**Abstract:**

This paper describes a repository of example sentences in the Native American indigenous language, Choctaw. The repository provides documentation support for the threatened language, and allows researchers and language teachers novel texts.

**Introduction:**

This paper introduces a general use corpus for the American indigenous language, Choctaw. The Choctaw language is spoken by the Choctaw tribe, who originally inhabited the southeastern United States. Ethnologue (<https://www.ethnologue.com/cloud/cho>) classifies the Choctaw language as “Threatened”, as “the language is used for face-to-face communication within all generations, but it is losing users.” For many speakers, the language is their second language. Choctaw is spoken by all ages in Mississippi, but is only spoken by older adults in Oklahoma, thus it is of time importance to document the language as mature speakers age.

Like many indigenous languages, the language only gained a written form after contact with Europeans, in this case by a missionary in the late 1800s. The first text published in Choctaw was the Bible, and this continues to be the longest text in the language today. Efforts during the 1900s aimed to forcefully assimilate Native Americans and suppress indigenous languages. For this reason, many Native Americans did not learn their ancestral language, and few works are publicly published in these languages (Henze et al.). As a result of this history, the Choctaw language has few published works and little text representation online.

While work has been undertaken to document the language, and has been studied by linguists (see Broadwell, Haas, Nicklas, and Davies, among others), no digital corpus of the Choctaw language exists. This corpus has gathered sparse text resources representing different dialects of the language from teaching materials, books, and scholarly material. Video and transcribed audio clips where the Choctaw language is spoken in complete phrases also comprise this corpus.

For most of the world’s languages, no data in a machine-readable format are available for human language technology applications. Compared to well-resourced languages such as English and French, natural language processing tools for Native American indigenous languages are still not abundant. The goal of this database is thus first to preserve a threatened language. The second contribution is to make this data set available to historians, linguists, and natural language processing scientist for novel research opportunities, while also documenting the language for language learners and teachers.

**Language and tribe:**

The Choctaw tribe originally resided in the southeast of the United States, in what today would be Alabama, Louisiana, and Mississippi (Munro). The language belongs to the Muskogean language family, which also includes Cherokee, Seminole, and Chickasaw.

[MAP]

The last census of the Choctaw nation counted more than 220,000 enrolled members (<https://web.archive.org/web/20120512040555/http://www.ok.gov/oiac/documents/2011.FINAL.WEB.pdf>). Of these total population, only 10,400 are fluent (<https://www.ethnologue.com/language/cho>).

The language has two reference grammars, the first by the missionary Cyrus Byington (1870), the second by George Aaron Broadwell (2006). Two dictionaries also exist, one written by Cyrus Byington (1915), while the other was released by the Choctaw Nation of Oklahoma in 2016.

Several dialects of Choctaw exist, as American Federal Indian policy in the 1830s forced the Choctaws, along with the Chickasaws, Creeks, Seminoles, and Cherokees to give up their lands in the southeast and relocate in the migration known as the “Trail of Tears” to Oklahoma. While most Choctaws moved, a small number resisted removal and remained in Mississippi (Broadwell). Today, there are speakers primarily in Mississippi and Oklahoma (Davies), but also concentrations in Alabama, Louisiana, Texas, and California.

Due to the diaspora and resulting geographical isolation, there are three major dialects: Oklahoma Choctaw, Mississippi Choctaw, and Mississippi Choctaw of Oklahoma (Ulrich). Dialect differences in modern Choctaw are largely restricted to a few lexical items, with few differences in syntax or phonology (Broadwell). For example, Oklahoma Choctaws typically write “possum” as “shukhʋta”, while Mississippi Choctaws prefer “shokkata”, although the spoken word is identical for both dialects.

While the diaspora occurred after a written system had been devised, “Choctaw has been written in a variety of orthographies, and the choice of a writing system has been controversial, with different groups favoring different orthographies” (Broadwell), however all use Latin script. The orthographies vary from each other primarily in the way that vowels are written and in the representation of 3 consonants (Broadwell)”. The earliest orthography, called “traditional orthography”, was developed by nineteenth-century missionaries in producing translations of religious texts. A second orthography used by the Mississippi Choctaws, which was created in the mid-1970s, and a final “modified traditional orthography” is used by linguists in discussions of the language.

The Choctaw has several qualities that make it “not only very different from English but rather unusual cross linguistically” (Munro). The first feature is the pattern of word order, with subject-object-verb order.

Mr. Parrish ʋt aiittvtoba ia tuk.

Mr. Parrish went to the store. (TO DO notation)

Also unique is that adjectives follow the nous they modify.

Issuba lusa mʋt balihinli tuk.

That black horse continued to run. (TO DO notation)

The Choctaw language is polysynthetic, in that a complex sentence can be conveyed in a single verb word (Munro). The language presents novel linguistic challenges, as several papers have argued both for and against Choctaw being a language that is independent of Case marking making it one of the few natural human languages to represent this phenomenon (Woolford).

Ikintakobo

They’re not lazy. (TO DO notation)

**Data collection:**

The long-term goal of this project is to make all existing textual data from Choctaw digitally available for research. The current repository comprises written and oral teaching materials, books, linguistic articles and grey literature (referred to as scholarly sources for the remainder of this paper), audio clips, and videos.

|  |  |
| --- | --- |
| **Type** | **Number** |
| **Teaching materials** |  |
| **Books** |  |
| **Scholarly sources** |  |
| **Audio clips** |  |
| **Videos** |  |

Data was collected by crawling Youtube for videos. All audio clips were created by the Choctaw Nation of Oklahoma’s School of Language distance learning program. Published teaching material was gathered from the Choctaw Nation of Oklahoma and the Mississippi Band of Choctaw Indians.

The only material that was gathered from internet websites that were not scholarly sources were religious texts, such as hymns. To the best of our knowledge, no blogs or webpages entirely in the target language exist. Material that was only available in English was not included, as the primary focus of this corpus is to provide resources in the target language.

**Data set:**

Texts in the data set are stored in a database and are separated by dialect and type. The types of text collected include short stories, phrases, and poetry. Most the data are in both English and Choctaw, while monolingual data is stored separately, and all data points list the source reference.

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Oklahoma Choctaw** | **Mississippi Choctaw** | **Mississippi Choctaws of Oklahoma** |
| **Short stories** | 17 | 11 | 0 |
| **Phrases** | 1958 | 69 | 0 |
| **Poetry** | 3 | 0 | 0 |
| **Correspondence** | 1 | 0 | 0 |
| **Religious texts** | 5 | 1 | 0 |
| **Scholarly resources** | 219 | 12 | 1 |
| **Audio** |  |  |  |
| **Video** |  |  |  |
| **Untranslated** | 45 | 0 | 0 |

Stories and poems were all gathered from teaching materials. The stories cover a range of topics, from cultural, to personal anecdotes. The Mississippi Choctaw stories have 3028 word tokens, while the Oklahoma Choctaw stories have 5954 word tokens.

To form the phrases portion of the data set, only grammatically complete phrases were included in the repository. The Oklahoma Choctaw phrases have 9834 word tokens. The Mississippi Choctaw contains 1694 word tokens.

Grey literature, such as unpublished theses and manuscript, and published linguistic articles form the scholarly resources section of the corpus. In total, content in the target language was pulled from sixteen published articles, one unpublished manuscript, and two reference grammar books. This section contains the only annotation for part of speech of the language, and comes from the annotation provided by the author.

All audio and video data is stored apart from the database. All audio clips are accompanied by a text document with the Choctaw transcription and English translation document.

[Chart with number of clips, total lengths of content]

**Use cases:**

The proposed use cases for the data set are numerous given the variety and lack of specialization of the data. Primarily the use cases are for academic research in linguistics, history, and natural language processing (NLP), as well as serving to preserve the language.

Our future work is to annotate a training set of the complex morphemes of the language for study and to build a predictive model. Morphemes, for example in possession and referential coding (Nicklas 1971), are unique and challenging. As the data set includes both text and audio, a morpheme predictive model could be developed for audio.

One purpose within NLP is machine translation (MT). The majority of data is translated in English, creating a well-formed one-to-one data set. As no system yet exists, a MT system would assist in generating new texts in Choctaw from English. The language presents interesting challenges in this domain, as morphologically-rich languages pose problems for MT systems from errors in word-alignment and multiple affixes. Current alignment models at word-level do not distinguish words and morphemes, and produce low-quality end translation due to misalignment (Li et al.).

Finally, the corpus serves as a repository for teaching and learning. As nearly all of the entries are bilingual, learners and teachers alike can benefit from the translations.

**Discussion and Conclusion:**

Besides the cultural and psychological value in documenting endangered languages, there is also the scientific merits of preserving human languages. Munro, for example, says, “Differences among languages provide invaluable insight into how cognitive processes are actualized in language; such differences are part of biological diversity. A language’s passing reduces the diversity of human expressive power.” (pg 37)

Future work on this corpus aims to seek private sources for better representation of the Mississippi Choctaws of Oklahoma, to represent the Louisiana dialect of Choctaws, and include the Bible for all available dialects.

**References:**