous in documentation of understudied and endangered languages. In this section, we consider the general and Kanien'kéha-specific considerations in designing accessible language documentation.

3.1 Summary of tools related to LingView

LingView is not the first effort to render language documentation files into more accessible formats. We contextualize LingView in relation to the strengths and weaknesses of past efforts. In doing so, we highlight opportunities to make LingView more effective in supporting speaker-learner communities' language goals.

The IATH ELAN Text-Sync Tool (ETST), presented in Dobrin and Ross (2017), is among the most similar proposals to the LingView approach advocated for in this paper. Like LingView, the primary objective of ETST is to present complex language documentation files in an approachable interface. As the authors underscore, ELAN's main functions are transcription and analysis, not the dissemination of language documentation into user-friendly formats. They echo Berez (2007) in drawing attention to ELAN's steep learning curve. ETST is offered as a means to access language documentation files – both their transcriptions and the audio/video media being transcribed – without requiring a background in computer programming or ELAN transcription.

Dobrin and Ross (2017) describe the full process for configuring ETST to display ELAN files. One of the primary differences and limitations of the IATH ETST compared to LingView is that it requires ELAN files to follow a specific template of hierarchically structured tiers. In other words, the lines of the interlinear glosses require a specific ordering, and those lines require specific labels: "Transcription" and "Gloss." Furthermore, only those two lines are displayed in ETST. Other lines, such as a breakdown of morphemic boundaries, are not displayed in ETST. Made concrete with an example from Kanien'kéha in (2), this glossed word illustrates what information in a standard interlinear gloss would be omitted in an ETST display. The example in (2) demonstrates how the

visibility of the morphemes and glosses could be invaluable for learners, in understanding why a short English expression like 'I went shopping' is so lengthy in Kanien'kéha.[1]

(2) Example form from Kanien'kéha. Note that only the top (orthography) and bottom (free translation) lines would be visible in the ETST interface. The second (morphemes) and third (gloss) lines would not be visible.

wakatkehrontakohò:ne

wak-at-kehnront-ako-hon-hne

1SG.PAT-SRFL-peddle-REV-STAT.PRF-RMT

'I went shopping.' (DeCaire 2023; pg. 7)

While ETST's two-line maximum might make language documentation more approachable, less is not always more in terms of usability. Consider the potential uses of language documentation in Kanien'kéha pedagogy and revitalization. Explicit training in different aspects of formal linguistics, such as phonologically conditioned allomorphy, is a part of contemporary Kanien'kéha learning programs – for instance, in pedagogical frameworks like the Root Word method (Green and Maracle 2019) and in the Ratiwennahní:rats immersion program at the Kanien'kehá:ka Onkwawén:na Raotitióhkwa Language and Cultural Center in Kahnawà:ke, Quebec. ETST, in displaying only two lines, would require a choice between presenting the surface (orthographic) form or the underlying (phonemic) form. Excluding the orthographic form would limit students' opportunities to practice Kanien'kéha spelling. Excluding the phonemic form would obscure allomorphy; the ETST output would not indicate that two forms are in fact allomorphs. Most importantly, either exclusion would disrupt the mapping between underlying and surface forms – the core relationship that must be grasped in order to acquire advanced phonological knowledge of the language.

ETST offers significant contributions in making language documentation more accessible: minimal programming prerequisites, lightweight storage requirements, and ability for online as well as offline presentation. In considering the limitations that it imposes on the granularity of transcriptions, we look to LingView instead in this project. In principle, LingView can display arbitrarily complex ELAN files. Informed by feedback from Kanien'kéha teachers and learners, we extend LingView to enable users themselves to set the degree of complexity. As in the case of phonologically conditioned allomorphy, pedagogically informed complexity can have a role in community-oriented language documentation resources.

For further discussion of ETST and earlier tools like the Customizable Presentation of ELAN Documentation (CuPED; Cox 2011) and their limitations, we direct readers to Dobrin and Ross (2017) and Pride, Tomlin, and AnderBois (2020).

3.2 Past Kanien'kéha-specific proposals

Documentation of Kanien'kéha began in earnest hundreds of years ago with the work of Jesuit missionaries, who began visiting Kanien'kehá:ka communities and learning the language of the people who resided there. For the most part, this work remains inaccessible to language learners today. Not only are these sources difficult to find, but once found, they often require linguistics

[1] The glossed data in this paper follow the Leipzig conventions, with the following additions: PAT for patient, SRFL for semi-reflexive, REV for reversive, and RMT for remote past.

training since Kanien'kéha orthography has significantly changed in the centuries since. Furthermore, Jesuit missionaries were French speakers and wrote their commentaries on Kanien'kéha grammar in French. Many Kanien'kehá:ka people today do not speak or read French, with the exception of only some in Kanehsatà:ke and Kahnawà:ke. In general, these materials have not yet been turned into any valuable teaching or learning material for Kanien'kéha students.

Kanien'kéha language documentation continues today. Through the works of modern scholars (Mithun 2005; Michelson 2023; Baker 1996; Bonvillain 1973, among others), a great volume has been published on the structure of the language, with a wealth of examples of L1 speech. Nonetheless, this work remains inaccessible to language learners for similar reasons: (a) the language used in academic texts can be very difficult to parse and understand for a language learner untrained in linguistics; (b) these works often require access to university library systems or academic journals, and are otherwise locked behind paywalls; and (c) in works written before the 1990s, the orthography used is different from the one practiced today in language learning environments.

For the most part, the most easily accessible forms of language documentation available to learners are audio recordings that have been posted either on SoundCloud or YouTube. Kanien'kehá:ka Onkwawén:na Raotitióhkhwa (KOR), based in Kahnawà:ke, has produced a weekly all-Kanien'kéha radio talk show since the 1990s, and this is the most expansive volume of audio recordings available to language learners. There are now other radio shows that have followed it, such as the Ahkwesáhsne All-Kanien'kéha Radio Show and Tsi Niiá:wen's in Kanehsatà:ke. While these are indeed valuable recordings for language learners, they require at least an intermediate to advanced fluency in the language before their value can be realized by students, primarily because there are no accompanying translations or explanations.

The LingView tool that we are presenting in this paper seeks to bridge this gap between existing language documentation and learner needs: it presents fluent speakers' speech along with time-aligned transcriptions, translations, and explanations of the words being used and the morphological composition of the words, with modern orthography that language learners are accustomed to using in their classrooms.

3.3 Motivating LingView

LingView, developed by Pride, Tomlin, and AnderBois (2020), is an open-source interface for displaying language documentation files. As a webpage or an offline desktop application, LingView offers users the ability to control which tiers are visible in time-aligned presentation with the audio or video file.

The core functions of LingView are described in Pride, Tomlin, and AnderBois (2020); we invite interested readers to read the original paper and code repository.[2] Like ETST that preceded it, LingView is motivated by the desire to render complex language documentation files into formats that are approachable and useful to the stakeholders in language documentation and revitalization. Originally developed in the documentation of the A'ingae language, LingView itself is a language-independent interface for viewing documentation files.

While our discussion focuses on the usage of LingView with ELAN files, LingView is also compatible with FLEx files. An explicit goal of LingView is to "[leverage] the complex, structured ELAN and FLEx files practitioners are already producing… to pass along this rich structure to

[2] The original LingView repository is available for authorized usage and replication, under the MIT license, at https://github.com/BrownCLPS/LingView.

users" (Pride, Tomlin, and AnderBois 2020, pg. 89). LingView employs greater structural flexibility than ETST to acknowledge how virtually any representational commitment – such as spelling principles or surface versus underlying forms – will never meet the needs of all communities who stand to benefit from a language documentation project.

Like ELAN itself, LingView is an open-source project: it is licensed to be freely duplicated and modified. While the tools for setting up LingView are publicly available and can be extended for individual projects' uses, the transcription data presented in LingView is not inherently publicly available. As we discuss in section 5, this latitude in data access is important for Kanien'kéha language documentation and revitalization: it provides a no-cost and transparent tool for creating resources that the community might want to keep under certain access restrictions.

Further illustrated in section 4, the core function of LingView is to display detailed linguistic transcriptions with user controls for which layers ("tiers" in the terminology of ELAN files) are visible on the display. Pride, Tomlin, and AnderBois (2020) identify five design principles in the development of LingView:

(3) LingView design principles, paraphrased from Pride, Tomlin, and AnderBois (2020; Sec. 2.2)

a. Multimedia compatibility: the transcription can be accompanied by audio, video, or neither.

b. End-user compatibility: a user can control exactly which layers or tiers of the transcription are displayed.

c. Non-proprietary input: no special formatting is needed for ELAN files, and the usage of ELAN involves no commercial gain for any party.

d. Practitioner-friendly pipeline: there are options to edit and inventory files en masse.

4 Contributions to LingView

Among the adaptations we have made with the open-source LingView repository from Pride, Tomlin, and AnderBois (2020), we focus this section on concerns about terminology and its accessibility. While this section takes the needs and situation of Kanien'kéha language revitalization as a case study, the foundational consideration is relevant cross-linguistically: to bridge the terminologies of formal linguistics and speaker-/learner-communities, emphasizing accessibility for speakers and learners.

4.1 Terminology and pedagogical value

The vast majority of L2 Kanien'kéha language learners begin their studies with no background in formal linguistics. However, as mentioned in section 3, language immersion schools do make great use of linguistic terminology in their instruction, especially in morphology. For instance, the curriculum in use at Onkwawén:na Kentyóhkhwa, Ratiwennahní:rats, Shatiwennakará:tats, Ratiwennenhá:wi and the Kanien'kéha Language Program draw heavily from linguistics, and students are required to familiarize themselves with the field's terminology. We anecdotally report that not all students are enthusiastic about this requirement in learning their language, despite the fact that these immersion programs have repeatedly shown that it is the most effective way to learn Kanien'kéha as a second-language adult learner.

Despite the heavy influence of linguistics in language learning classrooms, these concepts tend to go no further than providing a name for a feature of the language. Students may know a particular

feature of the language as the 'perfective', but they are unable to describe what that means. They will learn that one given conjugation of a verb is called the 'punctual aspect', but do not understand what aspect or tense are. The influence of linguistics in language learning classrooms does not necessarily translate into an ability for language learners to understand linguistic papers as they pertain to Kanien'kéha language use and documentation. For the most part, linguistic s concepts are used to name features of the language in the classroom, so that instructors and students may more easily reference and talk about the language that they are teaching or learning. Despite this only surface-level use of the terminology, it remains incredibly useful as a pedagogical tool in language learning classrooms, since English and Kanien'kéha have very few morphosyntactic similarities for students to otherwise reference.

4.1.1 Case study: pronominal prefixes and aspect suffixes

The use of linguistics in classroom environments has also led to new challenges for language learners. In some instances, the terminology being applied in some language learning classrooms does not align with formal descriptions. For example, in a number of language immersion programs, the stative form of a verb is called the perfective, which is a misconception deriving from the fact that the stative form is in fact a perfect construction. Likewise, the term 'stative' is used to refer to any present-tense interpretation of a verb, whether this occurs in the habitual present or the stative present. Some of these formal-pedagogical misalignments are contrasted in Table 1 below. These differences create additional challenges for language learners when they are searching for further resources outside of what is provided to them by their programs.

Table 1: Examples of tense and aspect mismatches in different Kanien'kéha-studying contexts.

Pedagogy: Language immersion	**Formal linguistics**
Stative	Habitual present or stative present
Perfective	Stative
Past tense or definite	Factual punctual aspect

Our objective is not to assert any label as the correct analysis. Rather, we point out that existing terminology varies significantly among Kanien'kéha-studying contexts. For a linguist interested in language documentation as a means of supporting the language goals of speaker-learner communities, addressing and bridging these terminology mismatches is central to making the documentation usable.

As mentioned earlier, Kanien'kéha makes use of an extensive system of pronominal prefixes in its word construction. Different language immersion programs make use of different terminology to refer to these pronominal prefixes, and many refer to them as a whole as 'pronouns.' While current linguistic works tend to refer to them as agent and patient pronominal prefixes, in many language learning environments they are called 'subjective' or 'objective,' and oftentimes, they are referred to as 'red' and 'blue' 'pronouns,' respectively. This latter choice is a reference to the colours that are often used to distinguish between the two different sets in language instruction materials. This once again creates a barrier-to-entry for Kanien'kéha language learners who seek additional materials in linguistic documentation to further progress their fluency, all the while still providing a valuable pedagogical tool for instructors and learners alike. We attempt to reconcile these differences in LingView, through a glossary function discussed in the next section, offering clarifications by providing (a) the current standard terminology used in formal linguistics for features of Kanien'kéha, and (b) what they are frequently termed in language immersion schools.

4.1.2 Implementing terminology choices in LingView

Bearing in mind the nuanced terminological decisions in morphemic glossing, we have adapted the Kanien'kéha LingView project to bridge the different terminological choices. Our objective is to make LingView approachable, educational, and easy to use for Kanien'kéha students. In ongoing pursuit of this objective, we have developed interactive glossing explanations for LingView transcriptions.

For linguists and learners alike, one of the hallmarks of Kanien'kéha is its polysynthetic morphology (Baker 1996, among others). The Root Word method of language instruction (Green and Maracle 2019) capitalizes on this internal complexity of Kanien'kéha words, and L2 learners report that their first instinct upon encountering a new Kanien'kéha word is to see its morphological breakdown. Linguists are like learners in this respect: linguistics papers and language documentation files employ multilinear glosses as in (2) to break down words and sentences for formal analysis.

Across formal linguistics, the Leipzig glossing conventions are the standard for labeling morphemes in interlinear glosses (Max Planck Institute 2015). This section noted the presence of linguistics terminology in Kanien'kéha classrooms. However, there are sometimes inconsistencies in terminology usage between classrooms and formal linguistics papers; at other times, there are inconsistencies between different language schools. We endeavour to bridge these terminological differences by embedding an interactive gloss dictionary in the Kanien'kéha LingView interface.

The original LingView interface contains a template for a dedicated Glossary page, in which documentarians can provide explanations of the glossing conventions. Inspired by this template, we have created a JSON file that maps Leipzig-style abbreviations to short and plain-language descriptions of the terms. Below, we have included several such abbreviation-description mappings. The description begins with the full spelling of the abbreviated term, followed by a non-technical definition and an English example where appropriate.

Table 2: Example abbreviation-description mappings from Kanien'kéha LingView.

Abbreviation	Description
BEN	Benefactive: The benefactive is used to say that someone/something was affected by the action either positively or negatively (e.g., *He cooked for me*).
FUT	Future: The future is a modal prefix that is used to refer to an action or event that has not taken place yet but will take place sometime in the future (e.g., *He will leave*).
JR	Joiner vowel: This marking refers to a vowel that gets inserted after a verb or noun root that ends in a consonant, and before another consonant. The vowel used is almost always a.

We have modified LingView's core display function so that when the cursor hovers over a word with Leipzig-style abbreviations that are included in the JSON mapping, a text box appears with the descriptions of all identifiable abbreviations. Figure 2 illustrates this function.

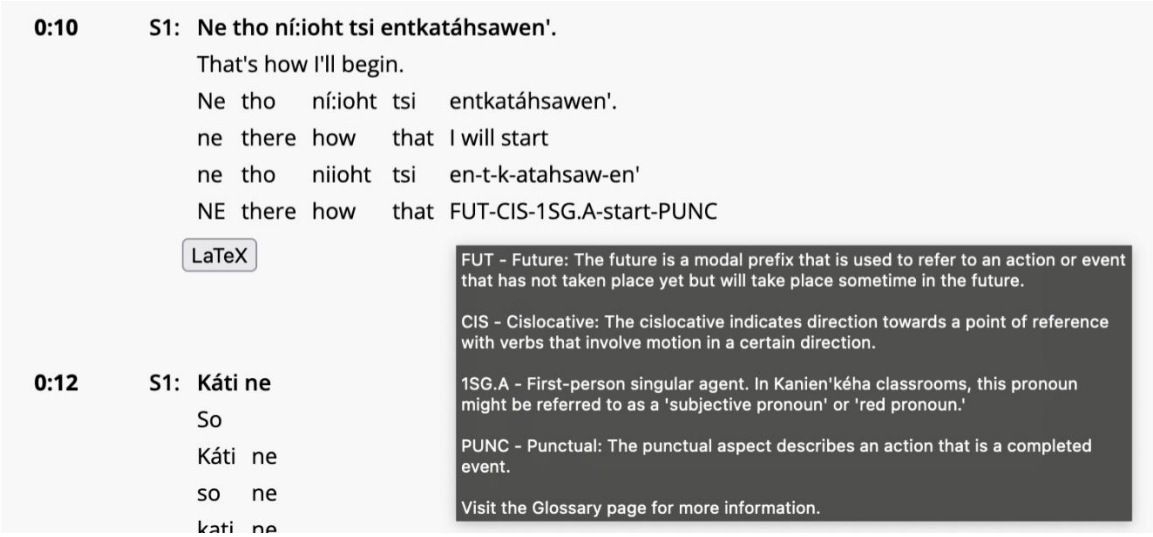


Figure 2: When the cursor hovers over a gloss on the website such as the gloss for the form *entkatáhsawen'*, a guide appears. In this case, the guide relates the 1SG.A pronominal prefix *k-* to classroom terms: 'subjective pronoun' or 'red pronoun.'

If there is an abbreviation in a transcription that is not yet included in the JSON mapping, it is effectively invisible to this functionality; the hovering glossing explanation will just show the identifiable abbreviations' corresponding descriptions.[3]

Figure 2 shows one specific instance of terminological differences across Kanien'kéha classrooms and formal linguistics documentation: the interactive pop-up explanations indicate that the morpheme labeled 1SG.A for 'first-singular agent' might be called a 'subjective pronoun' or 'red pronoun' in language classrooms. Just as the original LingView project sought to create accessible language resources without sacrificing the detail of formal glossing, these interactive explanations aim to bridge the terminologies of formal linguistics and language learning, without overwhelming the diverse user groups who stand to benefit from the website.

4.1.3 Site design and compatibility

With assistance from an L1 speaker and an advanced L2 teacher, we have created a preliminary version of the LingView interface's navigation in Kanien'kéha. That is, the text on the interface's menus and controls is itself in Kanien'kéha. English site navigation is also available. Pride, Tomlin, and AnderBois (2020) provide the template for additional site navigation languages. With Kanien'kéha experts' input, we include a Kanien'kéha version of LingView to parallel the Kanien'kéha site navigation of the Kawennón:nis online verb conjugator.[4] Our motivation was to further promote Kanien'kéha language learning: not only within the stories displayed in the interface, but in the usage of the interface itself being a means of language practice and study.

Like ETST, LingView was designed to allow for online and offline usage. However, in the years since the original release of LingView, the software dependencies enabling offline LingView

[3] At this time, the hovering glosses are simple HTML tooltips, which pose UX accessibility challenges. We are currently in the process of implementing more accessible functionality.
[4] The Kawennón:nis verb conjugator is available online at https://kawennonnis.ca.

had become outdated. As a result, the original LingView documentation for offline usage no longer supports the offline adaptation. Fortunately, we have been able to restore offline usage; the documentation for this offline capacity is in the Kanien'kéha LingView repository.

The benefits to an offline language documentation interface are numerous. From an infrastructure standpoint, an offline interface is usable in areas without stable internet. From an ethics standpoint, and what we explore further in section 5, an offline interface can convey greater control over who can access the data displayed in LingView. In its current form, LingView does not have granular access controls built in through user profiles or password protection as available for other language documentation resources. Access controls are an area for future work that Pride, Tomlin, and AnderBois (2020) also note.

4.2 Language learner and teacher feedback

In order to improve initial versions of LingView, we sought feedback from language learners and teachers alike. In general, the individuals that provided feedback thus far are either students or teachers at two-year adult immersion programs, although some advanced L2 learners also provided their feedback. The majority of those who provided feedback were from or reside in Ahkwesáhsne, although there were some responses from individuals in or from Kanehsatà:ke and Kahnawà:ke.

The feedback has been largely positive. Language learners remarked that they enjoyed being able to see morphological breakdowns of words that they were unfamiliar with, and they requested colour-coded morphemes. This preference is consistent with how the language is taught in schools, in particular the pronominal prefixes. They also enjoyed the time-aligned scrolling text, making it easier to follow along with the story that they were listening to. Some remarked that the option to modify playback speed (i.e., reduce it) would further make it easier to follow along. Some remarked that the glossary definitions were difficult to understand. The most consistent feedback that we received, however, was that there needed to be more stories added and that the visual appeal of the website could be improved. All language learners remarked that they were likely to use this website while studying. Teachers, on the other hand, remarked that they were unlikely to incorporate the website as part of their in-classroom instruction, instead saying that they would simply encourage their students to use it as part of their studying.

The process by which we collected feedback from students was done informally, with the exception of an online form that was sent out to a number of language stakeholders. Thus far, we have only been able to incorporate some of the feedback into website updates. In the future, having finished incorporating more of the initial feedback we've received, and to continue to improve LingView's value to language stakeholders, we would like to conduct a more robust study as to the opinions of both students and teachers alike.

5 Data sovereignty

Before concluding, we define the concept of data sovereignty, its relevance to the present LingView project, and the responsibilities it highlights for outsider linguists conducting language documentation work at large.

5.1 Overview of data sovereignty

Data sovereignty is a transnational concept, broadly defined as "the inherent and inalienable rights and interests of indigenous peoples relating to the collection, ownership and application of data about their people, lifeways and territories" (Kukutai and Taylor 2016, pg. 2). The concept is not

exclusively in reference to linguistics research, but linguistics's history of Indigenous knowledge exploitation puts data sovereignty in sharp focus.

Kukutai and Taylor (2016) offer a global overview of Indigenous-led efforts to advance and protect data sovereignty, including the OCAP® framework in Canada. OCAP® stands for Ownership, Control, Access, and Possession. In Kukutai and Taylor's (2016) volume, the First Nations Information Governance Centre (FNIGC) contextualizes OCAP® in the "recurring grievances about research and researchers over the years… [Non-Indigenous researchers] have frequently pre-empted meaningful community involvement by presenting completed research designs, often already funded, for community approval rather than collaborating from the start" (pg. 142).

Developed and administered by the FNIGC, OCAP® is a set of principles to express First Nations' jurisdiction over their own data. The framework is a registered trademark to protect the principles from becoming misconstrued or misused as virtue-signalling. To be explicit, our discussion of OCAP® here is not an endorsement from the FNIGC; our purpose is to draw attention to a framework that is particularly relevant and under-discussed in language documentation. Below, we summarize the four principles of OCAP®. We invite linguists researching, or hoping to research, First Nations languages to access further training resources at the FNIGC website.

(4) Our paraphrasing of OCAP® principles:

   a. Ownership: Cultural knowledge, in its various forms, is under the collective ownership of a community.

   b. Control: A community sets the terms of data collection, usage, and disclosure.

   c. Access: A community is required to have access to information about themselves.

   d. Possession: A community has mechanisms to express ownership and control over the data in their jurisdiction.

In seeking to practice more responsible language documentation and linguistics research practices, we are working to implement the principles of OCAP® with the Kanien'kéha LingView project in a way that is most appropriate for speaker-learner communities.

We highlight one aspect of the First Nations Information Governance Centre's OCAP® framework: possession. In this discussion, we echo the FNIGC in that the principles of OCAP® are intertwined. We cannot wholly isolate possession, but we focus on this principle due to space constraints in this paper.

In the context of Kanien'kéha language documentation, the OCAP® principle of possession refers to means of Kanien'kehá:ka-defined and Kanien'kehá:ka-owned means of controlling resources like LingView. Acknowledging the political complexity of implementing such a principle, we discuss two efforts to advance possession in Kanien'kéha LingView. We also emphasize that the discussions are ongoing in determining the appropriate usage of LingView in Kanien'kéha language documentation. We do not claim to act or decide on behalf of Kanien'kehá:ka leaders or communities.

First, we are actively working with a Kanien'kehá:ka cultural centre to host LingView data – both the lightweight interface and the data itself – on Kanien'kehá:ka land. Beyond a symbolic gesture, this choice gives Kanien'kehá:ka people physical control over the computing infrastructure that displays Kanien'kéha language to a broader audience. Connecting the data with the land itself is a means of self-determination.

Second, we are continuing to explore access-control mechanisms within LingView, such that there are granular degrees of access for language documentation files. It is a situation that is not unique to Kanien'kéha or to First Nations languages in general: there might be certain narratives that a community would like to be publicly available while other narratives are more suitable for a restricted audience. This restriction could be on the basis of sensitive narrative topics, be they political or cultural. Our proposals here are not to make those decisions but rather to make their implementation technologically feasible.

In implementing access control mechanisms, the availability of offline LingView immediately restricts access by limiting the interface and its contained documentation to a single computer. In the future, we look to offer gradable access control, such as through password-protected user profiles and different grades of access clearance as defined by the Kanien'kehá:ka communities themselves.

6 Conclusion

In this paper we motivated and described extensions to the open-source LingView interface for language documentation. We highlighted Kanien'kéha-specific considerations as a case study in language resource design, to serve the diverse interests of stakeholders in language documentation. Pronominal prefixes and aspect suffixes offer examples of bridging the terminology of formal linguistic description and classroom usage. Particularly in the context of the revitalization of a critically endangered language with limited classroom resources, making extant language documentation more learner-friendly increases the number of opportunities for learners to engage with native-speakers' speech — a demonstrated reported need for supporting advanced learners' progression in sustaining intergenerational transmission (e.g., DeCaire 2023).

Our Kanien'kéha project further represents an invitation to linguists at large who study under-documented languages to consider how data sovereignty factors into their own needs and goals. The ongoing Kanien'kéha LingView project aims to demonstrate to linguists that integrating data sovereignty is technologically feasible. Proactive, systems-level attention to data sovereignty stands to strengthen the research process itself, by centring interests of the speaker-learner community and promoting language research that meets demonstrated needs.

References (style: "References Section Heading")

Baker, Mark. 1996. *The Polysynthesis Parameter*. Oxford: Oxford University Press.

Baker, Mark. 2003. *Lexical Categories: Verbs, Nouns and Adjectives*. Cambridge: Cambridge University Press.

Berez, Andrea L. 2007. Review of EUDICO Linguistic Annotator (ELAN). *Language Documentation & Conservation* 1(2):283–289.

Bonvillain, Nancy. 1973. *A Grammar of Akwesasne Mohawk*. Ottawa: National Museum of Man.

Cox, Christopher. 2011. The ecology of documentary linguistic software development. Paper presented at the 2nd International Conference on Language Documentation and Conservation, Honolulu, February 11-13.

DeCaire, Ryan, Alana Johns & Ivona Kučerová. 2017. On optionality in Mohawk noun incorporation. In R. Maddeaux (Ed.), *Toronto Working Papers in Linguistics (TWPL)* 39.

DeCaire, Oheróhskon Ryan. 2023. *The role of adult immersion in Kanien'kéha revitalization*. Doctoral dissertation, University of Hawai'i at Hilo.

Dobrin, Lisa M. and Douglas Ross. The IATH ELAN Text-Sync Tool: A Simple System for Mobilizing ELAN Transcripts On- or Off-Line. *Language Documentation & Conservation* 11:54–102.

First Nations Information Governance Centre (FNIGC). 2014. Ownership, Control, Access and Possession (OCAP®): The Path to First Nations Information Governance. Ottawa, ON: The First Nations Information Governance Centre. https://fnigc.ca/ocap-training/.

First Nations Information Governance Centre (FNIGC). 2016. Pathways to First Nations' data and information sovereignty. In T. Kukutai and J. Taylor (Eds.), *Indigenous Data Sovereignty: Toward an Agenda*, 139-155. Canberra: Australian National University.

Green, Jeremy D. and Brian Maracle. 2019. The root-word method for building proficient second-language speakers of polysynthetic languages: Onkwawén:na Kentyóhkwa adult Mohwak language immersion program. In L. Hinton, L. Huss, and G. Roche (eds.), *The Routledge Handbook of Language Revitalization*, 83–93. New York: Routledge.

Kanien'kehá:ka Onkwawén:na Raotitióhkhwa. 2006. Haudenosaunee Communities in the 21st Century. https://www.korkahnawake.org/gift-shop/p/maps-amp-councils-posters.

Kukutai, Tahu and John Taylor. 2016. *Indigenous Data Sovereignty: Toward an Agenda*. Canberra: Australian National University.

Martin, Akwiratékha'. 2023. *Tekawennahsonterónnion: Kanien'kéha Morphology*. Kahnawà:ke, QC: Kanien'kehá:ka Onkwawén:na Raotitióhkhwa.

Max Planck Institute. 2015. Conventions for interlinear morpheme-by-morpheme glosses. Max Planck Institute for Evolutionary Anthropology. https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Michelson, Karin. 2023. Word classes in Iroquoian languages. In E. van Lier (ed.), *The Oxford Handbook of Word Classes*, 651-668. Oxford: Oxford University Press.

Mithun, Marianne. 1999. *The Languages of Native North America*. Cambridge: Cambridge University Press.

Mithun, Marianne. 2005. Mohawk and the Iroquoian languages. In P. Strazny (ed.), *Routledge Encyclopedia of Linguistics*, 699-702. New York: Routledge.

Mithun, Marianne. 2014. Gender and culture. In G. Corbett (ed.), *The Expression of Cognitive Categories: The Expression of Gender*, 131–160. Berlin: Mouton de Gruyter.

Pride, Kalinda, Nicholas Tomlin, and Scott AnderBois. 2020. LingView: A Web Interface for Viewing FLEx and ELAN Files. *Language Documentation & Conservation* 14:87–107.

Statistics Canada. 2016. The Aboriginal languages of First Nations people, Métis and Inuit. Available at: https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016022/98-200-x2016022-eng.cfm [Accessed on: 09/15/2025].

Statistics Canada. 2021. A diversity of Indigenous languages spoken across Canada, some growing and some in decline. Available at: https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2023029-eng.htm [Accessed on: 09/15/2025].