Cherokee Syllabary Texts: Digital Documentation and Linguistic Description

Jeffrey Bourns
Digital Scholarship Group, Northeastern University, Boston, MA, USA
j.bourns@northeastern.edu

Abstract

The Digital Archive of American Indian Languages Preservation and Perseverance (DAILP) is an innovative language revitalization project that seeks to provide digital infrastructure for the preservation and study of endangered languages among Native American speech communities. The project's initial goal is to publish a digital collection of Cherokee-language documents to serve as the basis for language learning, cultural study, and linguistic research. Its primary texts derive from digitized manuscript images of historical Cherokee Syllabary texts, a written tradition that spans nearly two centuries. Of vital importance to DAILP is the participation and expertise of the Cherokee user community in processing such materials, specifically in Syllabary text transcription, romanization, and translation activities. To support the study and linguistic enrichment of such materials, the project is seeking to develop tools and services for the modeling, annotation, and sharing of DAILP texts and language data.

2012 ACM Subject Classification Applied computing → Digital libraries and archives

Keywords and phrases Cherokee language, Cherokee Syllabary, digital collections, documentary linguistics, linguistic annotation, Linguistic Linked Open Data

Digital Object Identifier 10.4230/OASIcs.LDK.2019.18

Category Short Paper

1 Overview

The Digital Archive of American Indian Languages Preservation and Perseverance (DAILP) is an innovative language revitalization project that seeks to provide digital infrastructure for the preservation and study of endangered languages among Native American speech communities. DAILP is overseen by Northeastern University scholar Ellen Cushman, author of a recent study of the Cherokee Syllabary [2], and supported by the Digital Scholarship Group at Northeastern, the project's host institution [3]. The project's initial goal is to publish a digital collection of Cherokee-language documents to serve as the basis for language learning, cultural study, and linguistic research. Its primary texts derive from digitized manuscript images of documents recorded in the Cherokee Syllabary, a written tradition that spans nearly two centuries. Of vital importance to DAILP is the participation and expertise of Cherokee community members in the transcription, romanization, and translation of these texts. Further enhancements to DAILP texts will include phonemic romanization and free translation layers aligned with the Syllabary text, linguistic annotation, orthographic conversion functionality, parser development, and publication of project datasets as Linguistic Linked Open Data (LLOD). With project infrastructure in place, similar DAILP initiatives are envisioned for Ojibwe and other indigenous languages of North America. This paper describes resources, challenges, and early decisions informing the design and development of the DAILP Cherokee project.

2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 18; pp. 18:1–18:6
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Cherokee language and community

The Cherokee language (ISO 639-3, chr) belongs to the Iroquoian language family and survives as the sole representative of the Southern Iroquoian branch. Members of the distantly related Northern Iroquoian branch include Mohawk, Oneida, Onondaga, Seneca, Cayuga, and several further languages now extinct.

A recent report numbers speakers of Cherokee at approximately 12,300 people in the United States, including nearly 10,000 speakers of the Cherokee Nation community in northeastern Oklahoma and 1,000 speakers among the Eastern Band of Cherokee Indians in western North Carolina; to these estimates may be added an undetermined but relatively high percentage of speakers among the 7,500 members of the United Keetoowah Band of Oklahoma and Arkansas [5],[13]. Compared with other Native American languages, Cherokee has a relatively high number of speakers, but the language is spoken by few tribal members under the age of 40, and children at home no longer acquire Cherokee as their first language [12],[15]. Community efforts toward language revitalization include such initiatives as the establishment of Cherokee immersion schools since 2001, yet reversing the language shift will require more robust support for language learning and preservation. Vitality status currently assigned by UNESCO to Oklahoma Cherokee is “definitely endangered,” and North Carolina Cherokee is seen as “severely endangered” [11].

3 Cherokee Syllabary and written tradition

Among indigenous languages of North America, Cherokee is notable for its own writing system, the Cherokee Syllabary, and for a written tradition richly documented in this script. The Syllabary was devised in the early 19th century by Sequoyah, a Cherokee silversmith, who introduced the script to tribal leaders in 1821. In the years thereafter the Syllabary was quickly embraced by Cherokee society, which led to widespread literacy and official adoption by the Cherokee Nation in 1825. Compiled over nearly two hundred years, the documentary record of Cherokee Syllabary texts comprises newspapers, almanacs, religious tracts, hymns, laws, pamphlets, private correspondence, and also culturally sensitive materials, such as prayers and magic formulas recorded by traditional Cherokee doctors. Archival collections of Cherokee manuscripts have been preserved and cataloged by such institutions as Yale University and the Smithsonian’s National Anthropological Archives (NAA), and with the support of the Cherokee community, recent years have seen Syllabary manuscripts of cultural and historical interest digitized and published online [1].

4 DAILP goals and design

The DAILP initiative builds on digitization of historical Cherokee manuscripts. Under this approach, digitized Cherokee Syllabary documents provide the foundation for multi-layered text collections that can serve the diverse needs and interests of students and scholars of Cherokee language and culture. Project design is guided by the skills and requirements of the Cherokee community itself, particularly as these entail selection and preparation of texts and management of digital access. For gating and access to culturally-sensitive material, the DAILP collection will implement a system of protocols and permissions based on community-defined relationships and requirements. Archival Syllabary texts have been vetted and pre-selected by Cherokee translators for inclusion in the DAILP collection. Among these are numerous handwritten documents of uneven legibility for which automated processing via OCR is impractical. By design, DAILP workflows engage the Cherokee user community in processing these materials, specifically in Syllabary text transcription, transliteration, and translation activities. Among DAILP's initial goals are the design and development of an interface to support such tasks, informed by the skills and needs of project contributors.

Beyond these basic documentation activities lie more complex processing tasks. A key challenge for DAILP is support for the interpretation and annotation of text editions by contributors of varying levels of literacy and linguistic competence. To language learners and literate readers alike, historical Cherokee texts often pose significant difficulties due to the obscurity of lexical items, the morphological complexity of Cherokee language data, and the variety and ambiguity of Syllabary spellings. To support the interpretation and linguistic enrichment of such materials, the project is seeking to develop tools and services for the lexical and grammatical annotation of DAILP texts. Project editions thus annotated will also serve as a valuable source of primary language data for the development of further descriptive resources for Cherokee. Based on existing well-annotated datasets, recent contributions to Cherokee linguistics, and innovative language data management software, development of such infrastructure is currently underway.

5 DAILP language data

DAILP has acquired and enhanced several datasets of well-structured language data transcribed from descriptive resources for Oklahoma Cherokee. These datasets comprise Syllabary transcriptions, "simple phonetics" transliterations, phonemic representations, grammatical annotations, and English translations. The transcribed lexical data issues from three foundational sources for the study of Cherokee: *Cherokee-English Dictionary* [6], this dictionary's grammatical appendix [14], and *A Handbook of the Cherokee Verb* [7]. The main source is the dictionary, compiled by community linguist Durbin Feeling. Its appended grammatical outline is a rich source of annotated surface forms, and the verb handbook is similarly detailed and useful.

For phonemic representation, the dictionary and appendix use a romanized orthography known as the number system, which introduced a set of superscript numbers for marking Cherokee pitch patterns. Although unconventional, the number system is familiar and important to the community, thus DAILP plans to store and display surface forms transcribed faithfully from these sources in their original orthography. In addition to the number system transcriptions, a further DAILP dataset provides phonemic transcriptions of the same language data using conventional linguistic notation, which is practical for orthographic conversion functionality. Thus, for example, in addition to its Syllabary representation, the form for "I'm helping him" may be displayed as /jisdeliha/ (simple phonetics), /ji^{1}sde^{2}lị^{3}ha/(number system), or/jììsdeelíha/ (phonemic transcription) in the DAILP interface.

Crucially, these descriptive resources provide the project with an internally consistent generalization over Oklahoma Cherokee primary language data. Much in the way of many older manuscript traditions, spellings across historical Syllabary texts do not reflect an established standard. For DAILP's purposes, Syllabary spellings from the Feeling sources offer a practical standard under which orthographic and dialectal variants from DAILP texts may be subsumed. Surface forms in the Feeling sources are moreover linguistically conservative and preserve, e.g., final syllables, which are typically omitted in written sources. Especially valuable are Feeling's precise and consistent representations of vowel length and tonal configurations, which inform an important recent study of tonal behavior in Oklahoma Cherokee (TAOC) [16]. Together with the Feeling datasets, the specification of phonology in TAOC provides the DAILP project with a practical basis for parser development.

6 Linguistic resources for modeling Oklahoma Cherokee

For its modeling and annotation of project language data, DAILP has drawn mainly on two recent contributions to Cherokee linguistics: the systematic survey of phonology in TAOC, and a modern descriptive grammar of broader scope (CRG) [10]. Both TAOC and CRG offer valuable treatments of Oklahoma Cherokee, yet these works differ fundamentally in terms of orthographies, morphological analyses, terminologies, tagsets, and target audiences. A key early challenge for DAILP has been to identify and select from among these resources elements and approaches that are 1) practical for the design and implementation of DAILP tools and services, and 2) accessible and informative to a diverse community of users and contributors working with DAILP texts and language data.

For practical purposes, DAILP has made it a priority to deploy linguistic models and conventions that can straightforwardly support development of project infrastructure. Due to its primary reliance on TAOC for both example data and formulation of parser rewrite rules, DAILP has adopted the orthography, morphological analyses, and tags found in TAOC for the project's underlying representations, grammatical annotations, and specification of (morpho)phonology. In further support of this approach, the DAILP project has been fortunate to acquire a database of underlying lexical roots, stems, and affixes established by linguist Hiroto Uchihara, author of TAOC. By comparison with CRG, it should be noted, TAOC provides more granular morphemic segmentations of underlying forms. Accordingly, IGT examples presented in TAOC typically proceed from a deeper layer of derivation, and thus often require the application of more rules than CRG in order to generate well-formed surface forms. Despite this added complexity, the rigorous specification of phonology in TAOC is a significant windfall to project parser development, and DAILP's modeling decisions and dataset preparation reflect this practical advantage.

Designed for both linguists and language learners, CRG is an important descriptive resource for the study of Oklahoma Cherokee. For DAILP's purposes, the main value of CRG lies in its clear and concise explanations of Cherokee grammar and its many helpful examples. Given the complexity of Cherokee language data, ready access to the definitions and descriptions in CRG will be invaluable to users seeking to interpret and annotate DAILP texts, most practically via external reference to a published linguistic ontology. Ontology development moreover aligns with further interoperability goals of the project, based on best practices for Linked Data modeling and publication of DAILP datasets. Toward this end, DAILP is exploring development of Linguistic Linked Open Data (LLOD) tools for language-specific description of Cherokee, drawing on the domain knowledge of CRG as well as that of TAOC and several further resources. Due to the rich polysynthetic morphology of Cherokee, of particular interest to DAILP are such models as OntoLex and the Multilingual Morpheme Core Ontology (MMoOn Core) for representation of lexical and morphological language data [8].

7 Online Linguistic Database (OLD)

Due to multiple features well suited to the project, DAILP has installed and configured the Online Linguistic Database (OLD) as its language data management software. Created by linguist, developer, and DAILP project member Joel Dunham, the OLD is a program for creating collaborative language documentation web services [4]. The OLD was developed to meet the need for multi-user cross-platform tools for language documentation and analysis, and its software is designed specifically to support collaborative storing, searching, processing, and analyzing of linguistic data. Of special interest to DAILP is the OLD's well-documented utility in storing and analyzing language data from Blackfoot, a polysynthetic language of North America [4]. Likewise valuable is the OLD's parser development tool, which supports on-the-fly manual annotation based on user adjudication and selection of candidate parses. A further asset to DAILP is the OLD's orthographic converter, which enables users to select from among several familiar orthographies for the display of Cherokee phonemic representations. Project needs are also well served by the web services architecture of the OLD, which can interact seamlessly with the DAILP interface created for text processing by the user community.

8 Conclusion

As the pool of native speakers recedes and language shift encroaches on the Cherokee speech community, a sense of urgency attends the DAILP initiative. Interviewed for a recent article, Cherokee language translators working on NAA manuscripts report that these documents contain words and phrases that they hadn't heard in decades. A source for the same report estimates that nearly a third of lexical items attested in Smithsonian manuscripts are either no longer in current usage or else simply unknown [9]. Language revitalization is essential to the elucidation of historical Syllabary texts and to the discovery and preservation of Cherokee cultural and linguistic heritage. In partnership with the community, DAILP seeks to provide a durable window on this written tradition, and tools to help its linguistic heirs safeguard and illuminate its precious legacy.

References

1 Kilpatrick Collection of Cherokee Manuscripts. http://transcribe.library.yale.edu/projects/collections/show/2. Beinecke Library, Yale University.

2 Ellen Cushman. *The Cherokee Syllabary: Writing the People's Perseverance*. University of Oklahoma Press, 2012.

3 Digital Scholarship Group, Northeastern University. https://dsg.neu.edu/.

4 Joel Robert William Dunham. *The Online Linguistic Database: software for linguistic fieldwork*. PhD thesis, University of British Columbia, 2014.

5 David M. Eberhard et al. Cherokee. In David M. Eberhard, Gary F.Simons, and Charles D. Fennig, editors, *Ethnologue: Languages of the World*. SIL International, twenty-second edition, 2019.

6 Durbin Feeling. *Cherokee-English Dictionary*. Cherokee Nation of Oklahoma, 1975.

7 Durbin Feeling, Craig Kopris, Jordan Lachler, and Charles van Tuyl. *A Handbook of the Cherokee Verb: A Preliminary Study*. Cherokee National Historical Society, 2003.

8 Bettina Klimek. Proposing an OntoLex-MMoOn Alignment: Towards an interconnection of two linguistic domain models. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*, pages 1–16, 2017.

9 Robert Leopold. Articulating culturally sensitive knowledge online: A Cherokee case study. *Museum Anthropology Review*, 7(1-2):85–104, 2013.

10 Brad Montgomery-Anderson. *Cherokee Reference Grammar*. University of Oklahoma Press, 2015.

11 Christopher Moseley, editor. *Atlas of the World's Languages in Danger*. UNESCO, 3 edition, 2010.

12 Cherokee Nation. Ga-du-gi: A vision for working together to preserve the Cherokee language. Report of a needs assessment survey and a 10-year language revitalization plan. Technical report, Cherokee Nation of Oklahoma, 2003.

13 Endangered Languages Project. Cherokee. http://www.endangeredlanguages.com/lang/chr.

14 William Pulte and Durbin Feeling. Outline of Cherokee grammar. In *Cherokee-English Dictionary*, pages 235–354. Cherokee Nation of Oklahoma, 1975.

15 Elizabeth Seay. *Searching for Lost City*. The Lyons Press, 2003.

16 Hiroto Uchihara. *Tone and Accent in Oklahoma Cherokee*. Oxford University Press, 2016.